

# A Purine-Pyrimidine Classification Scheme of the Genetic Code



Thomas Wilhelm, Svetlana Nikolajewa

**Although containing the same information, our new classification scheme of the genetic code is simpler than the common representation as a three-dimensional matrix: it contains just 32 instead of 64 fields. Moreover, it shows known patterns in the code more clearly than the common scheme. Above all, with the help of our new scheme we could identify new patterns never seen before. This gives rise to some speculations about the origin and early evolution of the genetic code. We hypothesize that coding started in a binary doublet manner and developed via a quaternary doublet code to our contemporary quaternary triplet code. Most interestingly, it may be possible to discover traces of the old binary coding in present-day genomes.**

The genetic code specifies how the information contained in the nucleic acids DNA and RNA is translated into the correct sequence of amino acids building the highly specific proteins. Up to the three termination codons UGA, UA(G/A) (standard code), each nucleotide triplet stands for exactly one amino acid, the methionin codon AUG is also the start codon. The genetic code is comma-free and non-overlapping. It is usually represented as a three-dimensional matrix in which the four rows stand for the first base and the four columns for the second base. To show the third dimension (the

third base) in the plane figure, each of the 16 boxes is again divided into four fields, giving together 64 entries (Fig.1).

For a long time one assumed that the genetic code is universal for all life forms on earth. Today there are at least 16 slightly deviating different codes known ([www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi](http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi)). However, it is generally believed that all these deviations are later descendants of the earlier standard code. Not surprisingly, non-standard codes are only found in small genomes, nearly all of them in mitochondria known to have by far the smallest genomes.

Since the early days of the discovery of the genetic code non-random patterns have been searched in the code for providing information about its origin and early evolution. In 1965 Nirenberg finished his famous project of deciphering the code. At that time most scientists believed that the code is the result of pure chance and hence does not need any further evolutionary explanation. Crick [1] formulated the corresponding "frozen accident" hypothesis which was widely accepted for many years. However, today it is assumed that at least some hints of possible evolutionary scenarios can be found in our contemporary code. The top-down approach, which we are following here, analyzes patterns in the code 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.

It has been appreciated for a long time that the genetic code assigns similar amino acids to similar codons. Two different rationales have been presented: first, mutation and translation error minimization [2], and second, similar amino acids tend to directly interact with similar RNA sequences [3]. It was stated that "the canonical code is at or very close to a global optimum for error minimization" [4]. It has also been proposed that instead of the actual codons, some of their derivatives, such as the anticodons or codon-anticodon duplexes were the original amino acid binding motifs. It is also possible that the original amino acid recognition took place at the tRNA acceptor stem. Szathmáry [5] proposed that amino acid-RNA allocation took place even

		2nd base					
		U	C	A	G		
1st base	U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	3rd base	U C A G
	C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg		U C A G
	A	AUU Ile AUC Ile AUA Ile AUG Ile	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg		U C A G
	G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly		U C A G

Fig. 1: The common representation of the standard genetic code (mRNA triplets in the mRNA reading direction (5'–3')). Shaded regions show codon families.

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.

However, the first codon position seems to be correlated with amino acid biosynthetic pathways and to their evolution as evaluated by synthetic “primordial soup” experiments. The second position is correlated with the hydrophobic properties of the amino acids, and the degeneracy of the third position could be related to the molecular weight or size of the amino acids [6]. Lagerkvist [7] observed that codon families (the amino acid of a codon family is uniquely determined by the first two nucleotides of a codon) have a much higher probability to appear in the left part of the common illustration scheme (cf. Fig. 1). He also 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.

### The New Classification Scheme of the Genetic Code

Most amino acid properties show no clear pattern in the common scheme of

the genetic code. Recently we proposed a new classification scheme [8 and [www.imb-jena.de/~sweta/genetic\\_code](http://www.imb-jena.de/~sweta/genetic_code)], based on a binary purine(1)-pyrimidine(0) coding (Fig. 2). It shows known regularities more clearly than the common scheme and it even highlights some new patterns.

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

Fig. 2: The purine(1)-pyrimidine(0) classification scheme of the standard genetic code. The third base is given in parenthesis. If there are differences between the standard code and any other code, the number of deviations from the standard code is indicated. 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 some bacteria the 21<sup>st</sup> amino acid, selenocysteine, can also be encoded by UGA. Shaded regions show codon families. The point in the center indicates the perfect point symmetry corresponding to Halitsky's family – nonfamily symmetry operation [9]. The thick horizontal line marks the symmetry axis for codon-anticodon symmetry.

