TWO TYPES OF AMINOACYL-TRNA SYNTHETASES COULD BE ORIGINALLY ENCODED BY COMPLEMENTARY STRANDS OF THE SAME NUCLEIC ACID

SERGEI N. RODIN[1,2] and SUSUMU OHNO[1]

[1] *Beckman Research Institute of the City of Hope, 1450 E. Duarte Road, Duarte, California 91010–3000, U.S.A.**

[2] *Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, av.acad.Lavrentjeva 10, Novosibirsk 630090, Russia*

(Received September 19, 1994)

Abstract. The lack of even a marginal similarity between the two aminoacyl-tRNA synthetase (aaRS) classes suggests their independent origins (Eriani *et al.*, 1990; Nagel and Doolittle, 1991). Yet, this independence is a puzzle inconsistent with the common origin of transfer RNAs, the coevolutionary theory of the genetic code (Wong, 1975, 1981) and other associated data and ideas. We present here the results of antiparallel 'class I versus class II' comparisons of aaRSs within their signature sequences. The two main HIGH- and KMSKS-containing motifs of class I appeared to be complementary to the class II motifs 2 and 1, respectively. The above sequence complementarity along with the mirror-image between crystal structures of complexes formed by the opposite aaRSs and their cognate tRNAs (Ruff *et al.*, 1991), and the generally mirror (*'head-to-tail'*) mapping of the basic functional sites in the sequences of aaRSs from the opposite two classes led us to conclude that these two synthetases emerged synchronously as complementary strands of the same primordial nucleic acid. This conclusion, combined with the hypothesis of tRNA concerted origin (Rodin *et al.*, 1993a,b), may explain many intriguing features of aaRSs and favor the elucidation of the origin of the genetic code.

1. Introduction: The Problem

It is hard to overrate the prospects which a comparative study of aminoacyl-tRNA synthetases (aaRS) promises for the elucidation of the origin of the genetic code. Indeed, twenty aaRSs, one for each amino acid, are responsible for the first step in translation – selection and charging of the cognate amino acid to all of the isoacceptor tRNAs with the corresponding anticodons. In fact, it is the aaRS which provides the correct codon-to-amino-acid assignment. On the other hand, common features of the aminoacylation step include a requirement for ATP in addition to the specific amino acid and tRNA substrate pair. Thus it was long supposed that all aaRSs had a common evolutionary root at least for their ATP-binding catalytic domain.

However, standard primary sequence analyses (Eriani *et al.*, 1990; Nagel and Doolittle, 1991; Cusack *et al.*, 1991) as well as comparisons of crystal structures of aaRSs complexed with cognate tRNAs (Rould *et al.*, 1989; Cusack *et al.*, 1991a, b; Ruff *et al.*, 1991) unambiguously revealed two distinct 'mutually exclusive' classes

* Address for correspondence.

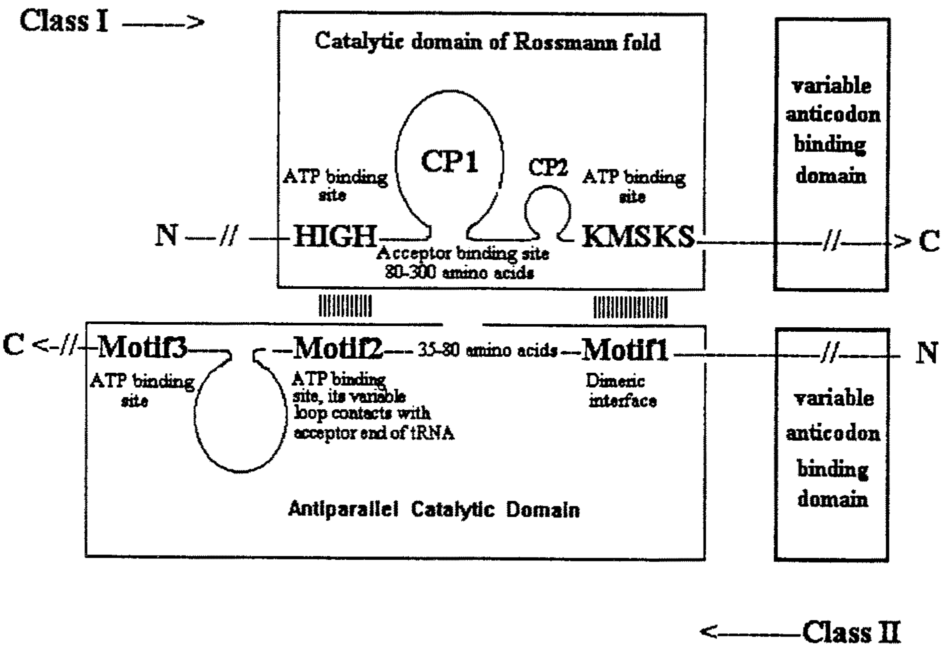
Origins of Life and Evolution of the Biosphere **25:** 565–589, 1995.

TABLE I
The genetic code table with indicated assignment of aminoacyl-tRNA synthesases to class I (boxed amino acids) and class II (shaded amino acids)*

	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	G
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	G
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	G
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	G

* adapted from (Carter, 1993)

of 10 synthetases each (Table I). In class I aaRSs, two short characteristic 'signature' sequences were identified: the first segment of 11 residues or so long that ended in the common tetrapeptide HIGH and the second segment that included the common pentapeptide KMSKS (for short we shall refer to these regions simply as the HIGH and KMSKS motifs). These two highly conserved motifs are contained in the catalytic domain of the Rossmann fold, a widely observed nucleotide binding region based on a six-stranded parallel β-sheet split in the middle by the quantitatively and qualitatively hypervariable connective polypeptides CP1 and CP2 (Figure 1). The HIGH itself is located in the N-terminal half of the fold, after the first β-strand whereas the KMSKS is in a loop following the C-terminal β-strand in the second half of the fold. Both HIGH and KMSKS motifs contribute to the ATP-binding site, the glycine G of the former and the distal lysine K of the latter being particularly important (Hou *et al.*, 1991; Carter, 1993). In C-proximity to HIGH is the site for activation of the specific amino acid, as was shown for *E.coli* MetRS and IleRS (Burbaum and Schimmel., 1992). Crossover CP1 links the N-terminal outer β-strand to the central β-strand of the C-terminal half and specifically interacts with the 3′CCA end of tRNAs (Perona *et al.*, 1991). Six class I aaRSs also have the short CP2 with presumably hydrolytic proofreading functions (Eriani *et al.*, 1991).



Catalytic domain of Rossmann fold

variable anticodon binding domain

—Motif2— 35-80 amino acids —Motif1

ATP binding site, its variable loop contacts with acceptor end of tRNA

variable anticodon binding domain
Antiparallel Catalytic Domain

Motif 1: + G Φ x x Φ x x **P** Φ Φ

variable loop
Motif 2: + Φ Φ ∓ Φ x x x F **R** x E/D ....∇..........+ Φ x x - F x x x - Φ x Φ Φ

Motif 3: G Φ G Φ G Φ **E R** Φ Φ Φ Φ Φ

Fig. 1. Schematic simplified representation of class I versus class II aaRSs antiparallelly mapped to each other. Class-defining signature motifs are given such that the signature HIGH motif of class I stand against motif 2 of class II and KMSKS against motif 1. CP – hypervariable in size connective polypeptides. Class I N-terminal sequences preceding the HIGH motif are relatively short to cover C-terminal part of class II aaRSs including the distal part of motif 2, motif 3 and variable intermediate sequence. Under the scheme are shown the three signature motifs of class II aaRSs. Diagnostic for all class II enzymes is the motif 3, while AlaRS has putative motifs 1 and 2 conserved only at the secondary structure level (Ribas de Pouplane *et al.*, 1993; Shi *et al.*, 1994) which are not still reliably identified in GlyRS. Here and all further figures, amino acids are given in the conventional single-letter symbols: L – leucine, I – isoleucine, M – methionine, V – valine, Y – tyrosine, Q – glutamine, E – glutamic acid, C – cystein, R – argnine, F – phenylalanine, S – serine, P – proline, T – threonine, A – alanine, H – histidine, N – asparagine, D – aspartic acid, K – lysine, and G – glycine. Amino acid groups are designated: x – any residue, (∓) – polar residue, (+) – positively charged residue, (—) – negatively charged residue, ϕ – hydrophobic residue. Bold are invariant amino acids.

In class II aaRSs, neither segments bearing even a slight resemblance to HIGH or KMSKS signature motifs nor Rossmann-like folds were found. Instead, three signature motifs of their own were identified (see Figure 1). All three of these are also contained in the catalytic domain of the new fold built around a six-stranded antiparallel β-sheet flanked by 3 α-helices. Motif 1 nucleates the dimer interface, probably involving cooperative interactions between two ATPs and adenines of CCA ends of tRNAs (Ruff *et al.*, 1991). All invariant residues of motifs 2 and 3 form a conserved active site for the binding of ATP and 3′CCA end of the tRNA, while the variable loop within the motif 2 and the residues situated more to the right of the active site recognize the amino acid (Eriani *et* al., 1990; Cusack *et al.*, 1991; Ruff *et al.*, 1991).

