THE PURINE-PYRIMIDINE CLASSIFICATION
SCHEME REVEALS NEW PATTERNS IN THE
GENETIC CODE

S. L. Nikolaeva, T. Wilhelm*
Institute of Molecular Biotechnology, Jena
Beutenbergstr. 11, D-07745,
Jena Germany

Abstract
We present a new classification scheme of the genetic code, based on a binary representation of purines (A, G = 1) and pyrimidines (U, C = 0). On the basis of this logical organization we detect codon regularities (strong, mixed and weak codons), patterns of amino acid properties and show symmetry characteristics of the genetic code. We demonstrate the “codon - reverse codon” pattern and use it to remove the only ambiguity which had remained in our recently proposed new classification scheme of the genetic code. The purine-pyrimidine scheme has now found its final form.
We show that the “codon - reverse codon” pattern is correlated to aspects of the tRNA anticodon distribution in all organisms. It is well known that there is no tRNA with an anticodon for any of the STOP codons; we found that there is also no tRNA containing reverse STOP anticodons. Finally, we show that the new classification scheme also provides hints about the early evolution of the genetic code.

Keywords
Genetic code, Amino acid properties, tRNA.

Introduction
The genetic code is a set of rules by which living organisms convert genetic information from the DNA into life. The alphabet of RNA contains four letters, which can be divided into two pairs with similar biochemical characteristics: purines (adenine and guanine) and pyrimidines (cytosine and uracil). We propose a new classification scheme of the genetic code, based on a binary representation of purines and pyrimidines. The new scheme allows us
- to show symmetry characteristics of the genetic code,
- to detect patterns of amino acid properties and codon regularities (strong, mixed and weak codons),
- to understand why there are 22 tRNAs in mammalian mitochondria,
- to speculate about evolution of the genetic code, starting with a binary doublet code and developing via a quaternary doublet code to our contemporary quaternary triplet code.

Two Representations of the Genetic Code
During translation a triplet of nucleotides (codon) translates the genetic information from mRNA into a single amino acid. Since there are 20 different amino acids and $4^3 = 64$ possible codons the genetic code is redundant.
Table 1 shows the common classification scheme of the genetic code: the four rows stand for the first base in the codon, the four columns represent the second base and the right site indicates the third base in the triplet. The dark regions indicate “family codons”, where the encoded amino acid is independent of the third position.
Based on the purine-pyrimidine classification of the bases (purines A, G = 1 and pyrimidines U, C = 0) we developed a new scheme of the genetic code (Table 2). It consists of 8 rows numbered from 000 up to 111, due to the 2³=8 binary representations of all possible codons. Each row contains again 8 possibilities, for instance codon 000 (three pyrimidines) represents the 8 codons: CCC, CCU, ..., UUC and UUU. Because of the third position degeneracy, the number of columns can be reduced to four. Each field of Table 2 stands for two codons, where the third bases are given in parentheses. For instance, the codon CCC means that the two codons CCC and CCU encode for Proline. Therefore the new scheme contains only 32 fields instead of 64, although the information content of both tables is the same.

Next we concentrate on the first two bases of the codons (similar to Lagerkvist, 1978, 1981). The four combinations of the first column (CC*, GC*, CG*, and GG*) always imply 6 hydrogen bonds in complementary base-pairing with the corresponding anticodon of the tRNA. Accordingly, they are called strong codons. Their third base does not matter for a determination of the corresponding amino acid (family codons). In the next two columns the first two bases yield 5 hydrogen bonds (mixed codons). The upper half of these two columns contains family codons and the lower half not. Finally, the weak codons (last column) only have just 4 hydrogen bonds in complementary base pairing of the first two nucleotides.

If the third position is needed for the determination of the correct amino acid (column 4 and the lower halves of columns 2 and 3), it is sufficient to determine if there is a purine or a pyrimidine, with the only exception of the two fields Trp/Stop and Met/Ile, where the translation machinery has to analyze the third purine base exactly.

