Journal of Bioinformatics and Computational Biology
Vol. 4, No. 2 (2006) 609–620

Imperial College Press
www.icpress.co.uk

THE NEW CLASSIFICATION SCHEME OF THE GENETIC CODE, ITS EARLY EVOLUTION, AND tRNA USAGE

SWETLANA NIKOLAJEWA, MAIK FRIEDEL, ANDREAS BEYER and THOMAS WILHELM*

Theoretical Systems Biology
Institute of Molecular Biotechnology Beutenbergstr
11, Jena, D-07745, Germany
**wilhelm@imb-jena.de*

Received 19 September 2005
Revised 9 December 2005
Accepted 23 December 2005

We present a new classification scheme of the genetic code. In contrast to the standard form it clearly shows five codon symmetries: codon-anticodon, codon-reverse codon, and sense-antisense symmetry, as well as symmetries with respect to purine-pyrimidine (A versus G, U versus C) and keto-aminobase (G versus U, A versus C) exchanges. We study the number of tRNA genes of 16 archaea, 81 bacteria and 7 eucaryotes to analyze whether these symmetries are reflected in the corresponding tRNA usage patterns. Two features are especially striking: reverse stop codons do not have their own tRNAs (just one exception in human), and **A**** anticodons are significantly suppressed. Our classification scheme of the genetic code and the identified tRNA usage patterns support recent speculations about the early evolution of the genetic code. In particular, pre-tRNAs might have had the ability to bind their codons in two directions to the corresponding codons.

Keywords: Genetic code; evolution; tRNA.

1. Introduction

The genetic code specifies how the information contained in the nucleic acids is translated into the correct sequence of amino acids. It is usually represented as shown in Fig. 1. Since the early days of the discovery of the genetic code, patterns have been searched for gaining insights into its origin and early evolution.[8] It is known that the genetic code assigns similar amino acids to similar codons. Two different rationales have been presented: first, mutation and translation error minimization,[3,10] and second, similar amino acids tend to directly interact with similar RNA sequences.[42] It has also been stated that instead of the actual codons, some of their derivatives, such as the anticodons[9,14] or codon-anticodon duplexes[2] were the original amino acid binding motifs. Recently, a new mechanism has been proposed for the association of amino acids with their codons and the origin of

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

Fig. 1. The common representation of the standard genetic code (codon families are shaded).

the genetic code.[6] It could explain two other long-known regularities of the genetic code. The first codon position seems to be correlated with amino acid biosynthetic pathways and to their evolution as evaluated by synthetic "primordial soup" experiments.[32,39] The second position is correlated with the hydrophatic properties of the amino acids. Codons with **U** as the second base code for the most hydrophobic amino acids and those having **A** as the second base are associated with the most hydrophilic amino acids.[32] Lagerkvist[16,17] observed that codon families (the amino acid of a codon family is determined by the first two nucleotides of a codon alone) have a much higher probability to appear in the left part of the common illustration of the genetic code (Fig. 1).

Recently, it was found that special purine-pyrimidine patterns of DNA binding sites facilitate recognition by restriction enzymes.[21] Here we also show that the genetic code is largely determined by purine-pyrimidine coding.

The following section introduces a new classification scheme of the genetic code based on the purine-pyrimidine coding, which demonstrates different codon symmetries that do not appear in the standard scheme. Section 3 presents an analysis of tRNA frequencies in 104 species (tRNA usage patterns), corresponding to the five symmetries in the new scheme. For each codon the number of genes coding for the tRNA with the complimentary anticodon (according to the Watson–Crick base paring) is counted. Note that this analysis differs from the codon adaptation index,[12,28] which additionally takes cytoplasmic tRNA concentration into account. The observed features of tRNA usage allow us to extend our earlier speculations concerning the evolution of the genetic code.[37]