Activated aminoacyl may attach to the tRNA either on the 2′OH or the 3′OH group of the tRNA terminal ribose. Intriguingly, the enzymes aminoacylating on the 2′OH end are all (except for PheRS) from class I and those specific for the 3′OH are all from class II (Eriani *et al.*, 1990; Ruff *et al.*, 1991). As one would expect, the anticodon-binding domains of both aaRS classes are rather more variable than those of their activation catalytic centers.

All attempts to identify even weak sequence similarity between the two classes proved futile. This was accepted as evidence of their independent origins in two archaic translation systems that merged later in the course of evolution (Eriani *et al.*, 1990; Nagel and Doolittle, 1991). Yet, the apparently independent origins of these two enzymes elicit a number of unsolvable problems, namely:

1. While it may seem reasonable for earliest life to have begun with the proteins made up solely of 10 amino acids activated by one aaRS class only (Eriani *et al.*, 1990), it follows that conserved motifs of a given aaRS class in this scenario should then consist predominantly of residues which are activated by the same class of aaRSs. This expectation is contradicted by the facts. The published aaRS alignments (Eriani *et al.*, 1990; Nagel and Doolittle, 1991; Cusack *et al.*, 1991a,b; Hou *et al.*, 1991) allow us to conclude that 67% and 70% of the complete HIGH- and KMSKS-containing signature motifs of class I aaRSs are comprised of residues that can only be activated by members of class II. In turn, more than half of the residues comprising three signature motifs of class II aaRSs can only be activated by class I.

2. The coevolutionary theory of the genetic code (Wong, 1975, 1980, 1981) postulates two major groups of amino acids, based on the connectedness of their biosynthetic pathways rather than the similarity of chemical properties. One group comprises Ile, Met, Thr, Lys, Asp and Asn whereas the other Pro, His, Glu, Gln and Arg so that each of the groups contains members of class I and class II. Moreover, amongst presumably earliest prebiotically synthesized proteinous amino acids, i.e. Gly, Ala, Ser, Asp, Glu, Val, Leu, Pro, and Thr (Wong, 1981) are members of the two opposite classes. Accordingly, also activated by the two ‘exclusive’ enzymes are those amino acids which have likely entered in translation later as products of

the 'inventive biosynthesis' (*ibid.*), i.e. Arg, His, Met, Trp, Asn, Gln, Lys, Phe, Tyr, and Cys.

3. Alternative theories of the origin of the genetic code with the starting tRNA synthetases made of ribozymes (Weiner and Maizels, 1987; Szathmary, 1993) imply that basic amino acids like Arg, Lys, and His were first ones which could produce short basic peptides preferentially interacting with the negatively charged backbone of RNAs. It was shown that ribozymes do selectively bind positively charged Arg (Yarus, 1988). However, this scenario is again inconsistent with the hypothesis of two independent aaRS classes since arginine is activated by the aaRS from class I, while the opposite class II enzymes activate lysine and hystidine.

4. Because of the abundance of palindromes, complementary base triplets occur in RNA and both strands of DNA with nearly equal frequencies. This is very likely an ancient feature which antedated the origin of the genetic code (Ohno, 1991; Ohno and Yomo, 1991). However, a larger group of amino acids encoded by complementary triplets is activated by aaRSs from the opposite classes (Table I). Furthermore, the hypothesized archaic translation system which utilized only class I pre-aaRSs had to be completely unable to read the triplets with the central C base (Table I: 2nd column) while at the same time using C as the central anticodon base. In turn, the second independent system based on usage of exclusively class II aaRSs somehow had to 'ignore' 15 of 16 codons with the central U base, the only exception being Phe's codon, UUU (Table I: 1st column) while again using U as the central anticodon base.

5. Were the two classes of aaRSs of independent origin, their sequential apparition could be postulated. However, phylogenetic analysis did not reveal any reliable difference in the evolutionary age of the two classes (Nagel and Doolittle, 1991).

6. Two evolutionarily distinct classes of aaRS imply that the two sets of cognate tRNAs also have independent origins. However, all tRNAs have a number of invariant elements in their primary sequences and, consequently, share a uniform cloverleaf-like secondary and L-shaped tertiary folding, all these similarities pointing to a single common progenitor of tRNAs.

In view of the above, it would appear that the explanation of the origin of the two classes of aaRSs should be sought anew. Inasmuch as our previous studies indicated the simultaneous concerted origin of tRNAs with complementary anticodons (Rodin *et al.*, 1993a,b), we decided to see if the same held true in regard to the origin of two classes of aaRSs.

2. Methods

We analyzed the complete set of 20 *E. coli* aaRS genes and 25 genes of other procaryotic and eucaryotic species, all retrieved from NCBI GenBank (Release 80.0 December 1993). Class I was represented by 21 sequences, and Class II, by 24 sequences.

AaRSs are among the oldest proteins, each class showing considerable divergence in size and primary structure mainly in regions beyond their catalytic domains. Thus, because of the general limited similarity, even between isozymic aaRSs, the diagnostic signature motifs within the two catalytic domains were the only regions suitable for 'class 1 vs class 2' antiparallel comparisons. Furthermore, the signature motifs are not absolutely frozen and may contain segments of flexible length such as the insert between P and HIGH from class I (Figure 2) or the loop between halves of the motif 2 from class II (Figure 1). Therefore, to align all aaRSs from one class even within their conserved signature regions, one has to assume local short gaps (see Figures 2, 4). At any rate, as a starting step, we focused our analysis on the signature motifs and their nearest flanking sequences taking into account the secondary structures of the two enzymes (Cusack *et al.*, 1991b; Hou *et al.*, 1991). This approach seems justifiable also because both classes are thought to have originated from the minimalist precursors of the catalytic modules while more specific and variable parts of tRNA binding modules have been added later in evolution (Schimmel *et al.*, 1993; Moras, 1993).

For each signature motif, our analysis included a search for optimum sequence alignment, calculation of its average consensus sequence, then building its complementary (antisense) image and finally, a comparison of the latter with the signature motifs of the opposite class. In parallel, frequencies of codons TTA, CTA and TCA which are complementary to three stop codons in antisense strands were calculated in all 3 reading frames. A symmetrically equivalent approach began with construction of the antisense sequence for every individual signature gene fragment, the generation of their consensus sequences and finally, evaluation of the extent to which the latter is complementary to the sense signature motifs of the opposite class.

To test the results on randomness, we subjected the pairs of presumably complementary motifs to a jumbling procedure as described in (Doolittle, 1987), i.e. compared the authentic scores of differences between the signature sequences with the distributions of such scores obtained for their jumbled random derivatives of the same base and amino acid composition. To avoid 'fixed sequence randomization', we generated jumbles serially, i.e. jumble jumble 1 as jumble 2, then jumble jumble 2 as jumble 3, etc. (Doolittle, 1987). Two versions of the jumbling test were used, one for parallel alignments of class I sense and class II antisense sequences with the matrix of base mismatches $\| d(ij) \|$, where $d(ii) = 0$ and $d(ij) = 1$ $(i \neq j)$, and the other for antiparallel alignments of two sense sequences from the opposite classes with the matrix of base mispairings $\| c(ij) \|$, where $c(ij) = 0$ if bases i and j are complementary to each other (i.e. A-T, G-C, and R-Y), and $c(ij) = 1$ if they are not. For any pair including either at least one gap or arbitrary base X, both d(ij) and c(ij) values were set equal to 1.0. For each original pair of real consensus sequences (aligned either in parallel or in antiparallel) as well as of their jumbled versions, the corresponding direct or complementary distance was calculated simply as an additive sum of the individual distance values d_i or c_i, i.e. $D = \sum d_i$

or $C = \sum c_i$, where 'i' is the alignment position number, i = 1, 2, ..., N; N is a length of the aligned sequences. Finally, the generated distribution of the distances between randomized motifs was compared with the distance shown by 'real' pairs of consensus motifs to evaluate a statistical significance of their complementarity. For additional control, we have also jumbled amino acid sequences of these class-defining motifs. Although the sense-antisense transformations of the genetic code (Zull and Smith, 1990; Konecny *et al.*, 1993) allows to calculate the complementary distance between antiparallelly aligned polypeptides, we have used a more simple approach: all pairs of randomized consensus amino acid sequences were aligned in parallel but because one sequence was sense while the other antisense, we calculated the 'routine' direct distance on the basis of the five-level amino acid class hierarchy exactly as described in (Smith and Smith, 1990) with the values of the difference matrix || d(ij) || range from d(ii) = 0 (for identical residues) to d(ij) = 3 (for pairs including at least one 'wild-card' amino acid of any type, X) and maximum d(ij) = 6 (for gaps).

Accordingly, we used the VOSTORG package of programs (Zharkikh *et al.*, 1991) designed for the processing of sequence data: the routine ALSEQ based on the Needleman-Wunch algorithm for aligning a pair of sequences in combination with the multiple alignment and manual editing programs (MUTAL and SEQUED), and SEQU routine for building the complementary (−) strands. Also, the improved versions of CONSENS and JUMBLE programs (propietary of A. Rodin) compatible with VOSTORG format were used for building consensus sequences and testing the observed local sequence complementarity as caused by chance.

