Heredity (2008) 100, 341–355

www.nature.com/hdy

REVIEW

On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases

SN Rodin[1] and AS Rodin[2]
[1]*Theoretical Biology Department, Beckman Research Institute of the City of Hope, Duarte, CA, USA and* [2]*Human Genetics Center, School of Public Health, University of Texas, Houston, TX, USA*

If the table of the genetic code is rearranged to put complementary codons face-to-face, it becomes apparent that the code displays latent mirror symmetry with respect to two sterically different modes of tRNA recognition. These modes involve distinct classes of aminoacyl-tRNA synthetases (aaRSs I and II) with recognition from the minor or major groove sides of the acceptor stem, respectively. We analyze the anticodon pairs complementary to the face-to-face codon couplets. Taking into account the invariant nucleotides on either side (5′ and 3′), we consider the risk of anticodon confusion and subsequent erroneous aminoacylation in the ancestral coding system. This logic leads to the conclusion that ribozymic precursors of tRNA synthetases had the same two complementary modes of tRNA aminoacylation. This surprising case of molecular mimicry (1) shows a key potential selective advantage arising from the partitioning of aaRSs into two classes, (2) is consistent with the hypothesis that the two aaRS classes were originally encoded by the complementary strands of the same primordial gene and (3) provides a 'missing link' between the classic genetic code, embodied in the anticodon, and the second, or RNA operational, code that is embodied mostly in the acceptor stem and is directly responsible for proper tRNA aminoacylation.
Heredity (2008) **100,** 341–355; doi:10.1038/sj.hdy.6801086; published online 5 March 2008

Keywords: complementary symmetry; RNA world; genetic code; tRNA aminoacylation

Introduction

The idea of the genetic code was one of the most important and captivating implications of the discovery of the double-helical structure of DNA (Yanofsky, 2007). It has transpired that the genetic code comprises a (nearly) universal assignment of nucleotide triplets (codons) to corresponding amino acids (Table 1). The deciphering of the code piqued the interest of the scientific community in a much more challenging problem—the origin(s) of the code.

An 'origin-of-code' scenario, which explains the existence of genetic coding material and the associated amino acids for its expression is worth mentioning if, and only if, it evades the proverbial 'chicken-or-egg' conundrum. The hypothesis of a direct stereochemical affinity ('key–lock') between an amino acid and a codon (Woese, 1965) meets this criterion, whereas the hypothesis of code 'adaptor' molecules (Crick, 1958) cannot get around the chicken-or-egg conundrum. The essence of this hypothesis is that an adaptor is able to recognize simultaneously an amino acid and a cognate codon or codons, but then the burden of explanation is merely passed onto this homunculus. In principle, the existence of such adaptors does not rule out at least weak direct stereochemical recognition between an amino acid and its codon(s); however, it certainly makes this affinity unnecessary (Crick, 1958, 1968).

The discovery of tRNAs confirmed Crick's hypothesis, but raised the following problem: tRNAs implement the code *via* complementary replica of the codon, the anticodon. However, the anticodon is located in the middle of the tRNA cloverleaf, at the maximum possible distance from the CCA end, where the cognate amino acid will be attached (Figure 1a). Because of this separation, tRNA molecules cannot self aminoacylate; instead, there are 20 amino acid-specific aminoacyl-tRNA synthetases (aaRSs) that perform this function. Thus, it is these aaRSs that actually define the familiar matrix of the genetic code, by linking the specific amino acids and tRNAs with the corresponding anticodons.

The aaRSs are proteins. Moreover, they are pleiotropic proteins that are directly involved in the synthesis of all other proteins, and this means that the aaRSs represent the chicken-or-egg paradox at its most puzzling. The problem is further aggravated by the distinctive partitioning of the protein aaRSs (p-aaRSs) into two classes (I and II) of 10 members each (Eriani *et al*., 1990). Despite performing exactly the same function, tRNA aminoacylation, these two enzyme classes share no homology—either in primary sequence or at higher 2D and 3D levels (Eriani *et al*., 1990). However, their modes of tRNA

Correspondence: Dr SN Rodin, Theoretical Biology Department, Beckman Research Institute of the City of Hope, 1500 East Duarte Road, Duarte, CA 91010-3000, USA.
E-mail: srodin@coh.org
Received 30 May 2007; revised 17 September 2007; accepted 12 November 2007; published online 5 March 2008

Table 1 The conventional representation of the genetic code with indicated assignments of p-aaRSs to class I (yellow) and class II (blue)

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

I class aaRS
II class aaRS

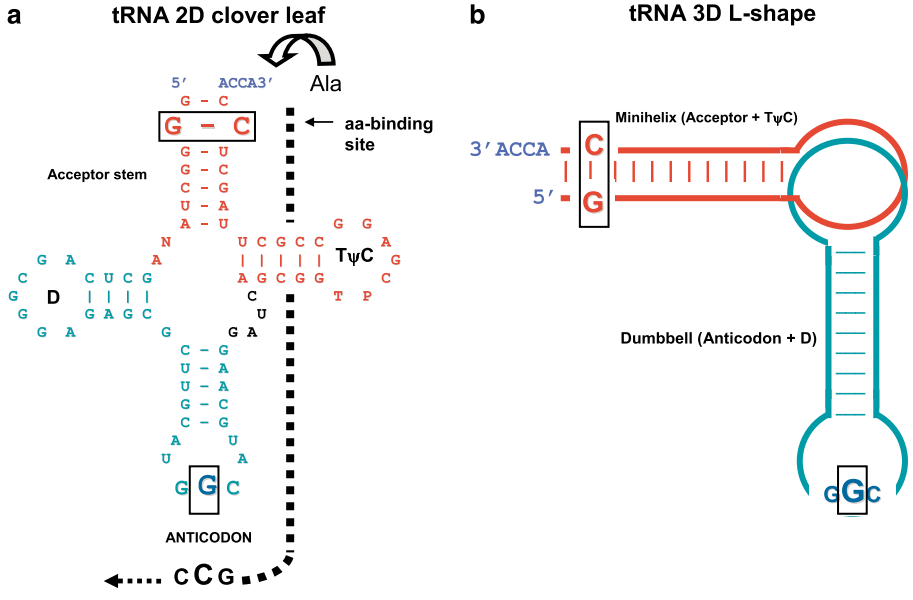


Figure 1 tRNA cloverleaf (**a**) and L-shaped 3D structure (**b**). tRNA molecules consist of two halves, the minihelix (acceptor stem plus TψC arm) on top, and the 'dumbbell' (anticodon arm plus D arm) at the bottom. An *Escherichia coli* $tRNA^{Ala}$ (with GGC anticodon) is shown, with the characteristic 3G:U72 'wobbling' complementary base pair that determines the identity of the Ala tRNAs across all species (Hou and Schimmel, 1988). The second base of the anticodon and the second base pair of the acceptor (showing dual complementarity—see text) are enlarged and boxed. Dotted line on the right of Figure 1a denotes the hypothetical aminoacylating ribozyme (adapted from Rodin and Rodin, 2007).

recognition almost perfectly complement each other in a symmetric, mirror-like fashion; class I recognition occurs from the minor groove side of the acceptor stem, whereas class II recognition occurs from the major groove side (Rould *et al*., 1989; Cusack *et al*., 1990; Eriani *et al*., 1990; Ruff *et al*., 1991; Carter, 1993). We believe that the two different types of groove recognition imply particular distributions of codons and amino acids, and these distributions, together with certain additional considerations, unambiguously point to the essential role that the primordial double-stranded (sense–antisense) translation played in the formation of the very core of the genetic code.

Accordingly, in this review article, we will concentrate on the properties of tRNAs and aaRSs that are associated with complementary codons. In particular, the

rearrangement of the genetic code that puts complementary codons face-to-face with each other discloses an intriguing, highly nonrandom pattern. This pattern (1) shows a key potential selective advantage arising from the partitioning of aaRSs into two classes, (2) supports the hypothesis that the two classes of aaRSs were originally encoded by the complementary strands of the same ancestral gene (Rodin and Ohno, 1995), and (3) might provide a 'missing link' between the classic genetic code embodied in the anticodons and the operational code that is embodied mostly in the acceptor stem and is directly responsible for aminoacylation (Schimmel *et al*., 1993).

The problem of two codes

The 3D structure of tRNAs is formed by two domains (Figure 1b): the top domain (minihelix), to which the cognate aaRS attaches a specific amino acid at the 3′CCA end, and the bottom domain (dumbbell), with the anticodon positioned in the very center, which determines amino acid specificity. The minihelix and dumbbell comprise the characteristic L shape, mainly due to additional base pairings between the D and TψC loops. To a striking extent, these two domains appear to be functionally independent of each other. The tRNAs of at least ten amino acids can be charged successfully with the correct amino acids by the cognate p-aaRSs when truncated to a minihelix, or even a smaller piece that contains the 3′CCA end (reviewed by Schimmel and Beebe, 2006). Reciprocally, the truncated aaRSs (in extreme cases, a truncated aaRS is unable even to reach the anticodon) maintained the same tRNA-aa specificity (Schimmel and Beebe, 2006).

The RNA operational code

This striking anticodon-independent, yet amino acid-specific aminoacylation of tRNAs led to the idea of there being a second, RNA operational, code that is localized mainly in the acceptor stem of the tRNAs (in the vicinity of the amino acid attachment site) and is recognized by the corresponding module of p-aaRSs (see also de Duve 1988; Schimmel *et al*., 1993).

To an unexpected extent, the operational code determines which aaRS is cognate for a given tRNA (Schimmel *et al*., 1993), and it is the operational code again that brings the classic code, which is associated with anticodons, into action. The question then arises: are the two codes independent by origin? Of great significance in this regard is the observation that the replication initiation sites of RNA genomes resemble the minihelix with the 3′CCA terminus (Weiner and Maizels, 1987, 1999). It was, therefore, proposed (Schimmel *et al*., 1993) that:

1. a mini- (or even micro-) helix tRNA precursor might have had the ability to interact specifically with amino acids, probably long before it merged with the anticodon-containing dumbbell (Schimmel *et al*., 1993).
2. although the present-day operational code is implemented by the protein aaRSs, the original precursor of the operational code (presumably implemented by the ribozymic aaRSs) might have been older than the classic genetic code, and, if so
3. the classic code (Table 1) might have been a 'frozen accident' of sorts (Crick, 1968).