2. The New Classification Scheme of the Genetic Code

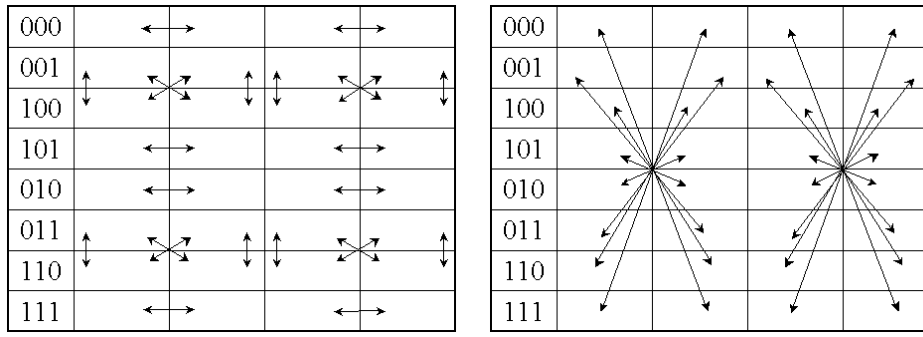
In contrast to the common representation of the genetic code our scheme is based on a binary encoding of the four bases **A**, **G**, **U**, **C**. There are three possibilities of a binary base coding,[41] according to:

(i) weak-strong bases (**A**,**U** = 1; **G**,**C** = 0),
(ii) keto- and amino bases (**G**,**U** = 1; **A**,**C** = 0), and
(iii) purines and pyrimidines (**A**,**G** = 1; **C**,**U** = 0).

In such a simplified code eight different binary triplets exist: $000, 001, \ldots, 111$. Each of these binary triplets represents eight different codons, e.g. in our purine-pyrimidine coding scheme 000 stands for $\mathbf{CCC}, \mathbf{CCU}, \ldots, \mathbf{UUU}$. 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 codon position it only matters if it is a purine or a pyrimidine. Interestingly, the only two exceptions are the start (**AUG**) and stop (**UGA**) codons. Given the purine-pyrimidine coding, there are two possibilities to sort the first two bases per row: one can use either of the remaining two binary codings, according to the weak and strong bases or according to keto- and aminobases, as a sort criterion inside the rows. We have

	Strong *6 hydrogen bonds*	**Mixed** *5 hydrogen bonds*	**Mixed** *5 hydrogen bonds*	**Weak** *4 hydrogen bonds*
000	*Pro* CC (C/U) Proline	*Ser* UC (C/U) Serine	*Leu* CU (C/U) Leucine	*Phe* UU (C/U) Phenylalanine
001	*Pro* CC (A/G) Proline	*Ser* UC (A/G) Serine	*Leu* CU (A/G) Leucine	*Leu* UU (A/G) Leucine
100	*Ala* GC (C/U) Alanine	*Thr* AC (C/U) Threonine	*Val* GU (C/U) Valine	*Ile* AU (C/U) Isoleucine
101	*Ala* GC (A/G) Alanine	*Thr* AC (A/G) Threonine	*Val* GU (A/G) Valine	*Ile/Met* AU (A/G) Isoleucine/Methionine
010	*Arg* CG (C/U) Arginine	*Cys* UG (C/U) Cysteine	*His* CA (C/U) Histidine	*Tyr* UA (C/U) Tyrosine
011	*Arg* CG (A/G) Arginine	*Stop/Trp* UG (A/G) Tryptophan	*Gln* CA (A/G) Glutamine	*Stop* UA (A/G)
110	*Gly* GG (C/U) Glycine	*Ser* AG (C/U) Serine	*Asp* GA (C/U) Aspartic acid	*Asn* AA (C/U) Asparagine
111	*Gly* GG (A/G) Glycine	*Arg* AG (A/G) Arginine	*Glu* GA (A/G) Glutamic acid	*Lys* AA (A/G) Lysine

Fig. 2. The purine(1)-pyrimidine(0) classification scheme of the genetic code. The third base is given in parenthesis. Shaded regions show codon families. The dashed horizontal line marks the symmetry axis for codon-anticodon symmetry and the dashed vertical line the mirror symmetry of purine (**G** ↔ **A**)-pyrimidine (**C** ↔ **U**) exchange. The point in the center indicates the point symmetry corresponding to keto-aminobase exchanges (**G** ↔ **U**, **A** ↔ **C**).