3. Results

3.1. Class I versus class II comparisons of the signature sequences aligned in antiparallel to one another

3.1.1. *HIGH Signature Sequence of Class I Versus Motif 2 of Class II*

Figure 2 shows the alignment of class I genes of aaRSs within the HIGH region. The alignment has two key sites occupied by the HIGH and conservative proline (P). The latter plays an essential role in aminoacylation (Burbaum and Schimmel, 1992) and is common for all class I aaRSs except for CysRS of *E. coli* which has no proline in an upstream site proximal to the HIGH tetrapeptide. Instead, either Ile as in (Burbaum and Schimmel, 1992) or the two-residue distally located Val, as in Figure 2, might be mapped at the proline's site. This P precedes the HIGH by 4–8 amino acids in different synthetases so that the local gaps of 1-2-residue-long should be introduced to allow a reasonable alignment within the variable interval between P and HIGH. The result is rather close to the alignment built by Nagel and Doolittle (1991). Indicated above every individual gene fragment is the deduced polypeptide sequence. Immediately under the alignment are shown consensus nucleotide and amino acid sequences.

```
A:              5'----->3'

                  G   I   T | V |$  Y   $   $   D   L   C | H   I   G   H | G
CYS.ec          -GGA-ATC-ACC|GTG|  TAT---------GAT-CTC-TGT|CAT-ATC-GGT-CAC|GGG-
                  G   A   Q | P |           S   G   E   L | T   I   G   N | Y
TRP.ec          -GGC-GCA-CAG|CCC|  --------TCA-GGT-GAA-TTG|ACC-ATT-GGT-AAC|TAC
                  G   F   D | P |   T       A   D   S   L | H   L   G   H | L
TYR.ec          -GGC-TTC-GAT|CCT|  ACC-----GCT-GAC-AGC-TTG|CAT-TTG-GGC-CAT|CTT
                  G   F   D | P |   T       A   D   S   L | H   I   G   N | L
TYR.bc          -GGG-TTT-GAC|CCG|  ACG-----GCG-GAC-AGT-TTG|CAT-ATC-GGG-AAC|TTG
                  M   I   P | P |P  N   V   T   G   S   L | H   M   G   H | A
VAL.ec          -ATG-ATC-CCG|CCG|\/AAC-GTG-ACC-GGC-AGT-TTG|CAT-ATG-GGT-CAC|GCC
                  V   I   P | P |   N   V   T   G   K   L | H   L   G   H | A
VAL.bs          -GTC-ATC-CCG|CCG|  AAC-GTC-ACC-GGC-AAA-TTG|CAC-TTG-GGG-CAT|GCG
                  D   G   P | P |   Y   A   N   G   S   I | H   I   G   H | S
ILE.ec          -GAT-GGC-CCT|CCT|  TAT-GCG-AAT-GGC-AGC-ATT|CAT-ATT-GGT-CAC|TCG
                  D   G   P | P |   Y   C   S   G   R   I | H   L   G   T | A
ILE.mt          -GAC-GGC-CCA|CCC|  TAC-TGC-AGC-GGA-AGG-ATT|CAC-CTT-GGA-ACA|GCA
                  C   A   L | P |   Y   A   N   G   S   I | H   L   G   H | M
MET.ec          -TGC-GCA-CTG|CCG|  TAC-GCT-AAC-GGC-TCA-ATC|CAC-CTG-GGC-CAT|ATG
                  V   T   T | P I,Y Y   V   N   A   E   P | H   L   G   H | A
MET.tt          -GTG-ACC-ACC|CCC|\/TAC-GTG-AAC-GCC-GAG-CCG|CAC-CTG-GGC-CAC|GCC
                  S   M   L | P |   Y   P   S   G   R   L | H   M   G   H | V
LEU.ec          -TCT-ATG-CTT|CCC|  TAT-CCT-TCT-GGT-CGA-CTA|CAC-ATG-GGC-CAC|GTA
                  R   F   P | P |   E   P   N   G   Y   L | H   I   G   H | A
GLN.ec          -CGT-TTC-CCG|CCG|  GAA-CCG-AAT-GGC-TAT-CTG|CAT-ATT-GGC-CAT|GCG
                  R   F   A | P |   S   P   T   G   Y   L | H   V   G   G | A
GLU.ec          -CGC-TTC-GCG|CCG|  AGC-CCA-ACA-GGC-TAT-CTG|CAC-GTT-GGC-GGC|GCG
                  Y   S   A | P |   N   V   A   K   E   M | H   V   G   H | L
ARG.ec          -TAC-TCT-GCG|CCA|  AAC-GTG-GCG-AAA-GAG-ATG|CAT-GTC-GGT-CAC|CTG

---------------------------- --- ------------------------- --------------- ---
                      R     |   |   T                   A |         G     |
B:              -xxx-TTC-xxx|CCG|  AAC- $ - $ -GGC-xxx-YTG|CAY-ATT-GGY-CAY|GYC
                            |   |   ↓                     |               |
C:                          |   |   A                     |               |
D:                X  F/I  X | P | N/Y  $   $   G   X  L/M| H  I/M  G   H |A/V
                                   (K)        +
                                              ↓
                                              -

                                                                    3'<-----5'
                      T             A                   T           C
E:              -xxx-AAG-xxx-GGC---TTG- $ - $ -CCG-xxx-RAC-GTR-TAA-CCR-GTR-CRC-
                                    ↓
F:                                  T
G:                X [E/D  X   R     V  ]        A   X  Q/H[M/V N/H A/T V/M]R/H
                    [             (F/L)]

H:                X [E/D  X   R     F  ]        X   X  (∓)[ Φ  (∓)  Φ   Φ ](+)
```

Fig. 2. Consensus signature HIGH sequence and its complementary image: **A**: Aligned within the HIGH region are the complete set of 10 *E. coli* (ec) sequences and five representatives of other species, namely: bc – *Bacillus caldotenax*, bs – *Bacillus Stearothermophilus*, mt – *Methanobacterium thermoautotrophicum*, ss – *Saccharomyces cerevisiea*, and tt – *Thermus thermophilus*. As in (Nagel and Doolittle, 1991), some sequences contain short local deletions and insertions (designated by symbol $) within the variable region between P and HIGH. Above each sequence of codons is shown the deduced polypeptide sequence in single-letter abbreviations of amino acids (see Fig. 1). Bold are conservative residues. **B**: (+) Consensus sequence of the above alignment, naturally derived from only the complete set of *E. coli* synthetases. When two bases occurred with equal (or nearly equal) frequencies, either both of them or R (purines) and Y (pyrimidines) were assigned to such positions. **C**: Assumed base substitutions. **D**: The deduced consensus polypeptide with the mutant residue (K) given in brackets. The vertical arrow in the center shows the complementary transformation of (+) consensus sequence into its (–) counterpart (**E**). The symmetric base substitutions and consensus (–) polypeptide are indicated in (**F**) and (**G**) respectively. For comparison with (**G**), on the bottom is shown the consensus motif 2 of class II aaRSs (**H**). Boxed are the most conservative fragments of the compared signature motifs. Amino acid groups are designated as in Figure 1.

TrpRS of *E. coli* has the shortest variable segment of 4-residue-long between P and HIGH thus determining the minimum HIGH-containing a 12 residue long signature motif (Figure 2). The mirror image of the consensus polypeptide of this minimum length obviously resembles the common motif 2 of class II aaRSs (Figure 2). Symmetrically, Figure 3 shows how, in turn, the consensus motif 2 of the class II enzymes is complementarily converted into the unambiguously HIGH-like sequence. Though not long, this mirror correspondence looks rather convincing. In fact, for the alignment in Figure 2 only one base transversion **C** → **A** at the 3rd position of the codon which follows the conservative proline P is needed to generate the complementary phenylalanine F and thus convert the entire first signature HIGH motif of class I into the first half of signature motif 2 from class II. The conservative proline P of class I aaRSs turns into the invariant arginine R of the common FRXE tetrapeptide in the class II motif 2.

With eloquent symmetry (see Figure 3), among class II aaRSs, transversions at the 3rd position of only two codons allow for the common part of motif 2 preceding its variable loop to convert into the HIGH-like signature motif of class I. Thus, by complementary transformation not only does the conservative tetrapeptide HIGH generate its mirror counterpart $\phi\phi\mp\phi$ of the opposite synthetase but more degenerate $\phi\phi\mp\phi$ in its turn quite faithfully reproduces the HIGH. The same is true for the other pair of mutually complementary tetrapeptides: that of class I with the invariant proline (Figure 2: F/Ix**PN**/Y) and its symmetrical counterpart with the invariant arginine (Figure 3: F**R**xE/D) from class II synthetases. Accordingly, even direct antiparallel comparisons of individual aaRS genes still show an impressive complementarity between the HIGH-containing motif of class I and the second motif of class II, base mispairings being notably concentrated at 1st and 3rd positions of codons (Figure 4a).