Yet, the codons-to-amino acids assignment does not seem to have been shaped by pure chance. Noticeably, similar triplets tend to encode similar amino acids (Table 1). And, although the frozen accident-based scenario does not entirely rule out the gradual selection-driven optimization of the genetic code along the 'similar codons for similar amino acids' lines (Crick, 1968), it is clear that direct stereochemical affinity between amino acids and cognate nucleotide triplets (anticodons and/or codons) would provide this agreement much more readily (Woese, 1965; Szathmary, 1993, 1999; Yarus, 1998). Furthermore, both Corey–Pauling–Koltun models of stereochemical C4N complexes of anticodons (noncovalently linked with the unpaired 73rd discriminator base) with appropriate amino acids (Shimizu, 1982) and aa-binding sites of RNA aptamers that were selected from random RNA pools during evolution *in vitro* (Yarus 1998; Caporaso *et al*., 2005; Yarus *et al*., 2005) show that this stereochemical affinity, although often weak, is quite real. Finally, it is only logical to suppose that the putative ribozymic precursors of synthetases, r-aaRSs, experienced precisely the same problem with the two codes.

Indeed, at first glance, the obvious advantage of r-aaRSs over their protein successors is the ability of the ribozymes to easily recognize the anticodon *via* a trivial complementary pairing. This anticodon-recognition site can be thought of as an anti-anticodon triplet—a codon-like triplet. However, for any r-aaRS to charge its cognate tRNAs with the correct amino acid, the r-aaRS must have possessed not only this anticodon-recognition site but also the specific aa-binding site. Note that the aa-binding site would have needed to be located close to the 3′ end of the tRNA (Figure 1), that is, again very far from the 'anti-anticodon' site. Therefore, it appears that if r-aaRSs did exist, they faced exactly the same 'remoteness' problem; in order to aminoacylate their cognate tRNAs, they would require catalysts of their own, that is, 'meta-r-aaRSs,' which, in turn, would inherit the same problem and require catalysts of their own, *ad infinitum*. Therefore, the advantages of direct recognition of anticodons by hypothetical r-aaRSs only readdress, rather than solve, the paradox.

This brings us to the only reasonable solution—a duplication of anticodon within the same tRNA molecule (Di Giulio, 1992), a duplication that actually means that these two, presently very different, codes (operational and classic) were originally one and the same (Rodin *et al*., 1996). And this, in turn, necessarily implies that the bottom (dumbbell) module of tRNA might have originated by duplication from the top (minihelix) module, or vice versa (Figure 1b); (Szathmary, 1999). The well-known internal sequence periodicity of tRNAs (Bloch *et al*., 1985) is consistent with this duplication model.

Concerted dual complementarity of second bases in two codes

The first three positions of the acceptor stem can be considered the best candidates for being the anticodon homolog, because they represent the major identity elements of tRNAs and they are located adjacent to the base-determinator and the 3′CCA site of amino acid

attachment. However, straightforward analysis failed to uncover any traces of homology in this case. We performed a different analysis based on the assumption that the earliest primitive translation might have been more strand symmetric, as compared to the later mechanism that translated the strands after they had differentiated into the sense/coding and antisense/non-coding strands. Accordingly, instead of looking at the individual tRNAs, we tested pairs of consensus tRNAs with complementary anticodons, and were richly rewarded: for the majority of these tRNA pairs, the complementarity of anticodons was accompanied by the complementarity of second bases in their acceptor stems (Figure 1; Rodin *et al*., 1996).

We interpreted this parallelism as a remnant of the common ancestry of two, classic and operational, codes (Rodin *et al*., 1996; Rodin and Ohno, 1997). However, only the central nucleotide of the putative anticodon's duplicate in the acceptor stem showed this concerted dual complementarity, so it remained speculative that the primordial operational code even possessed a triplet structure. In a general context, the question was how could the anticodon and the duplicate in the acceptor stem coevolve within the same tRNA molecule during the expansion of the genetic code?

We have re-examined the dual complementarity for ancestral tRNAs that were reconstructed from the updated compilation of more than 8000 tRNA gene sequences covering the three main kingdoms of organisms—eubacteria, archaebacteria and eukaryotes (Sprinzl and Vassilenko, 2005). Dual complementarity was detected for pairs of ancestral tRNAs with completely complementary anticodons and was not detected for pairs in which only the second bases of anticodons were complementary (Rodin and Rodin, 2006a). This surprising difference suggests that

1. the dual complementarity originated when the three-letter frame of translation had already been in full use.
2. Although coevolution of these two codes could have started with the duplicates of same trinucleotides, both classic and operational codes were originally highly ambiguous. Specifically, only the second bases could have actually encoded the groups of similar amino acids at the time when the first protein aaRSs began to replace their ribozymic precursors (for details, see Rodin and Rodin, 2006a).

The dual complementarity is consistent with the archaic in-frame translation of both strands—sense and antisense

Figure 2 illustrates this statement for a hypothetical primitive gene consisting of complementary GCC and GGC triplets that encode Ala and Gly, respectively. Ala and Gly were likely the first amino acids incorporated

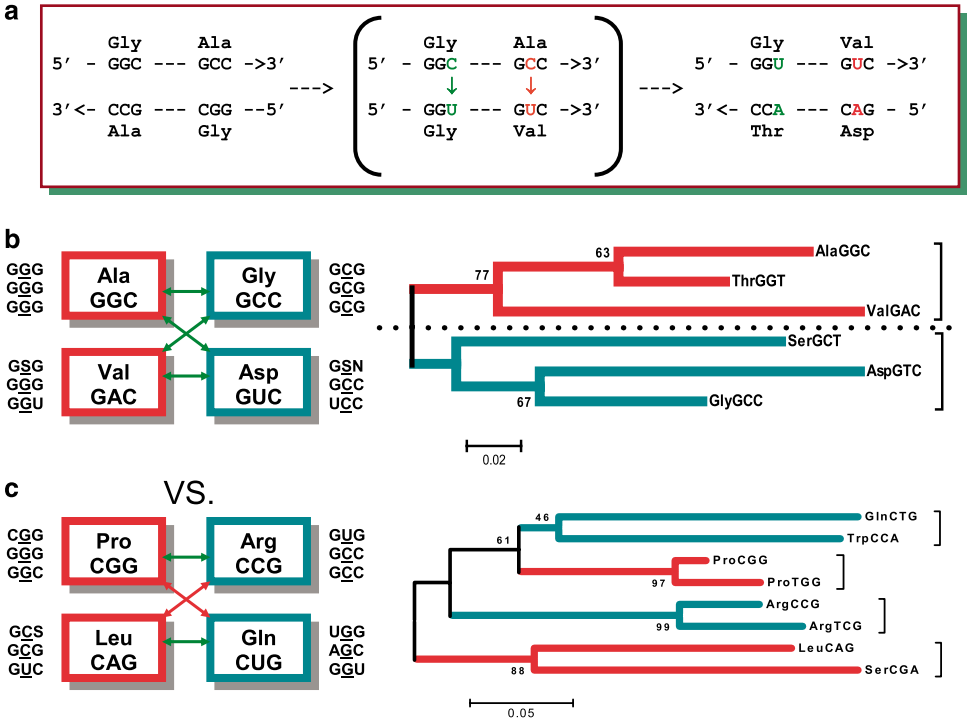


Figure 2 New codons could enter translation by complementary pairs. (**a**) The scheme illustrating how a C→U transition in one strand is accompanied by a G→A transition in the opposite strand. Missense mutations at the second codon position and conservative/silent mutations at the first/third codon positions are shown in red and green, respectively. (**b**) Shown in boxes on the left is the tetrad of complementary anticodons and cognate amino acids that (1) also have complementary second bases in the acceptor stems and (2) demonstrate such dual complementarity in all four canonical (G–C, A–U), wobbling G:U and A*C combinations of base pairing at the central position of anticodons (outlined by the green bidirectional arrows). Shown next to the boxes are the ancestral triplets at positions 1, 2 and 3 of the acceptor stem (with underlined second bases) for eubacteria, archaebacteria and eukaryotes (from top to bottom, respectively); S denotes either G or C. On the right is the Neighbor-Joining phylogenetic tree for ancestral tRNAs of archaebacteria generated by this tetrad of amino acids. The phylogenetic tree exhibits a distinct NRN vs *ИУИ* bilateral branching pattern. (**c**) In contrast to (**b**), an example of the tetrads of amino acids with complementary codons is shown that neither exhibits dual complementarity in wobbling combinations (outlined by the red bi-directional arrows) nor shows the bilateral NRN vs *ИУИ* branching pattern in the corresponding phylogenetic tree (adapted from Rodin and Rodin, 2006a, b).

into the genetic coding (Eigen and Schuster, 1979; Rodin and Ohno, 1997; Klipcan and Safro, 2004; Patel, 2005; Trifonov, 2005). Translation of both strands in the same frame suggests that at first the repertoire of codons expanded by complementary pairs rather than one-by-one. In Figure 2, the GCC→GUC (Ala→Val) transition in one strand is necessarily complemented by GGC→GAC (Gly→Asp) transition in the opposite strand (Figure 2a); accordingly, the evolving code gains a new pair of codons, GAC and GUC, for a new pair of amino acids, Val and Asp. Such a concerted recruitment of Val and Asp in translation would imply at least two duplication events for $tRNA^{Ala}$ and $tRNA^{Gly}$ genes, with the subsequent Val- and Asp-specific mutational 'tune-ups.' If the original pair of tRNAs with complementary anticodons, that is, $tRNA^{Ala}$ and $tRNA^{Gly}$, carried complementary second bases in the acceptor stem (and it is likely that they did (Figure 2b)), then there is a high probability that their duplicates with complementarily mutated anticodons, $tRNA^{Val}$ and $tRNA^{Asn}$, preserve this dual complementarity (Figures 2b and c) while gaining the specific identity elements for new amino acids (Val and Asp, in this case) elsewhere in their cognate tRNA molecules (for details, see Rodin and Rodin, 2006).