(a) (b)

Fig. 3. (a) The codon — reverse codon pattern, and (b) the sense — antisense codon pattern, in the purine — pyrimidine scheme of the genetic code. For instance, codon **GAU** is reverse to the codon **UAG** and codon **CGU** is the antisense codon of **ACG**. The codon symmetries are indicated by arrows.

chosen the weak-strong splitting to sort rows, because only this reveals the following regularities of the genetic code. First, all the codon families group together, i.e. they are not scattered all over the table. Secondly, the codon strength classification[16,17] directly corresponds to the columns in our scheme (Fig. 2). Thus, in the first column the first two bases form complementary base pairs with six hydrogen bonds (strong codons), in the second and third column with five (mixed codons), and in the fourth column with just four hydrogen bonds (weak codons).

In addition to its simplicity the new scheme uniquely shows five codon symmetries (Figs. 2 and 3) that are not obvious in other representations of the genetic code. Figure 2 reveals that the recently proposed family-nonfamily symmetry operation,[13] exchanging the amino bases (**A** $\leftrightarrow$ **C**) and the keto bases (**G** $\leftrightarrow$ **T**), corresponds to the point symmetry in our scheme. Moreover, the horizontal mirror symmetry corresponds to the codon-anticodon symmetry (weak (**A** $\leftrightarrow$ **U**) and strong base (**G** $\leftrightarrow$ **C**) exchanges) and the vertical mirror symmetry represents the purine-pyrimidine exchange symmetry (**G** $\leftrightarrow$ **A**, **C** $\leftrightarrow$ **U**). Figure 3(a) shows the symmetric codon-reverse codon pattern and Fig. 3(b) the sense-antisense codon pattern. Note that the last four patterns cannot be seen in the usual presentation of the genetic code (Fig. 1).

In our recently presented new classification scheme of the genetic code[37] there was one ambiguity left concerning the amino acid arrangement: the order of the second and third column was arbitrary. We now present four reasons for choosing the column order (mixed codons) as shown in Fig. 2:

(i) The codon-reverse codon symmetry, and
(ii) the sense-antisense symmetry are revealed only by the chosen order.
(iii) In each quadrant of the scheme the second position of the corresponding triplets is the same.

(iv) Strongly conserved groups of amino acids[34] are subsets of exactly one quadrant, e.g. the amino acids *M*, *I*, *L*, *V* belong to the upper right block in the table. The other conserved strong groups belonging to one block are *MILF*, *STA*, *NEQK*, *NHQK*, *NDEQ*, *HY*. The only exceptions are *QHRK* (*R* (*Arg*) is in another quadrant) and *FYW* (in three quadrants). In other words, reverse codon pairs tend to code for evolutionary similar amino acids, and each quadrant is enriched for amino acids with similar biochemical properties.

The new scheme of the genetic code has now its optimal form (Fig. 2). It shows five different triplet symmetries, including two additional symmetries that could not be seen in our first version of the scheme.[37]

3. Patterns of tRNA Usage

We studied tRNA usage of 16 archaea, 81 bacteria and 7 eucaryotes, using all information from the public database Genomic tRNA Compilation.[29] Different tables corresponding to the identified codon symmetries were composed, each containing all codons together with their symmetric codons. Table 1 shows the tRNA usage of all organisms, corresponding to the codon-reverse codon symmetry. Rows are sorted by the number of tRNA genes for a given anticodon (highest priority archaea, second priority bacteria).

Table 1. Codon-reverse codon pairs and the corresponding number of tRNA genes.