3.1.2. *KMSKS Signature Sequence of Class I Versus Motif 1 of Class II*

Figures 5 and 6 exhibit the same mirror relationship between consensus KMSKS-containing sequence of class I and motif 1 of class II. Only one transversion **C** → **G** at the 3rd position of the most frequent glycine codon GGC preceding the KMSKS pentapeptide by one or two residues is required in order to obtain as a complementary partner the motif 1 of class II aaRSs (Figure 5). As far as the class I signature pentapeptide KMSKS itself is concerned, it accurately converts into the complementary GFx(∓)(V/L/I) pentapeptide shared by all class II synthetases. For whole alignment in Figure 6, the single dissonance is associated with a residue following the KMSKS pentapeptide: the ideal complementarity dictates that hydrophobic amino acids (F, L, M, V, A) are welcome in this position as complementary to positively charged residues (H, K or R) with which the motif 1 of class II usually begins. While in most other species a hydrophobic leucine appears here, in *E. coli* KMSKS motifs, this position is occupied by a polar residue, except for CysRS and ArgRS.

```
A:                5'--->3'
                    E   R | V   F   E   I | N   R   N | F   R   N   E | G
LYS.ec            -GAA-CGC|GTA-TTC-GAA-ATC|AAC-CGT-AAC|TTC-CGT-AAT-GAA|GGT-
                    D   R | V   Y   E   I | G   R   Q | F   R   N   E | G
LYS.ss            -GAT-CGT|GTT-TAC-GAA-ATT|GGT-AGA-CAA|TTC-AGA-AAT-GAA|GGT-
                    D   R | Y   Y   Q   I | V   K   C | F   R   D   E | D
ASP.ec            -GAC-CGT|TAC-TAT-CAG-ATC|GTT-AAA-TGC|TTC-CGT-GAC-GAA|GAC-
                    E   K | V   F   C   I | G   P   V | F   R   A   E | D
ASP.rn            -GAG-AAG|GTG-TTC-TGC-ATT|GGG-CCA-GTT|TTC-AGA-GCT-GAA|GAC-
                    Q   R | L   W   Y   I | G   P   M | F   R   H   E | R
HIS.ec            -CAG-CGT|CTG-TGG-TAT-ATC|GGG-CCG-ATG|TTC-CGT-CAC-GAG|CGT-
                    I   R | I   I   A   P | G   R   V | Y   R   N   D | Y
PHE.ec            -ATT-CGT|ATC-ATC-GCG-CCT|GGC-CGT-GTT|TAT-CGT-AAC-GAC|TAC-
                    S   K | I   F   T   F | G   P   T | F   R   A   E | N
ASN.ec            -TCC-AAA|ATT-TTT-ACC-TTC|GGC-CCG-ACT|TTC-CGT-GCT-GAA|AAC-
                    L   R | M   A   E   F | G   S   C | H   R   N   E | P
THR.ec            -CTG-CGT|ATG-GCC-GAG-TTT|GGT-AGC-TGC|CAC-CGT-AAC-GAG|CCG-
                    I   R | M   A   E   F | G   H   V | H   R   H   E | Y
THR.bs            -ATT-CGC|ATG-GCC-GAA-TTC|GGA-CAC-GTG|CAC-CGC-CAT-GAA|TAC-
                    L   N | F   Y   Q   I | Q   T   K | F   R   D   E | V
PRO.ec            -CTG-AAC|TTC-TAT-CAG-ATC|CAG-ACC-AAG|TTC-CGC-GAC-GAA|GTG-
                    I   K | M   T   A   H | T   P   C | F   R   S   E | A
SER.ec            -ATT-AAG|ATG-ACC-GCC-CAC|ACC-CCA-TGC|TTC-CGT-TCT-GAA|GCC-
                  -------- --------------- ----------- --------------- ----
                       A                       A
B:                -xxx-CRT|ATG-TYC-GAx-ATY|GGC-CCx-xxx|TTC-CGT-RAY-GAA|xxx-
                          |              ↓|           |      ↓        |
C:                        |              G|           |      G        |

D:                  X  R/∓| Φ   Φ   ∓   Φ | G   X   X | F   R   ∓   E | X
                                       (M)                 (R)

                                            +
                                            ↓
                                            -
                       T                       T                 3'<---5'
E:                -xxx-GYA-TAC-ARG-CTx-TAR-CCG-GGx-xxx-AAG-GCA-YTR-CTT-xxx-
                                         ↓                   ↓
F:                                       C                   C

G:                  X   Φ [ H  E/G  Φ  D/N] A  X   X[  E   T  V/I   F ] X
                                       (H)                 (P)

H:                  X   Φ [ H   G    Φ   H ] Φ   X...X[N/Y/E P   X   F/I] X
```

Fig. 3. Consensus motif 2 of class II aaRSs and its complementary image. The motif 2 of AlaRS as too diverged is not included. The sequence alignment is identical to that in (Eriani *et al.*, 1990, Cusack *et al.*, 1991). All designations are the same as in Figs. 1 and 2. rn – *Rattus norvegicus.*

The symmetrical scheme in Figure 6 displays the reverse complementary transfer from the alignment of class II motifs 1 and its consensus to the common class I KMSKS signature motif. To secure exact correspondence between these two motifs, 4 base substitutions have to be assumed. This is not surprising, because compared to the KMSKS motif of class I aaRSs, its presumably complementary image, the motif 1 of class II synthetases is more variable. In particular, the central

```
                          A:

GlnRS(E.coli)
5'--->3'
   R  |  F   x   P   E  |       |  H   I   G   H  |  A
 -CGT | TTC-CCG-CCG-GAA | -...- | CAT-ATT-GGC-CAT | GCG-
  ||  | |||      || |   |       |  |   || ||| ||| | |||
 -GCC | GAG-CAA-TGC-CAC | -...- | TTT-GAG-CCG-GTA | TGC-
   P  | E/D  x   R   H? |       |  F   E   A   M  | R(+)
                    CTT                      3'<---5'
                     F                 ThrRS(E.coli)
```

```
                          B:

ValRS(R.norvegicus)
5'--->3'
  H   G   R   x  |  K   M   S   K   S  |  L       G   N
-CAT-GGC-CGC- $  | AAG-ATG-AGC-AAA-TCT | CTC--$--GGC-AAT-
 ||| ||   |      | ||   |   |  |||  || |  |      |||  ||
-GTA-TCC-CCA-AAG | TTG-AAG-GTA-TTT-CGG | CGC-GCG-CTG-GTA-
  M   P   T   E  |  V   E   M   F   G  | R(+) A   V   M
                                                 3'<---5'
                                          LysRS  (E.coli)

ValRS(R.norvegicus)
5'--->3'
  H   G   R   x  |  K   M   S   K   S  |  L       G   N
-CAT-GGC-CGC- $  | AAG-ATG-AGC-AAA-TCT | CTC--$--GGC-AAT-
 ||   |  ||      | ||   ||  |  ||| ||| | ||      |   ||
-GTC-ACC GCA-ACT | TTG-GGT-CTT-TTT-AGG | GAC-GAG-CAA-TTT-
  L   P   T   S  |  V   W   F   F   G  |  Q   E   N   F
                                                 3'<---5'
                                          AsnRS  (E.coli)
```