First pairs of complementarily encoded amino acids

Our belief was that pairs of tRNAs with G:U or A*C illegitimate pairings in their anticodons should also demonstrate significant dual complementarity, because G:U and A*C represent transitory mutational states in simultaneous sense–antisense in-frame coding, where one strand has already gained the mutation, whereas the opposite strand remains in the parental state (Figure 2a). However, we observed significant dual complementarity in pairs with such illegitimate pairing (G:U and A*C) at the flanking positions, but not in pairs with the illegitimate base pairs in the central position. Namely, out of 16 amino acid tetrads, only two (Ala(GCC), Gly(GGC), Val(GUC) and Asp(GAC)) and (Ala(GCG), Val(GUG), Arg (CGC) and His(CAC)) were complementary at the second position in their acceptor and anticodon in *all four* combinations: two legitimate G-C and A-U, one wobbling G-U and one A*C (Figure 2b). This means that (1) the central nucleotide-based skeleton of the genetic code (Table 1) was initially established for a few (four to six) amino acids, such as those from the above two tetrads and (2) subsequently, the code expanded mostly *via* conservative, or even silent, substitutions of the flanking nucleotides (Figure 2a). Interestingly, two basic amino acids, Arg and His, show significant stereochemical affinity to cognate triplets in selected RNA aptamers (Yarus *et al*., 2005). Gly, Ala, Asp and Val were the most preponderant of the abiotically synthesized amino acids (Miller, 1987). These four amino acids and their nearest one-transition-step-apart mutational derivatives generated the tRNA tree with a major N**R**N vs *И***Y***И* dichotomy (Figure 2b). This dichotomy is consistent with (1) the primacy of Gly, Ala, Asp and Val in nearly every scenario of the origin of the genetic code, (2) the double-strand coding-based expansion of the genetic code, and (3) the preservation of dual complementarity. Significantly, no other amino acid tetrad generated such a bilateral branching pattern (Figure 2c).

The problem of the r-aaRS→p-aaRS transition

Our Ariadne's thread in the labyrinth of possible evolutionary transitions from ribozymes to proteins is the concept that in the emerging genetic code and associated translation machinery, both complementary strands of ancestral genes could have been used not only as catalysts (Kuhns and Joyce, 2003), but also, later, as first templates for encoded protein synthesis—that is, the future coding (sense) and noncoding (antisense) strands were originally both coding (Eigen and Schuster, 1979; Fukuchi and Otsuka, 1992; Rodin and Ohno, 1995, 1997; Carter and Duax, 2002; Pham *et al*., 2007). Therefore, we looked for 'fingerprints' of this primordial strand symmetry not only in tRNAs, but also in aaRSs and in the organization of the genetic code itself. The complementary modes of tRNA recognition by class I and II p-aaRSs were of particular interest in this regard.

Complementarity-based subcode for two modes of tRNA aminoacylation

All class I synthetases, except TyrRS (Yaremchuk *et al*., 2002) and TrpRS (Yang *et al*., 2006), approach the acceptor helix of the tRNA from the minor groove side and attach the amino acid to the 2′OH of the terminal adenine A76; by contrast, all class II synthetases, except PheRS (Goldgur *et al*., 1997), approach the acceptor helix from the opposite (major groove) side and attach the amino acid to the 3′OH. The distribution of these two classes in the code table does not appear to be arbitrary. In particular, it is immediately clear that all amino acids from the second column of the genetic code table (NCN codons) belong to class II, whereas all but Phe from the first column (NUN codons) belong to class I (Table 1). The main chemical properties of amino acids are determined by their side-chain R-groups: the nonpolar aliphatic Gly, Ala, Pro, Val, Leu and Ile; polar uncharged Ser, Thr, Asn, Cys, Met and Gln; negatively charged Asp and Glu; positively charged Lys and Arg; and ring/aromatic His, Phe, Tyr and Trp. Interestingly, the two classes of aaRSs are equally represented in each R-group (Patel, 2005, 2007). However, the amino acids with larger R-groups belong to class I, whereas their counterparts with smaller R-group belong to class II (Patel, 2005). Moreover, the median hydrophobicities of the two classes are very different (Pham *et al*., 2007).

With the exception of LysRS, every synthetase class assignment is invariant throughout the eubacteria, archaea and eukarya, suggesting that class assignment has not altered since the universal common ancestor of the three major kingdoms was extant (Cusack, 1997). It is unknown, however, whether this invariance was preserved due to steric or other constraints associated with amino acids, or their tRNAs (Frugier *et al*., 1993; Sissler *et al*., 1997; Ribas de Pouplana and Schimmel, 2001a, b). The example of $tRNA^{Lys}$ is particularly revealing in this regard; in some archaebacteria, $tRNA^{Lys}$ is aminoacylated by class I LysRS (Ibba *et al*., 1997) instead of the 'regular' class II LysRS.

This double assignment of LysRS hints that either of the two enzyme classes is probably versatile enough to be able to aminoacylate tRNAs in all 20 cases. Why, then, are the p-aaRSs divided into two classes? Our recent

analyses (Rodin and Rodin, 2006b) suggest a potential explanation.

If the direction in which the aaRS approaches the tRNA acceptor stem is used as a criterion of classification (instead of the class I vs class II dichotomy), then a strikingly nonrandom pattern of tRNA aminoacylation is revealed (Rodin and Rodin, 2006b). This alternative classification is shown in Table 2, with the two different modes of tRNA recognition represented by yellow (minor groove side) and blue (major groove side). Also, in this table, the AGG and AGA codons are assigned to blue Ser or Gly instead of yellow Arg, as they are in some mitochondrial codes (Knight *et al*., 2001). Alternatively, one can obtain the same yellow vs blue pattern by assuming an Arg↔Lys swap between codons AGR and AAR in Table 1. Remarkably, this swap is consistent with the fact (brought to our attention by E Szathmary) that the Arg-specific binding sites of selected RNA aptamers contain Lys's AAA codons (Caporaso *et al*., 2005). Another observation is that despite the two swaps, Phe↔Tyr and Lys↔Arg, this representation of the genetic code (Table 2) does not violate an equality of the two modes of tRNA recognition, from the minor and major groove sides, in each R-group 'subclass' of amino acids (Patel, 2007).

Thus modified the first genetic code column (NUN codons) is uniformly yellow, the second column (NCN) is uniformly blue, and the two remaining columns (NAN and NGN) appear to complement each other almost perfectly at the flanking codon positions (Table 2). Specifically, in the fourth column (NGN), all yellow codons start with a pyrimidine (Y = C or U) and all blue codons except UGG (Trp) start with a purine (R = A or G). The third column (NAN) also shows the yellow/blue split, but in a complementarily mirror manner (R/Y) and this time at the third, not the first, codon position.

The flanking codon nucleotides are directly connected by the Watson–Crick pairings under only one specific coding scenario, that is, when both complementary gene strands encode proteins, in the same frame. This prompted us to rearrange the genetic code in a way that places complementary codons face-to-face with each other. In this representation, a remarkable mirror symmetry becomes evident (Figure 3). Moreover, this symmetry demonstrates the otherwise latent subcode for the two modes of tRNA recognition and, correspondingly, the two types of anticodon pairs. Specifically (Figure 3b):

(i) If two complementary codons contain YY vs RR at the second and adjacent (either first or third) positions, their aaRSs approach the tRNA acceptor from the same side of the groove (minor (yellow) for 5′N**AR**3′ × 5′**YU**И3′ codon pairs or major (blue) for 5′**RG**N3′ × 5′И**CY**3′ codon pairs).
(ii) If these positions are occupied by RY and YR, the modes of tRNA recognition are different, one from the minor groove side and the other from the major groove side, namely: minor (yellow) 5′**YG**N3′ vs major (blue) 5И**CR**3′, and mirror-symmetrically, major (blue) 5′И**AY**3′ vs minor (yellow) 5′**RU**N3′.

These two rules also hold for anticodons with G↔C, A↔U and R↔Y replacements.

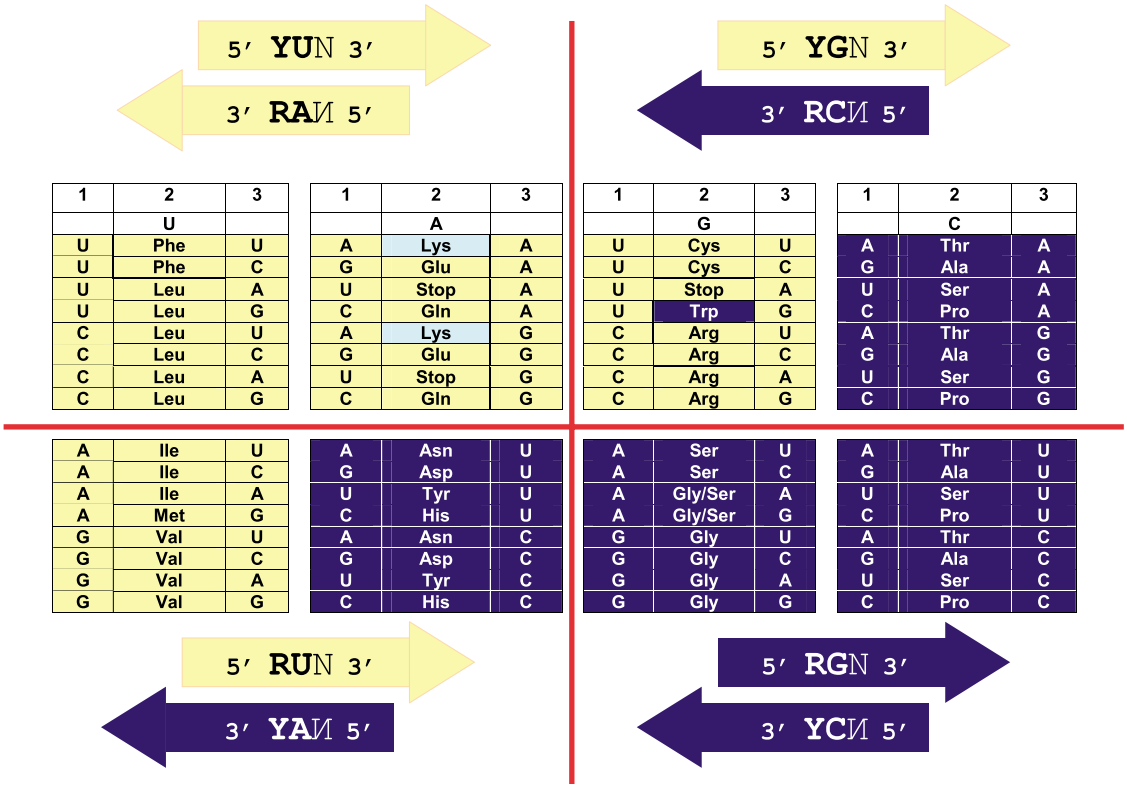
The distinction between (i) and (ii) makes sense. The YR and RY dinucleotides include CG, GC, UA and AU palindromes, each of which is indistinguishable from its complement. Thus, even single base shifts in the

Table 2 The genetic code representation in which yellow and blue colors mark the two modes of tRNA recognition—from the minor and major groove sides of the acceptor stem, respectively[a]

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

Minor groove side
Major groove side

See Rodin and Rodin (2006b) for details.
[a]Lys is shown in a lighter shade of blue to reflect its activation by class I synthetase in some archaebacteria (Ibba *et al*., 1997). Stop codons are colored yellow because the known cases of their 'capture' by amino acids are mostly from class I. Codons AGG and AGA are assigned to blue Ser or Gly, as they are in mitochondria (Knight *et al*., 2001).