There are three possible variants of a binary coding scheme for the genetic code: One could group the bases (i) according to base-pairs (A,U = 1, G,C = 0), (ii) according to keto- and aminobases (G,U = 1, A,C = 0), and (iii) according to purines and pyrimidines (A,G = 1, C,U = 0). In such a simplified code eight different binary triplets exist: 000, 001, ..., 111. Each of these binary triplets represents eight different codons, e.g. in our coding scheme 000 stands for CCC, CCU, ..., UUU. The purine-pyrimidine coding is superior to the other two variants, because it is the only one that allows the genetic code to be represented using just four columns (Fig. 2). The reason for this vast simplification in our scheme is that for the third position in each triplet it only matters if it is a purine or a pyrimidine.

Given the primary purine-pyrimidine coding, we have again two different possibilities to sort the first two bases per row: one can use either of the remaining two binary codings, according to base-pairs or according to keto- and aminobases as a sort criterion inside the rows. We have chosen the base-pairs for sorting inside rows, because only this reveals the following regularities of the genetic code: (i) All codon families group together, i.e. they are not scattered all over the table. (ii) More importantly, the codon strength classification directly corresponds to the columns in our scheme (cf. Fig. 2). Thus, in the first column the first two bases complementary pair with 6 hydrogen bonds, in the sec-

ond and third column with 5, and in the fourth column with just 4 hydrogen bonds. For all these reasons our classification scheme of the genetic code is superior to all similar ones.

Our new scheme shows some fascinating regularities. We can, for instance, better understand the number of different tRNAs in some organisms. In the simplest case one should expect one tRNA per coding field in our scheme. Exactly this happens in the case of vertebrate mitochondria. It is known that animal mitochondria contain exactly 22 different tRNAs. In vertebrate mitochondria UA1 and AG1 are stop codons. Thus there are exactly 22 fields for amino acids left: the 8 codon families plus 14 remaining fields. Interestingly, the 22 tRNAs in animal mitochondria correspond 1:1 to these 22 fields.

The amino acids of the nine "strong groups" (mutually evolutionary conserved, based on the alignment score matrix PAM250, cf. [http://bioinfolab.unl.edu/em-lab/documents/clustalx\\_doc/clustalw.txt](http://bioinfolab.unl.edu/em-lab/documents/clustalx_doc/clustalw.txt)) very closely group together in our scheme, more closely than in the standard scheme. That means neighboring amino acids in our scheme have a higher probability to be aligned to each other in genome comparisons than neighboring amino acids in the standard scheme.

Our new scheme also led us to detect hitherto unknown regularities of amino acid properties in the genetic code. Jungck [10] collected 15 different measures of amino acid properties. For all of these we arranged a table with 8 rows and 4 columns corresponding to our scheme. Amazingly, the column sums of nearly all measures are perfectly correlated to the corresponding codon-anticodon binding strength. For instance, the first column harbours more polar amino acids, the last column less polar ones and the mixed codon fields are in between. Similarly, the bulkiness and the specific volume increases continuously from the first to the last column.

## Evolution of the Genetic Code

The observed regularities inspire to some speculations about the early evolution of the genetic code. Thus the strong correlation between amino acid properties and codon strength implies that the first two positions together (and not the second position alone as speculated by others) must have been important for the amino acid – codon assignment in the early evolution of the code. It therefore also could be that just the first two nucleotides of a codon (or anticodon) show specific binding affinity to the corre-

sponding amino acid (maybe important in the process of the code formation).

Nowadays one assumes that "the code probably underwent a process of expansion from relatively few amino acids to the modern complement of 20" [11]. Can we find some hints in our scheme indicating coding of less than 20 amino acids in ancient times? Indeed, there is a high redundancy for each second row. This gives rise to the speculation that in the early days of code evolution just the first two bases of the triplet were coding. The reading frame, however, arguably always comprised three letters. In any way, a quaternary doublet can encode at most 16 amino acids, or 15 plus one termination codon (some bacteria exist that do not possess any stop codon). In this context it is interesting to note that Asn, Gln, Met, Trp, and Tyr seem to be newer amino acids.

Since the discovery of the genetic code it is speculated that the first genetic material contained only a single base-pairing unit [1]. Recently, for the first time a ribozyme was found composed of only one purine and one pyrimidine [12]. Assuming a binary doublet code, it is tempting to speculate which four amino acids, one per two consecutive rows, were the first encoded ones. In the first two rows Ser seems to be the oldest amino acid, and in the third and fourth row Ala. The 01-rows obviously contain no really old amino acid while the 11-rows contain more than one: Gly, Asp, Glu. 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 might also have occurred.

## Conclusions

We have found a concise scheme for the genetic code that is superior to similar schemes for different reasons. It shows nice patterns and symmetries and even so far unknown regularities in the code. We are now studying the fascinating question whether we still can find traces of doublet coding or even binary coding in contemporary genomes.

References are available from the authors.

**Dr. Thomas Wilhelm**  
**Svetlana Nikolajewa**  
Theoretical Systems Biology  
Institute of Molecular Biotechnology  
Beutenbergstr. 11  
07745 Jena, Germany  
wilhelm@imb-jena.de  
sweta@imb-jena.de

---

---