Fig. 4. **A**: Antiparallel comparison of the *E. coli* class I GlnRS signature HIGH motif versus the *E. coli* class II ThrRS signature motif 2. The illegitimate pairings between bases G and T are also allowed. This example contains the single mispairing between central bases of the hystidine codon CAC and withstanding codon GAG for glutamic acid. But, CAC is exotic rather than common codon at this location: most of class II genes contain here the phenylalanine's TTC codon which is an ideal complementary partner of glutamic acid with its codons GAA and GAG. **B**: The same two comparisons as (**A**) but involving the *Rattus norvegicus* class I ValRS signature KMSKS motif versus the *E. coli* class II LysRS and AsnRS signature motif 1.

```
A:       5'--->3'
          V   D   R   $   E | K   M   S   K   S | L   $   G   N   $   F   F   T
CYS.ec   GTT-GAC CGG ----GAG|AAG-ATG-TCC-AAA-TCG|CTG-----GGT-AAC-----TTC-TTT-ACC
          V   N   G       A | K   M   S   K   S | R       G   T       F   I   K
MET.ec   GTG-AAC GGC ----GCA|AAG-ATG-TCC-AAG-TCT|CGC-----GGC-ACC-----TTT-ATT-AAA
          P   D   G       R | K   M   S   K   T | L       G   N       V   V   D
MET.tt   CCG-GAC GGC ----CGC|AAG-ATG-TCC-AAG-ACC|CTG-----GGG-AAC-----GTC-GTC-GAC
          Y   T   G   M   S | K   M   S   K   S | K       N   N       G   I   D
LEU.ec   TAT-ACC GGC ATG-AGC|AAA-ATG-TCC-AAG-TCG|AAG-----AAC-AAC-----GGT-ATC-GAC
          D   E   G       Q | K   M   S   K   S | K       G   N       V   I   D
VAL.ec   GAC-GAA GGC ----CAG|AAG-ATG-TCC-AAA-TCC|AAG-----GGT-AAC-----GTT-ATC-GAC
          A   Q   G       R | K   M   S   K   S | L       G   N       G   V   D
VAL.bs   GCC-CAA GGG ----CGG|AAA-ATG-AGC-AAG-TCG|CTC-----GGC-AAC-----GGC-GTC-GAT
          A   H   G       R | K   M   S   K   S | L       G   N       V   I   D
VAL.rn   GCT-CAT GGC ----CGC|AAG-ATG-AGC-AAG-TCT|CTC-----GGC-AAT-----GTC-ATC-GAC
          A   D   G       R | K   M   S   K   S | L       K   N       Y   P   D
ILE.ec   GCC-GAT GGT ----AGA|AAG-ATG-TCT-AAA-TCC|TTG-----AAA-AAT-----TAC-CCT-GAT
          E   E   G       R | K   M   S   K   S | L       G   N       V   V   E
ILE.mt   GAA-GAG GGC ----AGG|AAG-ATG-AGC-AAG-TCC|CTG-----GGG-AAC-----GTT-GTT-GAA
          D   D   G       K | K   L   S   K   R | H   G   A   V   S   V   M   Q
GLU.ec   GAT-GAC GGT ----AAA|AAA-CTG-TCC-AAA-CGT|CAC-GGG-GCA-GTC-AGC-GTA-ATG-CAG
          N   L   E   Y   T | V   M   S   K   R | K   L   N   L   L   V   T   D
GLN.ec   AAT-CTG GAA TAC-ACC|GTC-ATG-TCC-AAG-CGT|AAG-TTG-AAC-CTG-CTG-GTG-ACC-GAC
          A   D   G       T | K   F   G   K   T | E   G   G   A   V   C   L   D
TYR.ec   GCA-GAT GGC ----ACC|AAA-TTT-GGT-AAA-ACT|GAA-GGC-GGC-GCA-GTC-TGC-TTG-GAT
          S   S   G       A | K   F   G   K   S | A   G   N   A   I   W   L   D
TYR.nc   TCC-TCC GGT ----GCC|AAG-TTT-GGT-AAG-AGT|GCG-GGT-AAC-GCC-ATT-TGG-CTT-GAC
          L   E   P   T   K | K   M   S   K   S | D   D   N   R   N   N   V   I
TRP.ec   CTG-GAG CCG ACC-AAG|AAG-ATG-TCC-AAG-TCT|GAC-GAT-AAT-CGC-AAT-AAC-CTG-ATC
          K   D   G       K | P   F       K   T | R   A   G   G       T   V   K
ARG.ec   AAA-GAC GGC ----AAA|CCG-TTC-----AAA-ACC|CGC-GCG-GGT-GGT-----ACA-GTG-AAA

         ───────────────────|───────────────────|───────────────────────────────
                            |                   |RA
B:       GAx-GAC GGC xxx-xxx|AAG-ATG-TCC-AAR-TCY|CTG-xxx-GGC-AAC-xxx-GTY-ATY-GAC
                  ↓         |                   |
C:                G         |                   |
D:       D/E  D   G   X   X | K   M   S   K   S |Φ/∓? X   G   ∓   X   Φ   Φ   ∓
                 (G)                      +
                                          ↓
                                          -

                                                 YT                      3'<---5'
E:       CTx-CTG-CCG-xxx-xxx-TTC-TAC-AGG-TTY-AGR-GAC-xxx-CCG-TTG-xxx-CAR-CAR-CTG
                  ↓
F:                C
        ┌──────────┐        ┌───────────────────┐                  ┌────────────┐
G:      │ Φ   V   A│  X   X │ L   H   G   F  G/R│Φ/∓? X   A   V   X│D/N D/N   V │
        │       (P)│        └───────────────────┘                  └────────────┘
        └──────────┘

        ┌──────────┐        ┌───────────────────┐                  ┌────────────┐
H:      │ Φ   Φ   P│  X   X │V/L/I ∓  X    F/Y G│(+)  X   X   Φ   X│  ∓   ∓   Φ │
        └──────────┘        └───────────────────┘                  └────────────┘
```

Fig. 5. Consensus signature KMSKS sequence and its complementary image. As in Fig. 2, a few local gaps were introduced to allow the optimum alignment. The similar gaps were assumed by Nagel and Doolittle (1991). All designations are the same as in Figs. 2 and 3. nc – *Neurospora crassa*. The question mark indicates the ambiguous residue (polar or hydrophobic) that follows the pentapeptide KMSKS.

residue x in the GFx(∓)(V/L/I) appears too arbitrary to reproduce as its complementary partner the central serine residue in the KMSKS. Nevertheless, both the proximal and especially distal lysines K in KMSKS pentapeptide do look complementarily related with the hydrophobic V (or I) and F, respectively (Figure 6). Also,

```
A:      5'--->3'
         I   R   Q  | F   M   V   A   R  | G   F   M   E   V  | E   T  | P   M   M
LYS.ec  ATC-CGT-CAA |TTC-ATG-GTC-GCG-CGC |GGC-TTT-ATG-GAA-GTT |GAA-ACC |CCT-ATG-ATG
         I   R   R  | F   L   D   Q   R  | K   F   I   E   V  | E   T  | P   M   M
LYS.ss  ATC-AGA-AGA |TTT-TTG-GAC-CAA-AGA |AAG-TTT-ATT-GAA-GTA |GAA-ACT |CCG-ATG-ATG
         V   R   R  | F   M   D   D   H  | G   F   L   D   I  | E   T  | P   M   L
ASP.ec  GTG-CGC-CGT |TTT-ATG-GAT-GAC-CAC |GGC-TTC-CTC-GAC-ATC |GAA-ACT |CCG-ATG-CTG
         F   R   E  | T   L   I   N   K  | G   F   V   E   I  | Q   T  | P   K   I
ASP.rn  TTC-CGA-GAA |ACG-TTA-ATT-ACC-AAA |GGT-TTT-GTG-GAA-ATC |CAG-ACT |CCT-AAA-ATA
         L   K   N  | V   L   G   S   Y  | G   Y   S   E   I  | R   L  | P   I   V
HIS.ec  CTG-AAA-AAC |GTG-CTC-GGC-AGC-TAC |GGT-TAC-AGT-GAA-ATC |CGC-TTG |CCG-ATT-GTA
         P   E   I  | E   D   D   Y   H  | N   F   D   A   L  | N   I  | P   G   H
PHE.ec  CCG-GAA-ATC |GAA-GAC-GAT-TAT-CAT |AAC-TTC-GAT-GCT-CTC |AAC-AAT |CCT-GGT-CAC
         L   H   R  | F   F   N   E   Q  | G   F   F   W   V  | S   T  | P   L   I
ASN.ec  CTG-CAT-CGC |TTC-TTT-AAC-GAG-CAG |GGA-TTT-TTC-TGG-GTT |TCA-ACG |CCA-CTG-ATT
         V   R   S  | K   L   K   E   Y  | Q   Y   Q   E   V  | K   G  | P   F   M
THR.ec  GTT-CGT-TCT |AAA-CTG-AAA-GAG-TAC |CAG-TAT-CAG-GAA-GTT |AAA-GGT |CCG-TTC-ATG
         S   R   E  | L   Q   T   N   A  | G   Y   D   E   V  | R   T  | P   F   M
THR.bs  TCG-CGT-GAG |CTG-CAA-ACA-AAT-GCC |GGC-TAC-GAT-GAG-GTG |CGT-ACG |CCC-TTT-ATG
         V   R   E  | E   M   N   N   A  | G   A   I   E   V  | S   M  | P   V   V
PRO.ec  GTG-CGT-GAA |GAG-ATG-AAC-AAC-GCC |GGT-GCG-ATC-GAG-GTG |TCG-ATG |CCG-GTG-GTT
         L   D   L  | H   T   E   Q   H  | G   Y   S   E   N  | Y   V  | P   Y   L
SER.ec  CTG-GAT-CTG |CAT-ACC-GAA-CAG-CAT |GGC-TAC-AGT-GAG-AAC |TAT-GTT |CCG-TAC-CTG
---------------------------------------------------------------------------------
        G                                                                 T
B:      CTG-CRT-xRY |xxx-xTx-RAY-RAx-CAC |GGY-TTY-RxY-GAR-RTT |xAx-AYT |CCG xTG-ATG
                    |                    |  ↓           ↓  ↓  ↓ |      |
C:                  |                    |  A           C  T  Y |      |

D:      L/V  ∓   ∓  | X   Φ   X   X   H  | G   F   X   E  V/I | X   Φ  | P   Φ  M/Φ
                                          (G)         (H)  (F)

                                                 +
                                                 ↓
                                                 -
                                                                          3'<---5'
        C                                                                    A
E:      GAC-GYA-xYR-xxx-xAx-YTR-YTx-GTG-CCR-AAR-YxR-CTY-YAA-xTx-TRA-GGC-xAC-TAC
                                           ↓           ↓  ↓  ↓
F:                                         T           G  A  A

G:      [Q/H M/T  Φ]   X   ∓   X   X   V  [ A  E/K  X  F/L  N]   X   X   [R   ∓   ∓]
                                          [(S)          (M) (K)]

H:      [ ∓   Φ   Φ]   X N(∓)  X   X   X  [ S   K   S  M/Φ  K]   X   X   [G/R ∓   ∓]
```

Fig. 6. Consensus motif 1 of class II aaRSs and its complementary image. The motif 1 of AlaRS are too diverged is not included. The sequence alignment is identical to that in (Eriani *et al.*, 1990, Cusack *et al.*, 1991). All designations are the same as in Figs. 2, 3 and 5.

supportive of this complementary partnership is the fact that although two base substitutions are needed to convert the consensus E from the GFxE(V/L/I) into the consensus M of the KMSKS, the precise complementary images of E are F and L, which are also often observed at the second position of KMSKS pentapeptide. Presented in Figure 4b are two examples of rather good complementarity which these signature motifs from the opposite classes (and even from very different species) show at the level of codons, again as in the previous case (Figure 4a) base mispairings occurring mostly in their 3rd position.