1	2	3
	U	
U	Phe	U
U	Phe	C
U	Leu	A
U	Leu	G
C	Leu	U
C	Leu	C
C	Leu	A
C	Leu	G

1	2	3
	A	
A	Lys	A
G	Glu	A
U	Stop	A
C	Gln	A
A	Lys	G
G	Glu	G
U	Stop	G
C	Gln	G

1	2	3
	G	
U	Cys	U
U	Cys	C
U	Stop	A
U	Trp	G
C	Arg	U
C	Arg	C
C	Arg	A
C	Arg	G

1	2	3
	C	
A	Thr	A
G	Ala	A
U	Ser	A
C	Pro	A
A	Thr	G
G	Ala	G
U	Ser	G
C	Pro	G

A	Ile	U
A	Ile	C
A	Ile	A
A	Met	G
G	Val	U
G	Val	C
G	Val	A
G	Val	G

A	Asn	U
G	Asp	U
U	Tyr	U
C	His	U
A	Asn	C
G	Asp	C
U	Tyr	C
C	His	C

A	Ser	U
A	Ser	C
A	Gly/Ser	A
A	Gly/Ser	G
G	Gly	U
G	Gly	C
G	Gly	A
G	Gly	G

A	Thr	U
G	Ala	U
U	Ser	U
C	Pro	U
A	Thr	C
G	Ala	C
U	Ser	C
C	Pro	C

Figure 3 Rearrangement of the genetic code that puts complementary codons head-to-head with each other (in the center). Yellow and blue mark the two modes of tRNA recognition—from the minor and major groove sides of the acceptor stem, respectively. Antiparallel (head-to-tail) oriented arrows show the four types of pairs of corresponding codons that comprise the subcode for these two modes in the early RNP world. 1, 2 and 3 denote codon positions, N and complementary *И* present all four nucleotides; R, purine (G or A); Y, pyrimidine (C or U). For details, see Rodin and Rodin, 2006b.

recognition of the corresponding tRNAs would substantially increase the risk of amino acid confusion. Such errors were likely quite frequent in the primordial RNA life catalyzed by ribozymes. The problem persists if the cognate synthetases spread the tRNA recognition beyond these anticodons in the same direction. If they spread the tRNA recognition in opposite directions, this would greatly decrease the risk of incorrect aminoacylation (Figure 4). For YY and RR dinucleotides at the first–second or the second–third positions, this risk is not immediately obvious, and consequently, the synthetases either both approach from the major groove side or both approach from the minor groove side.

If we assume that the tRNA cloverleaf recognition spreads from the anticodon center in the opposite directions, then the corresponding pair of aaRSs must bind to their cognate tRNAs from the opposite (major or minor groove) sides. This is precisely the case with the two classes of protein synthetases. However, the subcode rules (i) and (ii) apply to codons and anticodons, whereas the protein synthetases most likely evolved from minimal catalytic modules that interacted with the acceptor stem (Schimmel *et al.*, 1993). This contrast, in conjunction with the chicken-or-egg conundrum, suggests that the subcode for the two aminoacylations revealed by the 'yellow-blue' pattern (Figure 3) was initially established by two r-aaRSs (Figure 4). However, in our first report (Rodin and Rodin, 2006b), we overlooked what is quite possibly the most conclusive argument yet in support of this hypothesis. It is described below.

The two modes of tRNA aminoacylation are not always symmetric

In general, the revealed yellow/blue pattern (Figure 3) appears to be almost perfectly symmetric with respect to the discrimination between complementary anticodons. (The only deviation is caused by a 'major-groove-side' TrpRS. Yet, Trp is not that exceptional because it still represents class I, and is a relatively late amino.) For pairs of the first type (YY vs RR), the symmetry looks invariant to 'flipping' the colors. For pairs of the second type (RY vs YR), the direction in which the given r-aaRS 'spreads' also does not matter, that is, it is irrelevant which half-tRNA, from minor groove side or major groove side (yellow or blue), is involved. What matters, although, is that the complementary partner spreads in the opposite direction (Figure 4). We have tacitly accepted this isotropy before (Rodin and Rodin, 2006b), but this symmetry is deceptive. On closer inspection, the two directions cease to be equal if we take into account that in all tRNAs, regardless of their complementary partnerships, the anticodon triplet has adjacent UY dinucleotides (mostly UC) at its 5′ side, and adjacent RN dinucleotides (mostly AA) at its 3′ side.

For any given pair of complementary anticodons, there are four possible scenarios of their recognition by two putative r-aaRSs (Figure 4; see also Table 3):

1. both from the minor groove sides (yellow and yellow), $5' \times 5'$

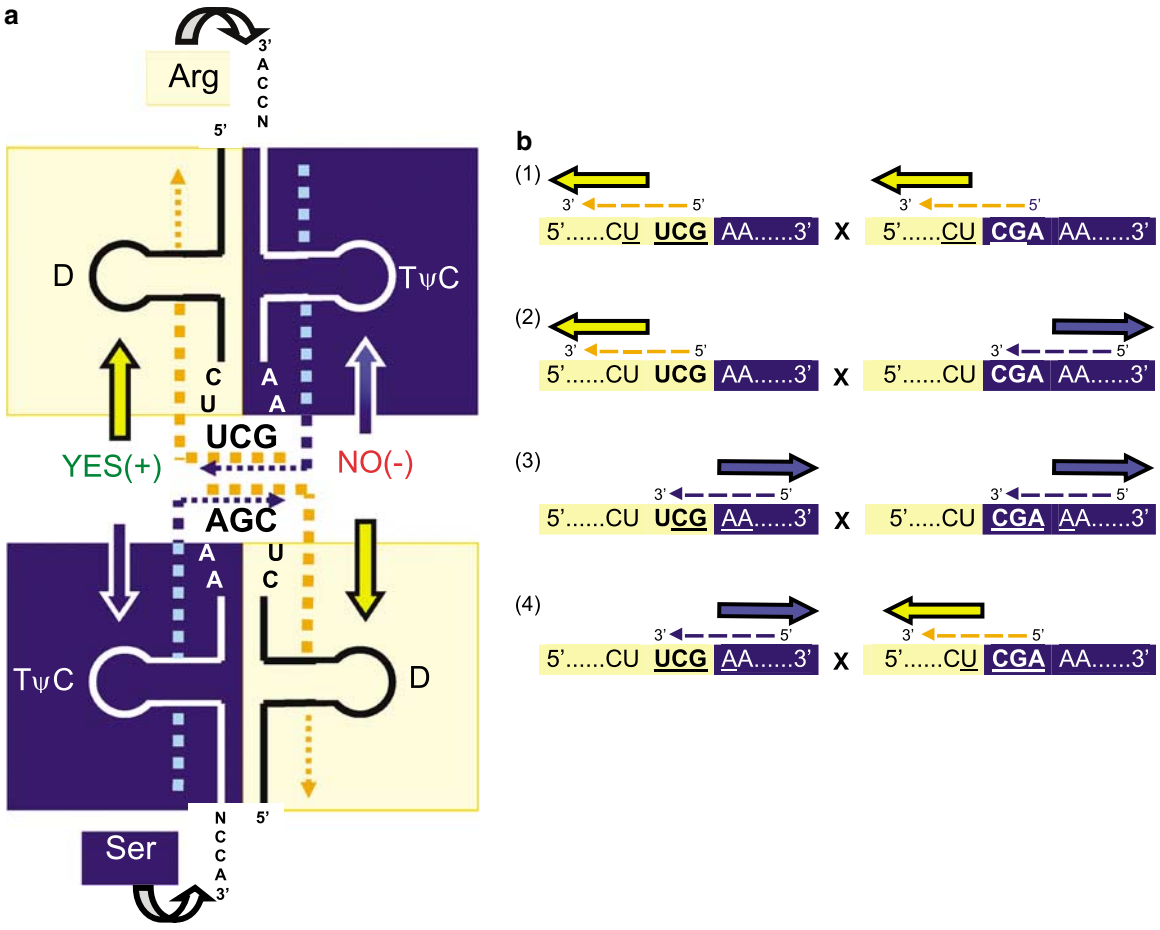


Figure 4 Pseudosymmetric recognition of tRNA complementary halves by two putative r-aaRSs. The $tRNA^{Arg}$ with anticodon UCG vs $tRNA^{Ser}$ with anticodon CGA pairing is shown, as an example of two tRNAs with complementary anticodons of the $5' \times 3'$ type (see text). Solid arrows indicate the two opposite directions in which the tRNA recognition spreads out beyond the anticodon triplet. Most probably, this recognition was based on complementary pairing. Dotted arrows indicate two r-aaRSs that recognize the tRNA substrate in the same, $5' \rightarrow 3'$, direction; their anticodon-binding sites are located at the opposite, 3′ and 5′, ends. As the two r-aaRSs recognize complementary halves of tRNA molecules, it is possible that originally they were also complementary to each other. (**a**) Complementary mirror representation of the tRNAs for Arg (UCG) and Ser (GCA). In each of the two cloverleaves, the bilateral halves are colored yellow and blue in accordance with the minor and major groove sides, respectively. The corresponding two dotted arrows, yellow and blue, denote their recognition by putative r-aaRSs. The scenario with the pair of arrows on the right (corresponding to scenario 4 in Figure 4b) is improbable because, in this orientation, the two aa-specific r-aaRSs encounter the same 5′AGCU3′ motif, thereby potentially confusing the cognate tRNAs. Conversely, the scenario with the arrows on the left (scenario 2 in Figure 4b) does not cause this problem and is, in fact, observed in nature for two p-aaRSs. (**b**): Four scenarios for the recognition of the Arg and Ser anticodon loops by two putative ribozymes. The similar (undistinguishable at R/Y resolution) or identical tetranucleotides are underlined. Scenarios (1), (3) and (4) all contain confusion-prone motifs for each given pair of anticodon loops. The only confusion-proof scenario is (2).