Interestingly, they correspond to the begin and end of translation.

**Two Symmetry Characteristics of the Genetic Code**

Our new representation of the genetic code shows symmetry characteristics. The bold horizontal line in Table 2 is the codon – anticodon symmetry axis. For instance, the Proline codons CCC and CCU have the anticodons GGG and GGA which code for Glycine.

The star in the center indicates the Halitsky family-non-family symmetry. Halitsky (2003) recently found that by exchanging the two keto bases A and C and the two amino bases G and U one gets a mapping of family codons to non-family codons.

**The Amino Acid Patterns**

Most amino acid properties show no clear pattern in the common representation of the genetic code. The purine-pyrimidine scheme highlights known regularities more clearly than the common scheme and it even reveals new patterns (Wilhelm and Nikolajewa 2004). Thus, it becomes clear that most amino acid properties (for instance hydrophobicity) are strongly correlated to the corresponding codon-anticodon binding strength. Additionally, we showed that the deviations of the standard genetic code are confined to specific regions in Table 2.

With the help of Table 2 we can also 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 the 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.

**Codon – Reverse Codon Patterns of the Genetic Code**

A reverse codon of any codon XYZ is defined as ZYX, where X,Y,Z can be any base. The arrows in table 2 represent pairs of “codon - reverse codons”. For instance, the reverse codon of CCC (Pro) is UCC (Ser). There exist 15 different amino acids in the rows 000, 101, 010 and 111 where the codon is reverse to itself, e.g. Lys (AAA), Tyr (UAU).

Considering codons and their reverse codons we report three observations.

1. Table 2 can be divided into four blocks (codon - reverse codon groups) of the same size, for instance the upper left block with Pro (P), Ser (S), Ala (A) and Thr (T). Each block shows the same arrow pattern. All strongly evolutionary conserved groups of amino acids (Thompson et al., 1994) are subsets of exactly one codon - reverse codon group, e.g. the MILV amino acids belong to the upper right block in the table. The other conserved strong groups belonging to one block are STA, NEQK, NHQK, NDEQ, QHRK, MILF, HY. The only exception is FYW.

2. We studied all known tRNA genes of 104 different organisms (Sprinzl et al., 1999). It is known that the STOP codons do not have any tRNA. We found that there are also no tRNA genes containing anticodons reverse to the STOP anticodons (ACT, ATC and ATT in Table 3). This is true for archaea (16), bacteria (81) and most eukaryotes (7). The only exception is *H. sapiens*, possessing one tRNA*αaa* gene with the anticodon ATT, but humans also have three different possible suppressor tRNA genes (Lowe and Eddy, 1997).

3. There are some codon - reverse codon pairs where there only exist tRNAs for one codon, but no tRNA for the reverse codon, e.g. all 104 studied organisms have at least one tRNA for Tyr with anticodon GTA (some organisms, for instance *H. sapiens*, have different tRNAs with the same anticodon, altogether there are 189 different tRNAs with the anticodon GTA in the 104 species), but no organism has a tRNA for His with anticodon ATG.

Table 3 lists different codon - reverse codon pairs. It is interesting to note that in the whole superkingdom bacteria the number of tRNA genes for the reverse part of table 3 is always zero, the few exceptions are only in eukaryotes and archaea. Another unique property of all bacteria is that there is no tRNA with anticodons for the self reverse codons UUU (Phe), UCU (Ser), UGU (Cys) and UAU (Tyr). Generally, table 3 shows that anticodons with an A at the third position are strongly preferred, whereas A** anticodons are significantly suppressed.