Antiparallel testing of the opposite aaRSs on mutual complementarity makes the identification of their signature motifs a bit more precise compared to their current definition. The next few specifications may serve as examples: (i) a hydrophobic residue preceding the HIGH (Figure 2) generates as a complementary partner a polar residue instead of an arbitrary one which follows the $\phi\phi\mp\phi$ in the motif 2 (Figure 3); (ii) similarly a hydrophobic residue usually follows the HIGH (Figure 2) and corresponds to positively charged residues (R or H) which the motif 2 generally starts from (Figure 3); (iii) a polar amino acid inside the $\phi\phi\mp\phi$ segment of the motif 2 (Figure 3) complementarily fits a hydrophobic residue between H and GH within the HIGH tetrapeptide (Figure 2) so that both of the tetrapeptides should be identified as $\phi\phi\mp\phi$ and HϕGH rather than $\phi\phi x\phi$ and HXGH or HIGH; etc.

3.1.3. *Jumbling Test*

Because the compared signature sequences are short, their observed mutual complementarity might be due to chance. To test the above, we repeatedly randomized (2000 to 5000 iterations) the signature sequences and for each of the jumbled sequence pairs calculated either the number of base mispairings between two sequences aligned in antiparallel or, analogously, the number of base mismatches between the sense strand of one sequence aligned in parallel with the antisense strand of the second sequence. In result, a distribution of 'jumbled distance values' was built and compared with the corresponding authentic value. We randomized the two separate pairs of signature motifs HIGH vs Motif 2 and KMSKS vs Motif 1, as well as their merged variant (HIGH + KMSKS) vs (Motif 2 + Motif 1), each in two representations, in nucleotide and amino acid sequences (the latter only in parallel alignment). Because the 1st and 3rd positions of codons were more variable, we also tested the simplified signature sequences comprised of only their second bases. In total, fifteen jumbling serial tests were carried out. Two of the tests are illustrated in Figure 7. In all of these tests, the result was essentially the same: authentic values of distance (indifferently, direct or complementary) were significantly less (deliberately beyond the 3-SD interval) than the mean values of the distributions generated by the corresponding jumbled sequences. Moreover, this result was insensitive to sequence insertions stretched up to 50 arbitrary triplets XXX, between the fixed (HϕGH vs motif 2) and (KMSKS vs motif 1) regions. Thus, at least these two pairs of 'mutually exclusive' signature motifs appear to be complementary with a high degree of confidence.

3.2. ON COMPLEMENTARITY OF AARSS BEYOND THEIR SIGNATURE REGIONS

The signature motifs of each class are common for most of its enzymes, whereas beyond the signature motifs, very low common sequence similarity is observed so that the two aaRS classes are partitioned further, into smaller groups of related sequences (Cusack *et al.*, 1991b). Since outside of the signature regions even par-

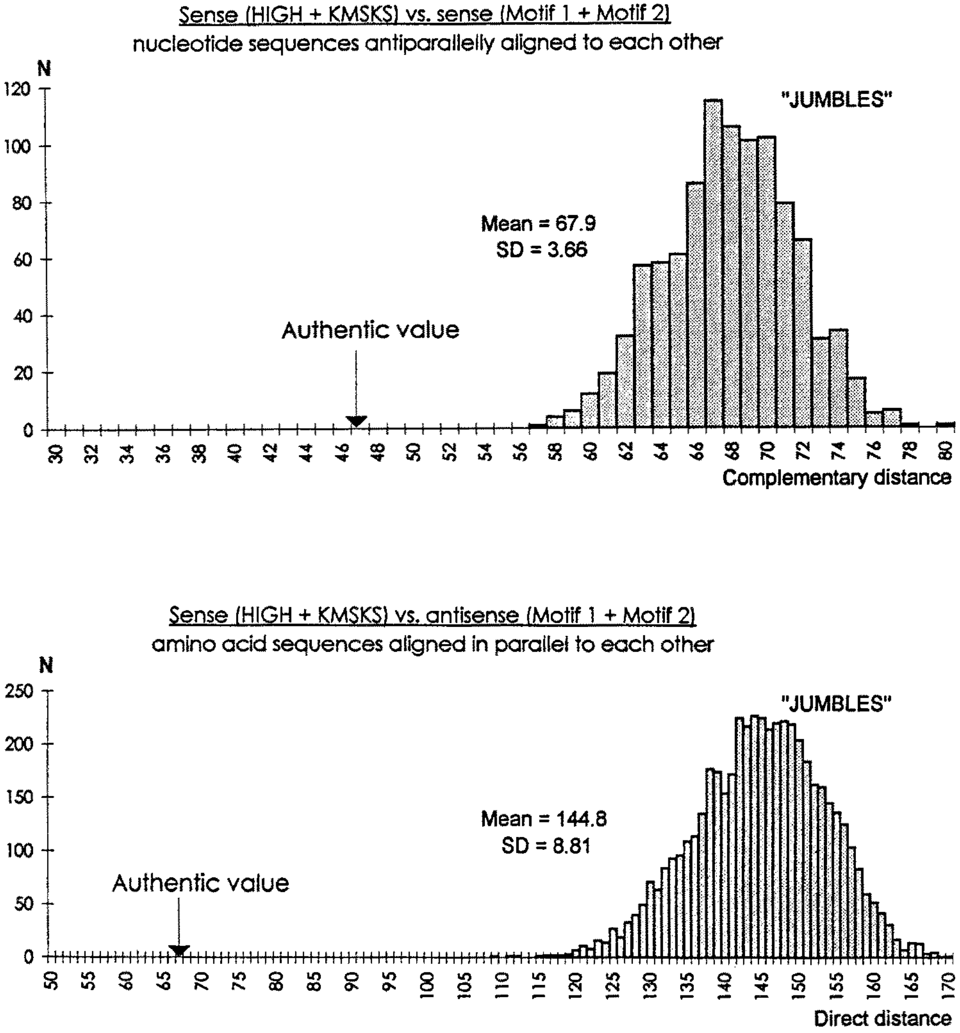


Fig. 7. Distribution of distance values generated by jumbled signature (HIGH + KMSKS) and (motif 2 + motif 1) sequences. Upper panel: the distribution of complementary distances obtained for randomized 99-base-long sequences, arranged in antiparallel. Lower panel: the distribution of direct distances obtained on the basis of "weighted" values of elementary differences between amino acids (see "Methods" and Fig. 1 in ref.: Smith and Smith, 1990) for randomized 33-residue-long sequences, derived from the merged signature (HIGH + KMSKS) and the inverse mirror image of the consensus (motif 2 + motif 1) sequence. In both these cases, the real class I vs class II comparison of signature sequences gives a value that appeared to be significantly less than those obtained with their jumbled derivatives ($P << 0.01$).