2. one from the minor and one from the major groove side (yellow and blue), $5' \times 3'$
3. both from the major groove sides (blue and blue), $3' \times 3'$
4. one from the major and one from the minor groove side (blue and yellow), $3' \times 5'$.

For the example shown in Figure 4, only the second scenario ($5' \times 3'$) allows the putative r-aaRSs to meet dissimilar sequences within their anticodon loops, thereby providing faultless discrimination between the cognate tRNAs at aminoacylation. In contrast, the other scenarios would contain similar tetranucleotides that included at least one anticodon (underlined in Figure 4b); moreover, these tetranucleotides would be undistinguishable for the two r-aaRSs at binary resolution, purine (R) vs pyrimidine (Y). The fourth scenario is actually the worst—the two putative r-aaRSs would encounter the identical tetranucleotides (5′UCGA3′) that included both anticodons, even without shifting. Thus, although the fourth scenario is a mirror copy of the second, it has a much higher potential for anticodon confusion, and subsequent erroneous aminoacylation.

In a similar fashion, we tested each of the 32 pairs of complementary anticodons for the risk of amino acid confusion under each of the four scenarios (Table 3; low- and high-risk cases denoted by pluses and minuses, respectively). As expected, the macrosymmetry of this test was well pronounced. For example, the $5' \times 3'$ and $3' \times 5'$ scenarios are equal with respect to the total plus/minus ratio of 24:8 (3:1). In sharp contrast, the $5' \times 5'$ and $3' \times 3'$ scenarios are generally more easily confused (the ratio is 1:1). However, the detailed analysis below indicates that the distribution of complementary

Table 3 Risk of confusion of complementary anticodons (at the purine/pyrimidine R/Y resolution) under four scenarios of tRNA recognition by two putative r-aaRS[a]

	Pairs of complementary anticodons		5′ x 5′ minor/minor	5′ x 3′ minor/major	3′ x 3′ major/major	3′ x 5′ major/minor
NAN x NUN	Phe(GAA)	Glu(UUC)	+	+	+	+
	Phe(AAA)[b]	Lys(UUU)	+	+	+	+
	[c]Leu(CAA)	Gln(UUG)	–	–	–	+
	[c]Leu(UAA)	stop(UUA)	–	–	–	+
	Leu(GAG)	Glu(CUC)	+	+	+	+
	Leu(AAG)[b]	Lys(CUU)	+	+	+	+
	[c]Leu(CAG)	Gln(CUG)	–	–	–	+
	[c]Leu(UAG)	stop(CUA)	–	–	–	+
	Ile(GAU)	Asp(AUC)[b]	+	+	+	+
	Ile(AAU)[b]	Asn(AUU)[b]	+	+	+	+
	Ile(UAU)	Tyr(AUA)[b]	–	+	–	–
	Met(CAU)	His(AUG)[b]	–	+	–	–
	Val(GAC)	Asp(GUC)	+	+	+	+
	Val(AAC)[b]	Asn(GUU)	+	+	+	+
	Val(CAC)	His(GUG)	–	+	–	–
	Val(UAC)	Tyr(GUA)	–	+	–	–
NCN x NGN	[c]Cys(GCA)	Ala(UGC)	–	–	–	+
	Cys(ACA)[b]	Thr(UGU)	–	–	–	+
	[c]Trp(CCA)	Pro(UGG)	–	+	–	–
	Stop(UCA)	Ser(UGA)	–	+	–	–
	Arg(GCG)	Ala(CGC)	–	–	–	+
	Arg(ACG)[b]	Thr(CGU)	–	–	–	+
	Arg(CCG)	Pro(CGG)	–	+	–	–
	Arg(UCG)	Ser(CGA)	–	+	–	–
	Ser(GCU)	Ala(AGC)[b]	+	+	+	+
	Ser(ACU)[b]	Thr(AGU)[b]	+	+	+	+
	Gly/Ser(CCU)	Pro(AGG)[b]	+	+	+	+
	Gly/Ser(UCU)	Ser(AGA)[b]	+	+	+	+
	Gly(GCC)	Ala(GGC)	+	+	+	+
	Gly(ACC)[b]	Thr(GGU)	+	+	+	+
	Gly(CCC)	Pro(GGG)	+	+	+	+
	Gly(UCC)	Ser(GGA)	+	+	+	+

[a]Pairs of complementary anticodons are ordered following Figure 3. Plus signs denote the pairs that have no identical tetra(or more)-nucleotides within the loop 3′YU-XYZ-RN5′, that is, they are distinguishable (under the corresponding scenario) by two putative ribozymes that recognize the complementary tRNA halves. Minus signs mark the opposite, indistinguishable, cases. For each pair, only a zero- or one base-long shift in one of two directions from the anticodon is allowed. Two simultaneous shifts (one in each anticodon loop) are considered highly unlikely (also see Figure 6 legend). The red rectangle encloses 16 RY- vs YR-type pairs that satisfy the second rule of the subcode for 2 aminoacylation (see text). If we postulate the assignment of NGN and NAN anticodons to major and minor groove sides, the number of conceivable scenarios of tRNA recognition is reduced from four to only two (enclosed by black rectangles). The actual evolutionary pathway is shown in green.
[b]5′ANN3′ anticodons usually do not exist—instead, the 5′GNN3′ anticodons recognize not only the legitimate 3′СИИ5′ codons but also the illegitimate wobbling 3′УИИ5′ codons. For the strictly legitimate nine pairs of the RY vs YR type, the +/− ratio is 7:2 (second scenario) vs 3:6 (fourth scenario)—the former being, therefore, seven times more 'secure'.
[c]See text for comments on these particular pairs.

anticodon pairs and cognate amino acids among these four scenarios is definitely asymmetric and nonrandom.

Ribozymic precursors of tRNA synthetases had the same two complementary modes of tRNA aminoacylation

Let us consider first the group of complementary 5′**RG**N3′ and 5′И**CY**3′ anticodons at the bottom of Table 3. These eight GC-rich pairs are reliably distinguishable under any of the recognition scenarios, as if there was no preference/selection at all. This suggests that the actual scenario chosen (3′ × 3′) might reflect the most fundamental aspects of translation. Remarkably, all amino acids from these pairs are believed to be the first or at least among the earliest candidates recruited in translation (Eigen and Schuster, 1979; Klipcan and Safro, 2004; Trifonov, 2005; Patel, 2005, 2007). Furthermore, in all of these major groove/major groove cases, the putative r-aaRSs grow and spread their recognition of tRNAs from the 3′ end, that is, moving first along the acceptor stem, then along the TΨC domain (together comprising the minihelix), next along the variable loop,

and eventually reaching the anticodon. This is perfectly consistent with the original replication-tag functions of the acceptor-like precursors of tRNAs (Weiner and Maizels, 1987, 1999; Maizels and Weiner, 1994); the idea that the ancient operational code is embodied in this part of the tRNA molecule (Schimmel *et al*., 1993; Schimmel and Beebe, 2006); and the concerted complementarity of the acceptor's second bases and complementarity of anticodons (Rodin *et al*., 1996, Rodin and Ohno, 1997).

The next 16 pairs of anticodons (enclosed in red in the middle of Table 3) contain RY and YR at the second and adjacent (either first or third) positions. The Arg(UCG)–Ser(CGA) pair (Figure 4) belongs to this group. The advantage of the $5' \times 3'$ scenario is clear—at the R/Y resolution, its $+/-$ ratio is 12:4 (3:1), whereas the mirror $3' \times 5'$ scenario yields 8:8 (1:1) (Table 3). Moreover, the selection of this scenario fits with the aforementioned earlier choice of the $3' \times 3'$ scenario. Indeed, when the ancient operational code allotted the major groove side r-aaRS to Ala, Thr, Pro and Ser in their complementary pairs with Gly (also major groove side), it made the subsequent reassignment of the same amino acids, Ala, Thr, Pro and Ser, to the mirror (minor groove side) r-aaRS (in pairs with Arg, Trp and Cys; see Table 3) exceedingly costly, and therefore unlikely.

At higher resolution, when r-aaRSs distinguish G from A and C from U, G–U is considered to be a weak, wobbling, bond, while A*C is a clear mismatch. At this level, the relative excess of low-risk cases for amino acid confusion under the $5' \times 3'$ scenario becomes even more pronounced, with a ratio of 15:1. The $5' \times 3'$ scenario is only less favorable for one minor/major groove pair of complementary anticodons (GCA (Cys) and UGC (Ala)) than the $3' \times 5'$ scenario, the actual loops being 5′CU–GCA–AA3′ and 5′UU–UGC–AA3′ (identical tetranucleotides are underlined), respectively. Note that the p-aaRSs for both of these amino acids do not need the anticodon for error-proof aminoacylation of their cognate tRNAs (Schimmel and Beebe, 2006).

At first glance, the four (out of eight) anticodon pairs of the YY vs RR type at the top of Table 3 are at variance with all of the above; they do not share any amino acids with the previous two groups, and their best discrimination can be achieved by the 'minus-free' $3' \times 5'$ scenario instead of the actual $5' \times 5'$ scenario (with four high-risk cases). However, two of these four minuses represent the pairs with stop codons! Advanced translation needs termination marks, and it seems rational that these triplets, which would otherwise be extremely confusable with their complementary partners, should be selected such that they do not convey genetic information, thereby being available for 'punctuation'. Moreover, all pairs of amino acids belonging to this group, other than Glu (CUC) × Leu (GAG), entered translation relatively late (Eigen and Schuster, 1979; Klipcan and Safro, 2004; Trifonov, 2005; Patel, 2005, 2007), possibly when the repertoire of potential assignments to the 3′ r-aaRSs (blue) had been filled, thus increasing the risk for a new amino acid to be confused with an old one.