Amino Acid Pairs	Codon Pairs	Anticodon Pairs	Number of tRNA Genes		
			archaea(16)	bacteria(81)	eucaryotes(7)
Cys	TGT	ACA	0	0	0
Phe	TTT	AAA	0	0	0
Tyr	TAT	ATA	0	0	1
Ser	TCT	AGA	0	0	28
Ile	ATA	TAT	0	5	16
Asp ↔ Stop	GAT ↔ TAG	ATC ↔ CTA	0 ↔ 0	0 ↔ 0	0 ↔ 0
Ser ↔ Stop	AGT ↔ TGA	ACT ↔ TCA	0 ↔ 0	0 ↔ 0	0 ↔ 0
Asn ↔ Stop	AAT ↔ TAA	ATT ↔ TTA	0 ↔ 0	0 ↔ 0	1↔ 0
Val ↔ Leu	GTT ↔ TTG	AAC ↔ CAA	0 ↔ 12	0 ↔ 93	18 ↔ 29
Ala ↔ Ser	GCT ↔ TCG	AGC ↔ CGA	0 ↔ 12	1 ↔ 64	36 ↔ 13
Gly ↔ Trp	GGT ↔ TGG	ACC ↔ CCA	0 ↔ 14	0 ↔ 111	0 ↔ 19
Pro ↔ Ser	CCT ↔ TCC	AGG ↔ GGA	0 ↔ 16	0 ↔ 99	16 ↔ 1
Ile ↔ Leu	ATT ↔ TTA	AAT ↔ TAA	0 ↔ 16	0 ↔ 107	16 ↔ 21
His ↔ Tyr	CAT ↔ TAC	ATG ↔ GTA	0 ↔ 16	0 ↔ 118	0 ↔ 55
Thr ↔ Ser	ACT ↔ TCA	AGT ↔ TGA	0 ↔ 16	1 ↔ 114	18 ↔ 21
Leu ↔ Phe	CTT ↔ TTC	AAG ↔ GAA	0 ↔ 16	8 ↔ 113	20 ↔ 25
Arg ↔ Cys	CGT ↔ TGC	ACG ↔ GCA	0 ↔ 16	114 ↔104	18 ↔ 46
Ala	GCG	CGC	12	28	15
Glu	GAG	CTC	12	30	19

Table 1. (*Continued*)

Amino Acid Pairs	Codon Pairs	Anticodon Pairs	Number of tRNA Genes		
			archaea(16)	bacteria(81)	eucaryotes(7)
Gly	GGG	CCC	12	46	13
Arg	CGC	GCG	14	12	0
Val	GTG	CAC	14	31	18
Glu ↔ Lys	GAA ↔ AAG	TTC ↔ CTT	14 ↔ 12	122 ↔ 59	21 ↔ 29
Pro	CCC	GGG	15	66	0
Thr	ACA	TGT	15	115	19
Val ↔ Leu	GTC ↔ CTG	GAC ↔ CAG	15 ↔ 12	92 ↔ 76	0 ↔ 10
Leu	CTC	GAG	16	90	2
His	CAC	GTG	16	106	16
Lys	AAA	TTT	16	120	25
Arg	AGA	TCT	16	121	21
Ala ↔ Pro	GCC ↔ CCG	GGC ↔ CGG	16 ↔ 12	73 ↔ 49	0 ↔ 11
Gly ↔ Arg	GGA ↔ AGG	TCC ↔ CCT	16 ↔ 12	113 ↔ 81	18 ↔ 14
Ala ↔ Thr	GCA ↔ ACG	TGC ↔ CGT	16 ↔ 12	118 ↔ 66	17 ↔ 18
Asp ↔ Gln	GAC ↔ CAG	GTC ↔ CTG	16 ↔ 13	121 ↔ 43	21 ↔ 26
Ser ↔ Arg	AGC ↔ CGA	GCT ↔ TCG	16 ↔ 14	105 ↔ 30	21 ↔ 20
Pro ↔ Thr	CCA ↔ ACC	TGG ↔ GGT	16 ↔ 15	109 ↔ 107	29 ↔ 0
Asn ↔ Gln	AAC ↔ CAA	GTT ↔ TTG	16 ↔ 17	132 ↔ 115	41 ↔ 23
Gly ↔ Arg	GGC ↔ CGG	GCC ↔ CCG	17 ↔ 11	116 ↔ 74	23 ↔ 8
Ile ↔ Leu	ATC ↔ CTA	GAT ↔ TAG	17 ↔ 17	106 ↔ 107	1 ↔ 12
Met ↔ Val	ATG ↔ GTA	CAT ↔ TAC	45 ↔ 16	285 ↔ 118	33 ↔ 12

The order in Table 1 shows best the following main observations. The first interesting pattern of tRNA usage refers to reverse STOP codons.[a] Of course, no species has a tRNA with an anticodon complementary to any termination codon. Intriguingly, there is also no tRNA with a specific anticodon for a reverse STOP codon. The only exception is *H. sapiens* with one tRNA^{Asn} with the anticodon **ATT**. The lack of specific tRNAs does not imply that no tRNA exists which can recognize reverse STOP codons. For instance, using base pairing allowed by Crick's wobble rules,[7] a tRNA with the **GTT** anticodon can recognize the reverse STOP codon **AAT**.