**Table 3. The amino acids codon- reverse codon pairs and the number of tRNA genes for the corresponding anticodons.**

<table>
<thead>
<tr>
<th>AA codon</th>
<th>anti-codon</th>
<th># tRNA</th>
<th>AA codon</th>
<th>anti-codon</th>
<th># tRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop UGA</td>
<td>TCA</td>
<td>0</td>
<td>Ser AGU</td>
<td>ACT</td>
<td>0</td>
</tr>
<tr>
<td>Stop UAG</td>
<td>GTA</td>
<td>0</td>
<td>Asp GAU</td>
<td>ACT</td>
<td>0</td>
</tr>
<tr>
<td>Stop UAA</td>
<td>GTA</td>
<td>0</td>
<td>Asn AAU</td>
<td>ATT</td>
<td>1</td>
</tr>
<tr>
<td>Tyr UAC</td>
<td>GTA</td>
<td>189</td>
<td>His CAU</td>
<td>AGG</td>
<td>0</td>
</tr>
<tr>
<td>Trp UGG</td>
<td>CCA</td>
<td>144</td>
<td>Gly GGU</td>
<td>ACC</td>
<td>0</td>
</tr>
<tr>
<td>Leu UUA</td>
<td>TAA</td>
<td>127</td>
<td>ile AUU</td>
<td>AAT</td>
<td>1</td>
</tr>
<tr>
<td>Leu UUG</td>
<td>CAA</td>
<td>134</td>
<td>Val GUU</td>
<td>AAC</td>
<td>18</td>
</tr>
<tr>
<td>Phe UUC</td>
<td>GAA</td>
<td>154</td>
<td>Leu CUU</td>
<td>AAG</td>
<td>28</td>
</tr>
<tr>
<td>Ser UCA</td>
<td>TGA</td>
<td>151</td>
<td>Thr ACU</td>
<td>AGT</td>
<td>19</td>
</tr>
<tr>
<td>Ser UCC</td>
<td>GGA</td>
<td>116</td>
<td>Pro CCU</td>
<td>AGG</td>
<td>16</td>
</tr>
</tbody>
</table>

**Evolution of the Genetic Code**

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, it has been noticed that there are some hints of possible evolutionary scenarios of the genetic code (Woese 1967). Our new scheme indicates additional patterns relevant for possible early code evolution. Jungck (1978) collected 15 different measures of amino acid properties. We observed the strong monotonic behaviour of these properties versus the codon strength (Wilhelm and Nikolajewa, 2004). It implies that both first triplet positions together and not the first or second position alone must have been important for the amino acid - codon code (Wilhelm and Nikolajewa, 2004).

**Conclusions**

We introduced the purine-pyrimidine classification scheme of the genetic code, which provides new insight into regularities of codons and amino acid properties. It helps to understand two hitherto unexplained observations:
first, the number of tRNA genes in mammalian mitochondria and second some aspects of the tRNA anticodon distribution in different organisms. Our findings suggest that evolution seems to avoid tRNAs with anticodons that are reverse to the anticodons of other tRNAs.

In our first publication we discussed the only ambiguity remaining in the arrangement of the new classification scheme: we could not find a definitive order of the second and third column (Wilhelm and Nikolajewa, 2004). With the here discussed “codon-reverse codon” pattern we can remove this ambiguity. Now all codons have a definite position in the purine-pyrimidine classification scheme of the genetic code (Table 2).

We speculated about the early evolution of the genetic code and are now studying the fascinating question whether we still can find traces of doublet coding or even binary coding in contemporary genomes. Recently we found such non-random binary patterns for DNA binding sites (Nikolajewa et al., 2005). The best studied restriction enzymes of type II contain the significant motif: two purines or two pyrimidines. For example EcoRI with recognition sequence GAATTC and BamHI with GGAATTC have common purine (1) - pyrimidine (0) pattern 111000.

Additional information can be found at http://www.imb-jena.de/~sweta/genetic_code.

Acknowledgments

We thank A. Beyer, R. Brockmann, M. Friedel and J. Hollunder for critically reading the manuscript. This work was supported by Grant no. 0312704E of the Bundesministerium fuer Bildung und Forschung, Germany.

References

http://lowelab.ucsc.edu/GRNAdb/