The actual evolutionary pathway (green in Table 3) includes only two amino acids (both likely to be latecomers) that are either undistinguishable (Cys) or barely distinguishable (Gln) from their complementary partners, even at the higher G/C/A/U level of recognition. It is hardly a coincidence that in many prokaryotes these amino acids represent two of three (Cys, Gln and Asn) indirect older routes for aa-tRNA synthesis: SepRS/SepCysS and Glu-tRNAGln through which Cys and Gln actually entered translation (Ibba *et al*., 2000; Di Giulio, 2002; O'Donoghue *et al*., 2005).

In general, there is a strong correlation between the group of abiotically synthesized amino acids (Eigen and Schuster, 1979; Miller, 1987) and the risk of their confusion with a complementary partner. In 16 out of 32 anticodon pairs, complementary anticodons are easily distinguished under any of the four tRNA recognition scenarios (Table 3). Remarkably, all abiotically synthesized amino acids (that is, Ala, Gly, Asp, Val, Leu, Glu, Ser, Ile, Thr and Pro) fall into this group. To achieve this by chance alone is exceedingly improbable.

Origin of the two modes of tRNA recognition is consistent with the stereochemical affinity of amino acids to their own coding triplets

In pools of RNA aptamers that were selected from presumably random sequences to bind to specific amino acids, the aa-binding sites contained cognate codons and/or anticodons at frequencies considerably greater than expected (Caporaso *et al*., 2005; Yarus *et al*., 2005). This striking association was reported for seven amino acids: Arg, Trp, His, Tyr, Ile, Leu and Phe. Nonspecific aptamers with cognate triplets and an affinity to the hydrophobic l-valine side chain have also been selected (Majerfeld and Yarus, 1994).

Six amino acids with the presumed stereochemical affinity to their own coding triplets—Arg, Trp, His, Tyr, Val and Ile—come from pairs of complementary anticodons 5′N**CR**3′ × 5′**YG***И*3′ and 5′N**AY**3′ × 5′**RY***И*3′ (red frame in the middle of Table 3). It is the minor/major, $5' \times 3'$ (yellow/blue), scenario of tRNA recognition that gives these six amino acids the lowest risk of confusion with complementary partners. Furthermore, in their aptamers, the Arg- and Tyr-binding sites contain not only a coding triplet *per se*, but also all of the pentanucleotides (or their complements) of the anticodon loop. This makes the confusion of Arg and Tyr with their complementary partners (Ser and Ile, respectively) unlikely, if and only if the correct tRNA-recognition scenario ($5' \times 3'$ in these two cases) is used. These pentanucleotides are 5′UGUAG3′ for the Ile × Tyr pair and 5′UCGAA3′ for the Arg × Ser pair (Figures 4a and b).

Remarkably, the above correlation does not seem to hold for the eight pairs of complementary anticodons 5′N**CY**3′ × 5′**RG***И*3′ and their amino acids, Gly, Ala, Thr, Pro and Ser (at the bottom of Table 3). To be more precise, neither positive nor negative results have been reported yet for this group of presumably the earliest amino acids in terms of attempts to select aa-binding RNAs. Yet, even the complete lack of stereochemistry between these amino acids and their cognate triplets is not discouraging. On the contrary, we would expect this if, as we proposed, the recognition of the corresponding pairs of tRNAs was encoded originally in the acceptor stem, was from the major groove side (that is, under the $3' \times 3'$ (blue/blue) scenario), and was independent of the anticodon domain. The strong dependence appears later, in the $5' \times 3'$ minor/major (yellow/blue) scenario, and thus substantiates the very existence of the two

complementary modes of tRNA recognition under aminoacylation.

Furthermore, if indeed the anticodon and first three paired bases in the acceptor helix had a common origin, then the ancient operational code contained not only proto-anticodons on one strand, but necessarily proto-codons on the opposite strand. However, the updated analysis of dual complementarity clearly points to a significant ambiguity of this ancestral double-stranded code (Rodin and Rodin, 2006a). Therefore, it seems reasonable that, in addition to examining individual amino acids for stereochemical affinities to their anticodons or codons, we also test (by both modeling and SELEX-like experiments) the binding preferences between (1) groups of similar amino acids and their coding triplets (Rodin and Rodin, 2006a), and (2) amino acids and codon–anticodon *pairs* (rather than just individual codons or anticodons) (Patel, 2007). Note, in this regard, that the Arg-binding site of the Tetrahymena group I self-splicing introns is located in the major groove of an rRNA precursor's P7 helix, with codon and anticodon triplets opposing each other on the complementary strands (Yarus, 1991, 1993). This said, when we speculate on the early coevolution of the two codes, it might be even more tempting to apply the above tests to the amino acids that are represented by pairs of complementary anticodons 5′**RG**N3′ × 5′И**CY**3′, that is, the eight blue/blue pairs at the bottom of Table 3.

It is probable that the pairs of anticodons 5′N**AR**3′ × 5′**YU**И3′ (at the top of Table 3) do not conform to the pattern because, again, they include stop codons and the amino acids that most probably entered translation relatively late (Phe and Gln). Gln is particularly indicative here because, as we have already mentioned, it is an indirect addition to the genetic code and it is the only amino acid whose binding sites in selected RNA aptamers do not contain coding triplets (Caporaso *et al*., 2005; Yarus *et al*., 2005).

Mimicry between two p-aaRSs and their ribozymic precursors

Most remarkably, for pairs of RY- and YR-containing anticodons (that is, pairs of type (ii) in the subcode for two aminoacylations), it is the 5′ × 3′ scenario which (1) is consistent with the earliest 3′ × 3′ scenario, (2) is more secure than the mirror 3′ × 5′ scenario and (3) is what is actually observed in extant class I and class II p-aaRSs. This crucial nonequivalence between the 5′ × 3′ and 3′ × 5′ scenarios suggests that the revealed subcode for two aminoacylations (Figure 3), and the existence of two complementary versions of p-aaRSs, must have been evolutionarily connected through the direct ribozymic precursors of p-aaRSs.

This intriguing connection highlights the importance of molecular mimicry in the RNA→RNP (RNA plus protein) transition (Nakamura, 2001; Liang and Landweber, 2005; Delarue, 2007), and strongly supports our earlier hypothesis that the two complementary recognition patterns of acceptor stems by the class I and class II p-aaRSs were inherited from the two isofunctional ribozymes (Rodin *et al*., 1996). This is also relevant to the possible origin of the two p-aaRSs from the complementary strands of the same ancestral gene (Rodin and Ohno, 1995) that conceivably, directly recapitulates the preceding complementarity of the two r-aaRSs. The nonrandom complementarity of signature motifs from the class I and II catalytic domains that are aligned in a 'head-to-tail' orientation (Rodin and Ohno, 1995), the real precedent of sense–antisense coding of class I and II aaRS homologs (Carter and Duax, 2002) and the recent artificial creation of the 130-residue minimal catalytic domain of TrpRS (that perfectly fits the minimal catalytic domain of class II aaRSs complementarily and retains the ability to specifically activate tryptophan; Pham *et al*., 2007) strongly support this hypothesis. Moreover, our present analysis shows that this r-aaRS→p-aaRS succession was quite efficient, even without a 'color change:' class I p-aaRSs directly followed their minor groove side, and class II p-aaRSs directly followed their major groove side ribozymic forerunners.

Also, Watson–Crick pairing-based recognition of the operational proto-code by r-aaRSs might imply a local distortion of the acceptor helix. Interestingly, interactions of typical class I protein aaRSs with tRNAs do cause serious changes of the acceptor stem end, including unwinding and disruption of base pairing (Rould *et al*., 1989; Carter, 1993).

Initially, the ancestors of the two p-aaRSs could have played a chaperone role by protecting the acceptor stem from both sides (Ribas de Pouplana and Schimmel, 2001a, b). As to their participation in coding, the updated analysis of the dual complementarity indicates that p-aaRSs began to replace isofunctional ribozymes long before all 64 codons received their final assignment, and yet most likely only after the complementary core of the code (Figure 3) had been established (Rodin and Rodin, 2006a, b). Since then, duplications of tRNA and p-aaRS genes, and their very specific coevolution, might have gradually reduced the code's ambiguity, as outlined in Ribas de Pouplana and Schimmel, 2001a, b; Carter and Duax, 2002; Rodin and Rodin, 2006a, b; Schimmel and Beebe, 2006; and Pham *et al*., 2007.

NCN and NUN codons: the first choice between two aminoacylation modes

Although the four scenarios of tRNA recognition in Table 3 have been evaluated for pairs of complementary anticodons, strictly speaking the 'plus' mark does not necessarily imply their simultaneous entry into the coding system. This means that the gamut of choices (colored green in Table 3) selectively favors any scenario for the formation of the genetic code that takes into consideration the partitioning of aaRSs into two classes. The recent phenomenological model of progressive differentiation-like reduction of codon ambiguity (Delarue, 2007) is no exception. This elegant model is also based on the pattern of tRNA aminoacylation by class I and II aaRSs, similar to the pattern in Table 2 except for its third (NAN) column. However, in contrast to our complementarity-based model, Delarue (2007) interpreted this pattern as a binary decision tree, emphatically asymmetric, like in a longitudinal differentiation process. Each decision is a choice between two options (first r-aaRSs, then their protein successors, p-aaRSs), but the reason why the minor (yellow) or major (blue) groove side is preferred in each particular case (step) remains unclear. In fact, the subcode for two complementary aminoacylations (Figure 3) and its

selective evaluation (Table 3, Figure 4) could provide such explanation. These two models will be compared in detail elsewhere; here, we focus on their mutual benefits associated with the choice between class I and II aaRSs, which is the first choice in Delarue, 2007.