The second striking pattern in tRNA usage is the significant suppression of tRNAs with **A** at the first anticodon position (Table 1). **A**** anticodons are fully excluded in archaea. In bacteria and eucaryotes there are some exceptions, but it can be observed that **AY*** anticodons do not apear in any species.

Another observation concerning tRNA usage is the significant suppression of **A***A self-reverse anticodons. No archaea and no bacteria species has a tRNA with such an anticodon.

4. The Reverse Recognition Conjecture

In this section we present a hypothesis that consistently explains the observed suppression of anticodons for reverse STOP codons. We conjecture that the

[a]tRNA genes specifically recognizing initiation codon (*Met*) are significantly overrepresented.

observed tRNA usage patterns reflect important features of the ancient translation machinery. Maybe, in the early days of translation, pre-tRNAs were able to recognize codons in both directions (Fig. 4). In order to guarantee termination (i.e. to avoid incorrect elongation) the reverse stop codons had (and have) no own tRNA.

In agreement with others,[6,24,40] we hypothesized in our previous work that the contemporary triplet code developed from an ancient doublet code.[37] However, in order to avoid the frameshift problem one has to assume a triplet reading frame also in doublet coding times.[37] In agreement, the triplet reading frame was recently substantiated because unpaired RNA loops with seven and eight nucleotides are the most stable ones.[40] Nevertheless, one still wonders about such a wasting of information, where the third base would not carry any information at all. We speculate that in the early days of translation pre-tRNAs could fit in two opposite directions to the corresponding mRNA (Fig. 4). This would resolve the wasting problem: if a codon could be recognized in both directions all bases would carry information, although in a given codon-anticodon pairing only two bases are analyzed. Three different facts support our speculation. First, ancient pre-tRNAs presumably only consisted of the anticodon loop, lacking the D- and T-loops.[27] Such pre-tRNAs would have been (almost) symmetrical and could thus bind in two directions. If the reverse recognition model is correct, the resulting polypeptide should be relatively independent of the pre-tRNA binding direction. This is supported by the special role of the central triplet base.[38,40] It is well-known that the second base has the strongest interaction with the bases of 16S RNA (the universally conserved and essential bases A1492, A1493, and G530[22,30]). Moreover, the middle base of the anticodon has particularly strong interactions with the correct aminoacyl-tRNA

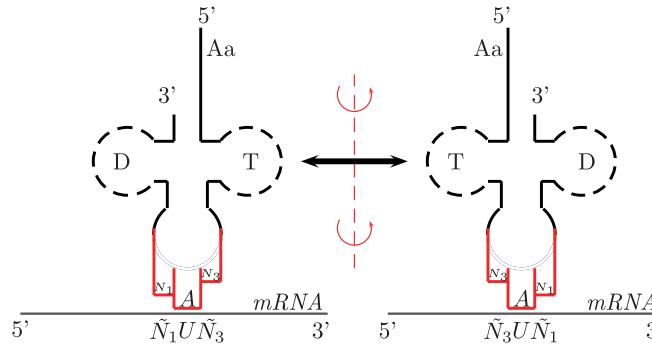


Fig. 4. Possible ancient codon-anticodon recognition with doublet coding and reverse pairing. The figure shows schematically the binding of the same pre-tRNA in normal and reverse direction to two different (reverse) codons. The proximity of the anticodon "fingers" to the codon represents the accuracy by which the respective nucleotides are recognized.[11] Additional figures can be found at www.imb-jena.de/~sweta/genetic_code_and_evolution/early_evolution_translation.html.

synthetase during amino acid attachment.[20] The second base is exceptional also in another respect: it is correlated to the main physical property of amino acids, the hydrophobicity.[32] The third fact supporting our "reverse recognition conjecture" is the above observation that reverse codon pairs generally encode evolutionary similar amino acids.[34] We suppose that this observation is a relict from old "reverse recognition times", where the reverse recognition should have a minimal effect on the resulting polypeptide.