allel alignment for all members of class I or class II appears ambiguous, it would be surprising (if possible at all) to find extensive complementarity in the antiparallel sequence comparisons. As an example, Figure 8 shows the 'best' of our attempts to extend the sequence complementarity in immediate upstream and downstream proximity to the complementary signature motifs. Unlike the commonly charac-

```
5'-----> 3'
   -Loop->|<---------------- α-helix ----------------------->|<--Loop-->|<-β-
            K   S   I     C   L   N   F   G   I   A   Q   D   Y   K           G   Q
Gln ec    -AAA-TCT-ATC---TGC-CTG-AAC-TTC-GGG-ATC-GCC-CAG-GAC-TAT-AAA---------GGC-CAG
            R   T   A     L   Y   S   W   L   F   A   R   N   H               G   E
Glu ec    -CGT-ACT-GCT---CTT-TAC-TCC-TGG-CTT-TTT-GCA-CGT-AAC-CAC-------------GGC-GAG
            V   P   L     L   C   L   K               R   F   Q   Q   A       G   H
Tyr ec    -GTT-CCA-TTG---TTA-TGC-CTG-AAA-------------CGC-TTC-CAG-CAG-GCG-----GGC-CAC
            M   G   A     L   R   Q   W   V   K   M   Q   D   D   Y   H   C   I   Y
Trp ec    -ATG-GGT-GCA---CTG-CGT-CAG-TGG-GTA-AAG-ATG-CAG-GAT-GAC-TAC-CAT-TGC-ATT-TAC
            R   S   T     I   I   G   D   A   A   V   R   T   L   E   F   L   G   H
Arg ec    -CGC-TCT-ACC---ATT-ATT-GGT-GAC-GCA-GCA-GTG-CGT-ACT-CTG-GAG-TTC-CTC-GGT-CAC
            R   T   F     V   A   F   D   V   V   A   R   Y   L   R   F   L   G   Y
Cys ec    -CGT-ACC-TTT---GTT-GCT-TTT-GAC-GTG-GTT-GCG-CGC-TAT-CTG-CGT-TTC-CTC-GGC-TAT
            R   N   Y     T   I   G   D   V   I   A   R   Y   Q   H   M   L   G   K
Leu ec    -CGT-AAC-TAC---ACC-ATC-GGT-GAC-GTG-ATC-GCC-CGC-TAC-CAG-CAT-ATG-CTG-GGC-AAA
            L   E   H     I   Q   A   D   V   W   V   R   Y   Q   R   M   R   G   H
Met ec    -CTG-GAG-CAC---ATC-CAG-GCT-GAT-GTC-TGG-GTC-CGT-TAC-CAG-CGA-ATG-CGC-GGC-CAC
            Y   T   T     V   V   A   D   F   L   A   R   W   H   R   L   D   G   Y
Met tt    -TAC-ACG-ACG---GTG-GTG-GCC-GAC-TTT-CTG-GCC-CGG-TGG-CAC-CGC-CTG-GAC-GGC-TAC
            F   Q   Q     T   I   M   D   T   M   I   R   Y   Q   R   M   Q   G   K
Val ec    -TTC-CAG-CAA---ACC-ATC-ATG-GAT-ACC-ATG-ATC-CGC-TAT-CAG-CGC-ATG-CAG-GGC-AAA
            W   D   T     T   L   Q   D   I   I   T   R   M   K   R   M   Q   G   Y
Val bs    -TGG-GAT-ACG---ACG-CTG-CAA-GAC-ATC-ATT-ACG-CGC-ATG-AAG-CGC-ATG-CAA-GGG-TAT
            W   N   K     I   M   K   D   S   Y   L   R   F   K   S   M   R   G   F
Ile mt    -TGG-AAC-AAG---ATA-ATG-AAG-GAC-TCC-TAC-CTC-CGC-TTC-AAA-TCA-ATG-AGG-GGT-TTC
            L   A   S     T   I   K   D   I   V   P   R   Y   A   T   M   T   G   H
Ile sc    -CTT-GCT-TCC---ACT-ATT-AAG-GAT-ATT-GTT-CCA-AGA-TAC-GCT-ACC-ATG-ACA-GGC-CAC
            X   A   C     Y   X   A   A   Y   T   Y   G   A   A   R   T   X   G   A
                |?  |?    |       |?  |   |   |   |?      |   *   |?  *       |   |
            X   T   G     R   Y   T   T   R   A   A   X   T   A   C   C   X   C   T
            D   F   G     S   M   M   L   L   Q   K   F   L   Q   P   S   Q   P   L
Asp ec    -CAG-TTT-TGG---CCT-GTA-GTA-GTC-GTC-GAC-AAA-CTT-GTT-GAC-GCC-CCT-AAC-GCC-GTC
            E   F   G     G   V   V   L   R   K   L   Y   L   E   P   A   I   R   L
Lys ec    -AAG-TTT-TGG---CGG-ATG-TTG-GTC-TGC-AAA-GTC-TAT-GTC-GAG-GCC-GCG-CTA-TGC-GTC
            D   L   G     G   V   V   L   Q   K   L   F   L   E   P   A   I   R   M
Lys sc    -TAG-GTT-TGG---TGG-CTG-TTG-GTT-AAC-AAA-GTT-CTT-GTT-AAG-ACC-TCG-TTA-AGA-GTA
            E   F   G     G   V   I   L   R   K   L   Y   L   E   P   A   I   P   L
2b Lys cj -GAG-TTT-TGG---TGG-ATG-ATA-TTC-AGA-AAA-TTC-TAT-TTC-AAG-ACC-TCG-TTA-CCC-ATT
            S   L   A     C   A   Y   T   E   G   N   L   Q   G   S       V   T   L
Asn ec    -CCT-GTT-ACG---CGT-TCG-CAT-CCA-AAG-CGG-CAA-GTT-GAC-CGG-TCT-----ATG-CCA-GTC
            I   P   P     Q   A               K   M   T   R   I   Q   V   G   S   T
Phe ec    -TTA-GCC-ACC---GAC-CCG---------AAA-GTA-CCT-CGC-CTA-GAC-ATG-CGG-TCT-CCA
            T   P   K     K   A   L           D   H   L   M   R   A   S   I   A   T
Phe sc    -CCA-CCC-GAA---AAA-CCG-GTT-----TAG-CAC-GTC-GTA-AGA-CCG-TCT-CTA-CCG-ACA
            I   P   L  |  G   R   V   L   N   T   L   P   V   E   A   T   P   I   L
Ser ec    -TTA-ACC-GTC-∇-TGG-CGC-GTG-GTC-CAA-TCA-GTC-GCC-TTG-AAG-ACG-GCA-ACC-CTA-GTC
            L   P   L  |  N   R   I   L   D   T   I   V   E   E   H   T   P   G   L
Pro ec    -GTC-GCC-GTC-∇-CAA-TGC-TTA-GTC-CAG-TCA-CTA-TTG-AAG-AAG-TAC-TCA-ACC-CGG-CTC
            L   P   L  |  Q   N   F   I   Q   V   H   G   P   C   N   M   P   K   I
Thr ec    -GTC-GCC-GTC-∇-GAC-CAA-CTT-TTA-AAC-ATG-CAC-TGG-CCC-CGT-CAA-GTA-GCC-GAA-TTA
2a          I   P   L  |  N   K   F   I   L   M   H   G   P   C   N   M   P   K   M
Thr bs    -TTA-GCC-GTC-∇-CAA-AAA-TTT-CTA-CTC-GTA-CAC-CGG-CCC-CGT-TAA-GTA-GCC-GAA-GTA
            L   P   L  |  K   G   F   I   L   C   H   G   P   C   N   M   P   K   L
Thr sc    -ATC-CCC-TTC-∇-AAA-TGG-CTT-CTA-GTC-TGT-TAC-CGG-ACC-CGT-TAA-GTA-GCC-AAA-ATT
            Q   E   Q  |  L   T   L   S   D   G   N   R   D   E   F   T   Y   M   E
His ec    -GAC-AAG-GAC-∇-GTC-TCA-GTC-CGA-CAG-CGG-TAA-CGC-TAG-GAG-TTT-CCA-CAT-GTA-GAG
            β  | <------ Loop -------->|<--- α-helix--->|<----- Loop ---->|<-β -
                                                                       3'<-----5'
```

Fig. 8. Antiparallel comparison of downstream continuation of class I HIGH signature motif (Fig. 2) versus upstream continuation of class II motif 2 (Fig. 3). Alignment of class II sequences is taken from (Cusack *et al.*, 1991b). Boxed are the subclasses which were used to construct the consensus 2nd bases of codons shown in the center. Bold are the residues which have the consensus central base of codons. Complementary bases are connected by lines, an asterisk is inserted between unmatched bases, the question mark indicates unreliable consensus symbol. Italicized are aaRSs which share very low similarity with other aaRSs and only presumably referred to the subclass 2a (HisRS) and the subclass 2b (PheRS). In subclass 2a, a large fragment (shown by ∇) between 3rd and 4th residues was removed from the alignment to retain the secondary structure similarity of the subclass 2a and 2b (see: Cusack *et al.*, 1991b).

teristic motif 2 region, already its immediate upstream neighborhood differentiates all class II aaRSs on two main subclasses, designated 2a and 2b as in (Cusack *et al.*, 1991b). Similarly, seven aaRSs of class I in downstream neighborhood of the HϕGH motif form a distinct subclass, whereas three other synthetases for amino acids Gln, Glu and Tyr have quite different sequences with gaps. In fact, all the different aaRSs were aligned due to their more conservative secondary structures (Cusack *et al.*, 1991b). Accordingly, in the center of the antiparallel class I vs class II comparison (Figure 8) are indicated the evenly reduced 'average' sequences which consist of consensus second bases of codons, built only for the above group of 7 cia ,s I enzymes and the subclass 2b. However, even such simplified sequences at best have only 9 pairs of complementary bases of 18 pairs compared. Thus, although the jumbling tests (not illustrated) for this region do not exclude the original class I versus class II sequence complementarity, the latter does not look as impressive as that found between the signature sequences as such. As to the both 5′ and 3′ extensions of the KMSKS vs Motif 1 comparison, they did not show a complementarity even when narrowed to the 2nd base of the codonds.