This choice is really self-evident (Table 2): the central C in codons (G in anticodons) leads to tRNA recognition from the major groove side (blue), whereas the central U in codons (A in anticodons) leads to tRNA recognition from the minor groove side (yellow). Recognition of tRNAs by r-aaRSs was likely based on complementary interaction with anticodons. However, this interaction looks absolutely symmetric with regard to G–C vs C–G or A–U vs U–A pairings. Therefore, the choice itself says nothing about why it should involve the first two columns of the genetic code table and not, for example, the third and fourth columns. The answer comes from our analysis of the risk of confusing complementary anticodons that is associated with adjacent 3′U and 5′A/G nucleotides. If our (or Nature's) primary motivation is to minimize the risk of confusion, then the optimal distribution of amino acids and cognate triplets among the two modes of tRNA recognition must necessarily be, as shown in Table 3, that is, starting from the major/major groove sides (blue/blue) pairs of complementary triplets.

If we again consider the possible recruitment of the GUC(Val)–GAC(Asp) pair into a primitive sense–antisense translation as the C→U–G→A derivative of the GCC(Ala)–GGC(Gly) pair (Figure 2a), then the two complementary expansions of the code, Ala→Val and Gly→Asp, may seem equivalent with respect to the fidelity of tRNA aminoacylation by r-aaRSs, but they are not (Figure 5). Indeed, r-ValRS with its anticodon-binding putative GUC site, can recognize not only its cognate GAC anticodon but also the Ala anticodon (GGC) due to U:G wobbling pairing (Figure 5). Importantly, such confusion of new (Val) and old (Ala) amino acids would pleiotropically affect all 'old' Ala codons, not just the one mutated ('new') individual codon. In contrast, r-AspRS, with its GAC site is unable to recognize the Gly anticodon (GCC) because of A*C mispairing (Figure 5). In this case, the G:U wobbling recognition also occurs, but it comes from the 'old' r-GlyRS, not from the 'new' r-AspRS, and therefore, does not bring the risks of the pleiotropic negative effects of Gly→Asp in many old Gly codons.

To avoid multiple mishaps along the above (Ala→Val) lines, a principally different mode of tRNA recognition is needed that would make r-ValRS much less (if at all) confusable with the already established r-AlaRS (this is apparently not required in the case of r-AspRS). And this is precisely what happens in reality: AspRS is of the same type as GlyRS—major (blue) groove type—whereas ValRS adopted the new, minor (yellow) groove, mode of tRNA recognition that safely distinguishes it from AlaRS (major groove).

Switching of r-aaRSs from the major to minor groove sides implies spreading out of tRNA recognition in the opposite (from anticodons) direction (Figure 4). Inevitably, the flanking positions of anticodons (and complementarily codons) will be replaced under rules (i) and (ii) of the subcode for two aminoacylations: the first position is changed for the third position and vice versa.

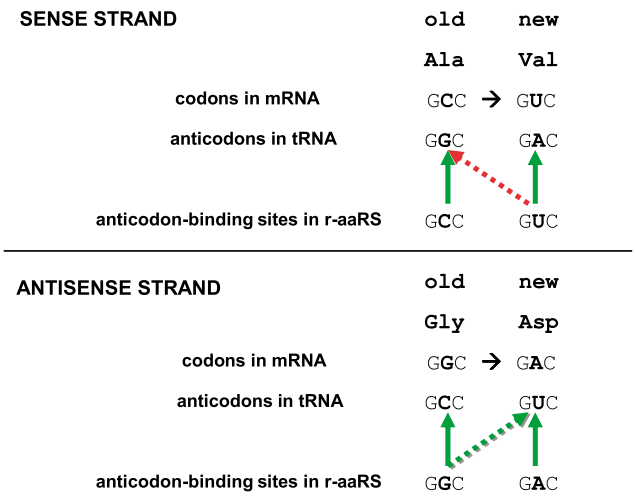


Figure 5 The fundamental asymmetry between complementary expansions of the genetic code lies in a risk of pleiotropic tRNA mis-aminoacylation by r-aaRSs. An example of two complementary expansions is shown with legitimate (solid arrows) and 'wobbling' (dotted arrows) recognitions of anticodons by r-aaRSs. Despite a symmetry of G-U and U-G wobbling pairings, the Ala(GCC)→Val(GUC) expansion is prone to pleiotropic, negative mis-aminoacylations of the old $tRNA^{Ala}$ by the new r-ValRS (marked by a red dotted arrow), whereas the complementary Gly(GGC)→Asp(GAC) expansion has no such disadvantage (marked by a green-dotted arrow).

The assignment of the minor and major groove sides of tRNA recognition to the (NUN) and (NCN) columns of codons (Table 2), coaxes out further choices for the evolution of the genetic code in the direction of the route that was actually chosen—by leaving only two options for each type of complementary pairs (shown by black frames in Table 3) out of conceivable four. In fact, this differentiation of NYN on blue NCN and yellow NUN, likely predetermined by the primal choice of the $3' \times 3'$ scenario for early amino acids such as Ala and Gly (Figure 5), makes the fourth, $3' \times 5'$, scenario very unlikely, and the advantages of the second, $5' \times 3'$, scenario even more convincing (Table 3). Thus, the asymmetric differentiation-based model (Delarue, 2007) and our 'symmetric' complementarity-based model (Rodin and Rodin, 2006b) supplement, rather than contradict, each other.

Concluding remarks

The genetic code-shaping processes—coevolution of anticodons with their putative duplicates in the acceptor stem of tRNAs, and the transition from r-aaRSs to p-aaRSs with the same two modes of tRNA recognition—are apparently interrelated. Both suggest that the codon repertoire is likely expanded by means of complementary pairs. This is beautifully reflected in the yin-yang-like mirror symmetric pattern of tRNA aminoacylation that is revealed in the genetic code table after its complementary transformation (Figure 3). In fact, the possible in-frame coding of two p-aaRSs by the two complementary strands of the same ancestral gene represents perhaps the most important variation on the theme.

Yet, our higher-resolution analysis of the mirror symmetric pattern (Figure 3) revealed the fundamental nonequivalence of different pairs of complementary anticodons, as far as the risk of their confusion during aminoacylation is concerned. The cause of such errors is located in the anticodon-flanking nucleotides U and R. These two nucleotides are almost invariant; hence, they do not affect the aa-specificity of tRNAs. This was also most likely the case for primordial tRNAs that were aminoacylated by r-aaRSs. However, Table 3 and Figure 4 show how important these U and R nucleotides can become if the risk of confusion of tRNAs is taken into account during the recognition of tRNAs by aaRSs from the minor and major groove sides. The cost of such confusion is the highest for complementary triplets, because they, more often than not, encode very different amino acids. It seems reasonable, therefore, to assume that each aa-specific tRNA had both aminoacylation options (either by minor groove or major groove r-aaRSs) available at first, and the priority in choosing the correct option was a lower risk of confusion with its complementary partner.

When looked at from this point of view, the two complementary, symmetric modes of tRNA recognition by aaRSs—from minor vs major groove sides—are not perfectly symmetric (Figure 4 and Table 3). Of particular interest in this respect is the difference between the two groups of complementary RR- and YY-containing pairs of anticodons, that is, the 5′**RG**N3′ vs 5′*И***CY**3′ pairs, which represent early amino acids and the 5′ N**AR**3′ vs 5′**YU***И*3′ pairs, which represent later amino acids. The example in Figure 6 shows why the recognition

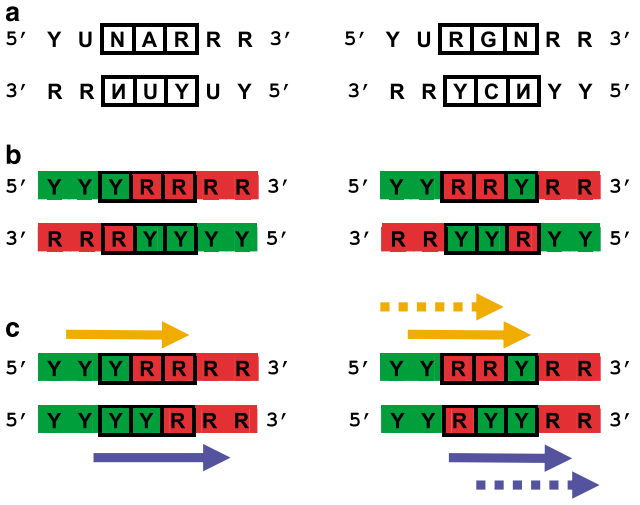


Figure 6 The difference between 5′NAR3′ × 5′YU*И*3′ and 5′RGN3′ × 5′*И*CY3′ pairs of complementary anticodons with regard to the risk of confusion by r-aaRSs. (**a**) The complementary anticodon loops in the antiparallel (head-to-tail) orientation to each other. (**b**) The same as (**a**), but at the purine/pyrimidine (R/Y) resolution. Red and green indicate purines and pyrimidines, respectively. (**c**) The same as (**b**), but in the standard parallel (head-to-head) orientation. Yellow and blue arrows correspond to the minor and major groove sides of tRNA recognition by putative r-aaRSs. Solid arrows indicate the identical tetranucleotides that cover the anticodon triplet in its entirety. Dotted arrows indicate identical tetranucleotides that do not actually cover the anticodon triplet under consideration, and therefore, should be ignored. By looking at the solid arrows, it becomes clear that the 5′NAR3′ × 5′YU*И*3′ pair has a much higher risk of confusion than the 5′RGN3′ × 5′*И*CY3′ pair.

confusion is virtually impossible for early amino acids and very likely for later amino acids. Not surprisingly, the stop codons UAA and UAG belong to the later group of amino acids.

In conclusion, we believe that the uncovered subcode for the two tRNA recognition modes (from the minor and major groove sides)—the subcode that is essentially associated with complementary anticodons and their adjacent U and R nucleotides—represents an ancient and very important milestone in the history of life. Furthermore, our analysis suggests that the two complementary modes of tRNA aminoacylation, mediated by the ancient ribozymes, constitute the missing link between the two fundamental components of the genetic coding system: the classic code embodied in the anticodon and the operational code embodied mostly in the acceptor stem. Originally, in conformity with an updated dual complementarity and the 3′ × 3′ scenario of tRNA recognition for earliest pairs of amino acids (Table 3), the anticodon loop structure might have evolved to fit an operational code (with a fixed second base in the acceptor stem) rather than the other way around. In this sense, the presumable antiquity of the operational code (Schimmel *et al*., 1993) is compatible with the logical primacy of anticodons (Szathmary, 1999; Rodin and Rodin, 2006b). This is also consistent with the model of early coding pentanucleotides 5′URNYA3′ that provided archaic ribosome-free translation (Crick *et al*., 1976).