It was speculated that the translation machinery of the last universal common ancestor (LUCA) is most similar to that of archaea,[35] so we expect that tRNA usage patterns in archaea reflect ancient translation. What could be the reason for the forbidden **A**** anticodon-tRNAs? Three different explanations can be given. First, sometimes **A** (and the simple derivative inosine **I**) at the third codon position misleadingly pair with the first anticodon position.[23] In order to prevent such a mistranslation **A**** anticodon-tRNAs could have been forbidden. A second possible explanation is the strong preference of **G** (instead of **A**) at the first anticodon position in order to recognize the corresponding pyrimidines. In the recently proposed evolution of wobble rules[35] **G** (at the first anticodon position) always recognizes **U** and **C**, in all discussed evolution stages. **A**** anticodons, in contrast, could not recognize any base in the early stages.[35] This would also be in agreement with earlier speculations about a binary coding scheme with just one purine and one pyrimidine.[26,37] Interestingly, Table 1 reveals that nearly all 16 **G**** anticodons have corresponding tRNAs in all species. The third explanation is based on an observation concerning initiation codons. Translation in eukaryotes can be initiated from codons other than **AUG**. A well-documented case (including direct protein sequencing) is the **GUG** start of a ribosomal P2A protein of the fungus *Candida albicans*.[1] Other examples can be found in the NCBI taxonomy database.[4,36] Interestingly, all nine different initiation codons have **U** at the second position (**AUG** (standard), **AUA**, **AUU**, **AUC**, **GUG**, **GUA**, **UUG**, **UUA**, **CUG**). Maybe, in earlier times ***U*** codons generally could initiate translation, starting with ***A*** anticodons. We speculate that the forbidden **A**** anticodons should protect the transcript against wrong translation initiation which would lead to a frameshift.

Moreover, we note that the three termination codons in the standard genetic code all have a purine at the second position. The alternative termination codons in non-standard codes are also ***R*** codons (**AGA** and **AGG** in vertebrate mitochondria[4,36]). Maybe, in "binary coding times"[37] ***Y*** codons could initiate translation, whereas ***R*** codons could terminate translation. This additionally supports our speculations of possible reverse codon recognition. Up to now the possibility of reverse recognition provides the only explanation that consistently integrates all of our observations. This model might be used as a plausible framework onto which research into translation evolution may be devised.

5. Discussion

We presented a new classification scheme of the genetic code. It has now its optimal form, no ambiguities in codon order are left. The scheme clearly shows all five different codon symmetries. We also studied the occurrence of tRNA genes in archaea, bacteria and eukaryotic species. tRNA usage, ordered according to the codon-reverse codon symmetry, shows three interesting facts:

(i) there are no specific tRNAs for reverse STOP codons;
(ii) **A**** anticodons are significantly repressed, **G**** anticodons are significantly utilized; and
(iii) **A***A** self-reverse anticodons are totally excluded in archaea and bacteria.

This led us to extend our earlier speculations on doublet coding.[37] We conjecture that in earlier times codon recognition could also have been carried out in the reverse direction with first recognizing the second base (Fig. 4). Our hypothesis is related to the recently proposed evolution scheme of the genetic code, where it was suggested that "... triplet codons gradually evolved from two types of ambiguous doublet codons, those in which the first two bases of each three-base window were read ('prefix' codons) and those in which the last two bases of each window were read ('suffix' codons).[40]" In contrast to this model our reverse recognition conjecture implies a parallel-stranded duplex structure of the two relevant codon-anticodon base pairs. Although such parallel structures are difficult to find in natural nucleic acids they have been observed in DNA[5] and mRNA,[33] and a corresponding crystal structure has been reported.[31] However, because RNA is unstable and difficult to synthesize, it was proposed that the first genetic material used a simpler backbone than ribose.[15] For such molecules the pairing strand direction is probably not as constraint as in DNA or RNA.

Of course, all discussed tRNA usage patterns depend on the completeness of the known tRNAs. If a significant number of tRNAs is still unknown, this might modify these patterns. However, the fact that tRNAs are systematically searched by robust computer algorithms[18,19,25] makes it very unlikely that such a significant number of tRNAs will be found in the future.

Acknowledgments

We thank R. Brockmann for critically reading the manuscript and an anonymous referee for valuable comments. This work was supported by Grant no. 0312704E of the Bundesministerium fuer Bildung und Forschung, Germany.

References

1. Abramczyk D, Tchorzewski M, Grankowski N, Non-AUG translation initiation of mRNA encoding acidic ribosomal P2A protein in Candida albicans, *Yeast* **20**: 1045–1052, 2003.

2. Alberti S, The origin of the genetic code and protein synthesis, *J Mol Evol* **45**:352–358, 1997.
3. Ardell DH, Sella G, No accident: genetic codes freeze in error-correcting patterns of the standard genetic code, *Phil Trans R Soc Lond B* **357**:1625–1642, 2002.
4. Benson DA, Karsch-Mizrachi I, Lipman DJ *et al.*, *GenBank Nucleic Acids Res* **28**: 15–18, 2000.
5. Borisova OF, Shchyolkina AK, Chernov BK, Tchurikov NA, Relative stability of AT and GC pairs in parallel DNA duplex formed by a natural sequence, *FEBS Lett* **322**:304–306, 1993.
6. Copley SD, Smith E, Morowitz HJ, A mechanism for the association of amino acids with their codons and the origin of the genetic code, *Proc Natl Acad Sci USA* **102**:4442–4447, 2005.
7. Crick FHC, Codon-anticodon pairing: the wobble hypothesis, *J Mol Biol* **19**:548–555, 1966.
8. Crick FHC, The origin of the genetic code, *J Mol Biol* **38**:367–379, 1968.
9. Dunnill P, Triplet nucleotide-amino-acid pairing; a stereochemical basis for the division between protein and non-protein amino-acids, *Nature* **210**:1265–1267, 1966.
10. Freeland SJ, Knight RD, Landweber LF, Hurst LD, Early fixation of an optimal genetic code, *Mol Biol Evol* **17**:511–518, 2000.
11. Freeland SJ, Hurst LD, The genetic code is one in a million, *J Mol Evol* **47**:238–248, 1998.
12. Friberg M, von Rohr P, Gonnet G, Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in Saccharomyces cerevisiae, *Yeast* **21**:1083–1093, 2004.
13. Halitsky D, 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, 2003.
14. Jungck JR, The genetic code as a periodic table, *J Mol Evol* **11**:211–224, 1978.
15. Knight RD, Landweber LF, The early evolution of the genetic code, *Cell* **101**: 569–572, 2000.
16. Lagerkvist U, "Two out of three": An alternative method for codon reading, *Proc Natl Acad Sci USA* **75**:1759–1762, 1978.
17. Lagerkvist U, Unorthodox codon reading and the evolution of the genetic code, *Cell* **23**:305–306, 1981.
18. Laslett D, Canback B, ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences, *Nucleic Acids Res* **32**:11–16, 2004.
19. Lowe TM, Eddy SR, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res* **25**:955–964, 1997; http://lowelab.ucsc.edu/GtRNAdb/
20. McClain WH, Schneider J, Bhattacharya S, Gabriel K, The importance of tRNA backbone-mediated interactions with synthetase for aminoacylation, *Proc Natl Acad Sci USA* **95**:460–465, 1998.
21. Nikolajewa S, Beyer A, Friedel M, *et al.*, Common patterns in type II restriction enzyme binding sites, *Nucleic Acids Res* **33**:2726–2733, 2005.
22. Ogle JM, Brodersen DE, Clemons WM Jr *et al.*, Recognition of cognate transfer RNA by the 30S ribosomal subunit, *Science* **292**:897–902, 2001.
23. Osawa S, Jukes TH, Watanabe K, Muto A, Recent evidence for evolution of the genetic code, *Microbiol Rev* **56**:229–264, 1992.