Acknowledgements

We thank Paul Schimmel, Charles Carter, Eors Szathmary, Apoorva Patel and Massimo Di Giulio for many thought-provoking discussions and valuable suggestions. We also thank anonymous reviewers for many useful criticisms and suggestions that have enhanced the quality and readability of the manuscript. Finally, we are greatly indebted to Keely Walker, Sarah Cheung and Christine Foreman for work on the manuscript.

References

Bloch D, McArthur B, Mirrop S (1985). tRNA-rRNA sequence homologies: evidence for an ancient modular format shared by tRNAs and rRNAs. *BioSystems* **17**: 209–225.

Caporaso JG, Yarus M, Knight R (2005). Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J Mol Evol* **61**: 597–607.

Carter Jr CW (1993). Cognition, mechanism, and evolutionary relationships in aminoacyl-tRNA synthetases. *Annu Rev Biochem* **62**: 715–748.

Carter Jr CW, Duax WL (2002). Did tRNA synthetase classes arise on opposite strands of the same gene? *Mol Cell* **10**: 705–708.

Crick FHC (1958). On protein synthesis. *Symp Soc Exp Biol* **12**: 138–163.

Crick FHC (1968). The origin of the genetic code. *J Mol Biol* **38**: 367–380.

Crick FHC, Brenner S, Klug A, Pieczenik G (1976). A speculation on the origin of protein synthesis. *Orig Life* **7**: 389–397.

Cusack S (1997). Aminoacyl-tRNA synthetases. *Curr Opin Struct Biol* **7**: 881–889.

Cusack S, Berthet-Colominas C, Härtlein M, Nassar N, Leberman R (1990). A second class of synthetase structure revealed by X-ray analysis of *Escherichia coli* seryl-tRNA synthetase at 2.5 Å. *Nature* **347**: 249–255.

De Duve C (1988). The second genetic code. *Nature* **333**: 117–118.

Delarue M (2007). An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA* **13**: 1–9.

Di Giulio M (1992). On the origin of the transfer RNA molecule. *J Theor Biol* **159**: 199–214.

Di Giulio M (2002). Genetic code origin: are the pathways of type Glu-tRNA(Gln) → Gln-tRNA(Gln) molecular fossils or not? *J Mol Evol* **55**: 616–622.

Eigen M, Schuster P (1979). *Hypercycle: a Principle of Natural Self-organization*. Heidelberg: Springer-Verlag.

Eriani G, Delarue M, Poch O, Gangloff J, Moras D (1990). Partition of aminoacyl-tRNA synthetases into two classes based on mutually exclusive sets of conserved motifs. *Nature* **347**: 203–206.

Frugier M, Florentz C, Schimmel P, Giege R (1993). Triple aminoacylation of a chimerized transfer RNA. *Biochemistry* **32**: 14053–14061.

Fukuchi S, Otsuka J (1992). Evolution of metabolic pathway by chance assembly of enzyme proteins generated from sense and antisense strands of pre-existing genes. *J Theor Biol* **158**: 271–291.

Goldgur Y, Mosyak L, Reshetnikova L, Ankilova V, Lavrik O, Khodyreva S *et al*. (1997). Crystal structure of phenylalanyl-tRNA synthetase from *T. thermophilus* complexed with cognate $tRNA^{Phe}$. *Structure* **5**: 59–68.

Hou Y-M, Schimmel P (1988). A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* **333**: 140–145.

Ibba M, Becker HD, Stathopoulos C, Tumbula DL, Söll D (2000). The adaptor hypothesis revisited. *Trends Biochem Sci* **25**: 311–316.

Ibba M, Morgan S, Curnow AW, Pridmore DR, Vothknecht UC, Gardner W *et al*. (1997). Euryarchael lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* **278**: 1119–1122.

Klipcan L, Safro M (2004). Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. *J Theor Biol* **228**: 389–396.

Knight RD, Freeland SJ, Landweber LF (2001). Rewriting the keyboard: evolvability of the genetic code. *Nature Rev Genet* **2**: 49–58.

Kuhns ST, Joyce GF (2003). Perfectly complementary nucleic acid enzymes. *J Mol Evol* **56**: 711–717.

Liang H, Landweber LF (2005). Molecular mimicry: quantitative methods to study structural similarity between protein and RNA. *RNA* **11**: 1167–1172.

Maizels N, Weiner AM (1994). Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc Natl Acad Sci USA* **91**: 6729–6734.

Majerfeld I, Yarus M (1994). An RNA pocket for an aliphatic hydrophobe. *Nature Struct Biol* **1**: 287–292.

Miller SL (1987). Which organic compounds could have occurred on the prebiotic earth. *Cold Spring Harbor Symp Quant Biol* **52**: 17–27.

Nakamura Y (2001). Molecular mimicry between protein and tRNA. *J Mol Evol* **53**: 282–289.

O'Donoghue P, Sethi A, Woese CR, Luthey-Schulten ZA (2005). The evolutionary history of Cys-$tRNA^{Cys}$ formation. *Proc Natl Acad Sci* **102**: 19003–19008.

Patel A (2005). The triplet genetic code had a doublet predecessor. *J Theor Biol* **233**: 527–532.

Patel A (2007). Towards understanding the origin of genetic languages. In: Abbott D, Davies PCW, Pati AK (eds). *Quantum Aspects of Life*. Imperial College Press: London.

Pham Y, Li L, Kim A, Erdogan O, Weinreb V, Butterfoss GL *et al*. (2007). A minimal Trp RS catalytic domain supports sense/antisense ancestry of class I and II aminoacyl-tRNA synthetases. *Mol Cell* **25**: 851–862.

Ribas de Pouplana L, Schimmel P (2001a). Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. *Cell* **104**: 191–193.

Ribas de Pouplana L, Schimmel P (2001b). Aminoacyl-tRNA synthetases: potential markers of genetic code development. *Trends Biochem Sci* **26**: 591–596.

Rodin S, Ohno S (1995). Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. *Origins Life Evol Biosphere* **25**: 565–589.

Rodin S, Rodin A, Ohno S (1996). The presence of codon–anticodon pairs in the acceptor stem of tRNAs. *Proc Natl Acad Sci USA* **93**: 4537–4542.

Rodin SN, Ohno S (1997). Four primordial modes of tRNA-synthetase recognition, determined by the (G,C) operational code. *Proc Natl Acad Sci USA* **94**: 5183–5188.

Rodin SN, Rodin AS (2006a). Origin of the genetic code: first aminoacyl-tRNA synthetases could replace isofunctional ribozymes when only the second base of codons was established. *DNA Cell Biol* **25**: 365–375.

Rodin SN, Rodin AS (2006b). Partitioning of aminoacyl-tRNA synthetases in two classes could have been encoded in a strand-symmetric RNA world. *DNA Cell Biol* **25**: 617–626.

Rodin SN, Rodin AS (2007). Evolution by gene duplications: From the origin of the genetic code to the human genome. In: Dobretzov N, Kolchanov N (eds). *Biosphere Origin and Evolution*. Springer: Berlin, pp 253–272.

Rould MA, Perona JJ, Söll D, Steitz TA (1989). Structure of *E. coli* Glutaminyl-tRNA synthetase complexed with $tRNA^{Gln}$ and ATP at 2.8 Å resolution. *Science* **246**: 1135–1142.

Ruff M, Krishnaswamy S, Boeglin M, Postersman A, Mitschler A, Rodjarny A *et al*. (1991). Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-trNA synthetase complexed with tRNA. *Science* **252**: 1682–1689.

Schimmel P, Beebe K (2006). Aminoacyl tRNA synthetases: from the RNA world to the theater of proteins. In: Gesteland RF, Cech TR, Atkins JF (eds). *The RNA World*. Cold Spring Harbor Lab. Press, Plainview, NY, pp 227–255.

Schimmel P, Giege R, Moras D, Yokoyama S (1993). An operational RNA code for amino acids and possible relation to genetic code. *Proc Natl, Acad Sci USA* **90**: 8763–8768.

Shimizu M (1982). Molecular basis for the genetic code. *J Mol Evol* **18**: 297–303.

Sissler M, Eriani G, Martin F, Giege R, Florentz C (1997). Mirror image alternative interaction patterns of the same tRNA with either class I arginyl-tRNA synthetase or class II aspartyl-tRNA synthetase. *Nucleic Acid Res* **25**: 4899–4906.

Sprinzl M, Vassilenko KS (2005). Compilation of tRNA sequences and sequences of tRNA genes. *Nucl Acids Res* **1:33**: D139–D140.

Szathmary E (1991). Codon swapping as a possible evolutionary mechanism. *J Mol Evol* **32**: 178–182.

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: amino acids as cofactors in an RNA world. *Trends Genet* **15**: 223–229.

Trifonov EN (2005). Theory of early molecular evolution: predictions and confirmations. in: Eisenhaber F (ed). *Discovering Biomolecular Mechanisms with Computational Biology*. Landes Bioscience, Georgetown, pp 107–116.

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**: 7383–7387.

Weiner AM, Maizels N (1999). The genomic tag hypothesis: modern viruses as molecular fossils of ancient strategies for genomic replication, and clues regarding the origin of protein synthesis. *Biol Bull* **196**: 327–330.

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

Yang X, Otero FJ, Ewalt KL, Liu J, Swairjo MA, Köhrer C *et al*. (2006). Two conformations of a crystalline human tRNA synthetase–tRNA complex: implications for protein synthesis. *EMBO J* **25**: 2919–2929.

Yanofsky C (2007). Establishing the triplet nature of the genetic code. *Cell* **128**: 815–818.

Yaremchuk A, Kriklivyi I, Tukalo M, Cusack S (2002). Class I tyrosyltRNA synthetase has a class II mode of cognate tRNA recognition. *EMBO J* **21**: 3829–3840.

Yarus M (1991). An RNA–amino acid complex and the origin of the genetic code. *New Biol* **3**: 183–189.

Yarus M (1993). An RNA-amino acid affinity. In: Gesteland RF, Atkins JF (eds). *The RNA World*. Cold Spring Harbor Lab. Press, Plainview, NY, pp 205–217.

Yarus M (1998). Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J Mol Evol* **47**: 109–117.

Yarus M, Caporaso JG, Knight R (2005). Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem* **74**: 125–151.