24. Patel A, The triplet genetic code had a doublet predecessor, *J Theor Biol* **233**: 527–532, 2005.
25. Pavesi A, Conterio F, Bolchi A *et al.*, Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions, *Nucleic Acids Res* **11**:1247–1256, 1994.
26. Reader JS, Joyce GF, A ribozyme composed of only two different nucleotides, *Nature* **420**:841–844, 2002.
27. Rodin S, Ohno S, Rodin A, Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors?, *Proc Natl Acad Sci USA* **90**:4723–4727, 1993.
28. Sharp PM, Li WH, The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res* **15**:1281–1295, 1987.
29. Sprinzl M, Vassilenko KS, Emmerich J, Bauer F, Compilation of tRNA sequences and sequences of tRNA genes. Last update 2003. www.uni-bayreuth.de/departments/biochemie/trna/.
30. Stahl G, McCarty GP, Farabaugh PJ, Ribosome structure: revisiting the connection between translational accuracy and unconventional decoding, *Trends Biochem Sci* **27**:178–183, 2002.
31. Sunami T, Kondo J, Kobuna T *et al.*, Crystal structure of d(GCGAAAGCT) containing a parallel-stranded duplex with homo base pairs and an anti-parallel duplex with Watson–Crick base pairs, *Nucleic Acids Res* **30**:5253–5260, 2002.
32. Taylor FJ, Coates D, The code within the codons, *BioSystems* **22**:177–187, 1989.
33. Tchurikov NA, Chistyakova LG, Zavilgelsky GB, *et al.*, Gene-specific silencing by expression of parallel complementary RNA in Escherichia coli, *J Biol Chem* **275**:26523–26529, 2000.
34. Thompson JD, Higgins DG, Gibson TJ, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**:4673–4680, 1994.
35. Tong KL, Wong JT, Anticodon and wobble evolution, *Gene* **333**:169–177, 2004.
36. Wheeler DL, Chappey C, Lash AE *et al.*, Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res* **28**:10–14, 2000.
37. Wilhelm T, Nikolajewa SL, A new classification scheme of the genetic code, *J Mol Evol* **59**:598–605, 2004.
38. Woese CR, *The Genetic Code: The Molecular Basis for Genetic Expression*, Harper & Row, New York, 1967.
39. Wong JT, A co-evolution theory of the genetic code, *Proc Natl Acad Sci USA* **72**: 1909–1912, 1975.
40. Wu HL, Bagby S, van den Elsen JM, Evolution of the genetic triplet code via two types of doublet codons, *J Mol Evol* **61**:54–64, 2005.
41. Yagil G, The over-representation of binary DNA tracts in seven sequenced chromosomes, *BMC GENOMICS* **5**(1):19, 2005.
42. Yarus M, Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin, *J Mol Evol* **47**:109–117, 1998.

Swetlana Nikolajewa received her Bachelor and Master degrees in Applied Mathematics from Rostov State University (RSU), Russia, in 1997 and 1999, respectively. From 2003 she is with the Institute of Molecular Biotechnology, Jena, Germany, where she does doctoral study at the Theoretical Systems Biology Group.

Maik Friedel is a student of Bioinformatics at the Friedrich-Schiller University, Jena, Germany. He is doing his diploma thesis at the Theoretical Systems Biology Group, Institute of Molecular Biotechnology, Jena, Germany.

Andreas Beyer received his degree in Applied Systems Science and his Ph.D. from the University of Osnabrück, Germany, in 1999 and 2002, respectively. Since 2002 he is working in the group of Thomas Wilhelm at the Institute of Molecular Biotechnology, Jena, Germany as a postdoctoral fellow.

Thomas Wilhelm received his Diploma degree in Biophysics, and his Ph.D. in Theoretical Biophysics from Humboldt-University Berlin, Germany, in 1992 and 1997, respectively. He is the head of the Theoretical Systems Biology Group at the Institute of Molecular Biotechnology, Jena, Germany.