According to the general Class I vs Class II antiparallel map (Figure 1), the distal half of motif 2 originally might complement an N-terminal segment that precedes the HϕGH motif in actual aaRSs. However, no conserved motifs are observed within the region and moreover its size varies greatly in different class I synthetases. For example, *E. coli* GluRS has only 5 residues preceding the HϕGH motif. In fact, when mapped 'head-to-tail', class I sequences lack sufficiently long sites to cover the distal part of motif 2 and further following motif 3 of class II enzymes. One may conclude that in class I precursors either such N-terminal domains once existed but then degenerated or, alternatively, the distal half of motif 2 and entire motif 3 emerged later in a common class II precursor.

3.3. Usage of Critical TTA, CTA and TCA Codons in AARS Genes

Because two DNA strands faithfully complement each other, the analysis of antisense aaRS sequences from the opposite classes resulted in the same mirror relationship between their consensus signature motifs. Only one detail is worth noting here. For any two proteins encoded by complementary DNA strands in the same reading frame, three codons – TTA (Leu), CTA (Leu) and TCA (Ser) – must not be used since they inversely complement stop codons TAA, TAG and TGA, respectively. If true, the complementary origin of the two synthetases requires that the above critical codons have been strictly suppressed in the corresponding two ancestral genes of class I and class II aaRSs. To test this, the two consensus genes being closest to the ancestors might be used. In turn, each of these two should be reconstructed from 10 individual consensus genes based upon as rich as possible spectrum of different species in order to reduce the smoothing influence of inter species divergence of aaRSs. The above was impossible: a complete set of aaRSs is available today only for *E. coli*, and only a few short signature motifs may more

TABLE II

Usage of codons TTA, CTA and TCA in aaRS genes[a]

	Expected	Observed						Base frequencies			
		RF-1		RF-2		RF-3					
aaRS	E	O	Z	O	Z	O	Z	A	T	G	C
Class I											
ArgRS ec								0.263	0.231	0.262	0.244
TTA	8.36	1	−2.6	5	−1.2	4	−1.5				
CTA	8.81	3	−2.0	6	−0.9	9	0.0				
TCA	8.81	0	−3.0	8	−0.3	13	+1.4				
CysRS ec								0.254	0.221	0.270	0.255
TTA	5.70	1	−2.0	6	+0.1	4	−0.7				
CTA	6.59	1	−2.2	1	−2.2	8	+0.6				
TCA	6.59	1	−2.2	5	−0.6	8	+0.6				
GlnRS ec								0.256	0.242	0.252	0.250
TTA	7.91	2	−2.2	12	+1.5	6	−0.7				
CTA	8.21	0	−2.9	7	−0.4	5	−1.1				
TCA	8.21	1	−2.5	12	+1.4	7	−0.4				
LeuRS ec								0.251	0.224	0.270	0.255
TTA	10.78	1	−3.0	19	+2.5	9	−0.5				
CTA	12.31	1	−3.3	7	−1.5	17	+1.4				
TCA	12.31	0	−3.5	17	+1.4	8	−1.2				
MetRS ec								0.244	0.244	0.253	0.259
TTA	12.55	0	−3.6	10	−0.7	6	−1.8				
CTA	13.44	1	−3.4	7	−1.7	10	−0.9				
TCA	13.44	3	−2.9	16	+0.7	14	+0.2				
TyrRS ec								0.253	0.232	0.275	0.240
TTA	5.77	4	−0.7	8	+0.9	3	−1.1				
CTA	5.97	0	−2.5	4	−0.8	4	−0.8				
TCA	5.97	0	−2.5	11	+2.1	6	0.0				
TyrRS bc								**0.329**	**0.236**	**0.228**	**0.207**
TTA	7.73	7	−0.3	4	−1.4	6	−0.6				
CTA	6.79	3	−1.4	6	−0.7	2	−1.9				
TCA	6.79	7	+0.1	14	+2.8	8	+1.0				
GluRS ec								0.237	0.268	0.256	0.239
TTA	8.00	1	−2.5	9	+0.4	6	−0.7				
CTA	7.14	0	−2.7	6	−0.4	7	0.0				
TCA	7.14	0	−2.7	12	+1.8	12	+1.8				
TrpRS ec								0.256	0.226	0.276	0.242
TTA	4.36	1	−1.6	4	−0.2	3	−0.6				
CTA	4.66	2	−1.2	2	−1.2	7	+1.1				
TCA	4.66	3	−0.8	4	−0.3	5	+0.2				
ValRS ec								0.249	0.209	0.277	0.265
TTA	10.30	0	−3.2	9	−0.4	9	−0.4				
CTA	13.08	0	−3.6	6	−1.9	6	−1.9				
TCA	13.08	0	−3.6	29	+4.4	13	0.0				

TABLE II

Continued.

	Expected	Observed						Base frequencies			
		RF-1		RF-2		RF-3					
aaRS	E	O	Z	O	Z	O	Z	A	T	G	C
Class II											
AsnRS ec								0.249	0.280	0.240	0.231
TTA	9.11	0	−3.1	6	−1.0	9	0.0				
CTA	7.53	0	−2.8	4	−1.3	5	−0.9				
TCA	7.53	3	−1.7	7	−0.2	4	−1.3				
AspRS ec								0.245	0.228	0.276	0.251
TTA	7.51	1	−2.4	9	+0.6	5	−0.9				
CTA	8.40	0	−2.9	4	−1.5	5	−1.2				
TCA	8.40	2	−2.2	16	+2.7	6	−0.8				
LysRS ec								0.264	0.262	0.250	0.224
TTA	9.09	5	−1.4	16	+2.3	5	−1.0				
CTA	7.78	1	−2.5	10	+0.8	2	−2.1				
TCA	7.78	2	−2.1	10	+0.8	11	+1.2				
LysRS cj								**0.355**	**0.325**	**0.205**	**0.115**
TTA	18.74	27	+2.0	16	−0.6	10	−2.1				
CTA	6.65	1	−2.2	12	+2.1	2	−1.8				
TCA	6.65	3	−1.4	3	−1.4	8	+0.5				
PheRS ec								0.255	0.229	0.274	0.242
TTA	4.38	2	−1.2	7	+1.3	4	−0.2				
CTA	4.63	0	−2.2	3	−0.8	2	−1.2				
TCA	4.63	3	−0.8	6	+0.6	6	+0.6				
PheRS sc								**0.345**	**0.279**	**0.194**	**0.182**
TTA	12.70	9	−1.0	14	+0.4	3	−2.8				
CTA	6.70	5	−0.7	9	+0.9	3	−1.4				
TCA	6.70	14	+3.4	9	+0.9	9	+0.9				
HisRS ec								0.234	0.232	0.298	0.236
TTA	5.31	4	−0.6	4	−0.6	8	+1.2				
CTA	5.42	1	−1.9	6	+0.3	3	−1.0				
TCA	5.42	1	−1.9	4	−0.6	4	−0.6				
ProRS ec								0.237	0.233	0.270	0.260
TTA	7.34	2	−2.0	7	−0.1	3	−1.6				
CTA	8.21	0	−2.9	8	−0.1	7	−0.4				
TCA	8.21	1	−2.5	8	−0.1	8	−0.1				
SerRS ec								0.260	0.241	0.258	0.241
TTA	6.30	2	−1.7	5	−0.5	1	−2.1				
CTA	6.50	0	−2.6	3	−1.4	7	0.0				
TCA	6.50	2	−1.8	4	−1.0	4	−1.0				
ThrRS ec								0.259	0.235	0.267	0.239
TTA	9.16	3	−2.1	18	+3.0	8	−0.4				
CTA	9.32	2	−2.4	1	−2.7	5	−1.8				
TCA	9.32	3	−2.1	12	+0.9	8	−0.4				

TABLE II

Continued.

	Expected	Observed						Base frequencies			
		RF-1		RF-2		RF-3					
aaRS	E	O	Z	O	Z	O	Z	A	T	G	C
AlaRS ec											
N-half								0.233	0.227	0.281	0.259
TTA	5.51	0	−2.4	6	+0.2	3	−1.1				
CTA	6.29	0	−2.5	2	−1.7	9	+1.1				
TCA	6.29	0	−2.5	9	+1.1	7	+0.3				
C-half								0.233	0.225	0.318	0.224
TTA	4.89	5	+0.1	7	+0.9	0	−2.2				
CTA	4.87	0	−2.2	4	−0.4	2	−1.3				
TCA	4.87	3	−0.8	9	+1.9	9	+1.9				

[a] As in (Goldstein and Brutlag, 1989), O – the observed number of each of the critical codons in each RF, E – its expected number. E was calculated by multiplying the corresponding base frequencies and the total number of codons in the sequence, n. Z is the normal deviate value, i.e. (O − E) divided by the square root of E[1 − (E/n)]. The table covers 11,732 total codons.

or less reliably represent the earliest precursors of the two aaRS classes. Accordingly, Table II gives observed and expected numbers of the critical TTA, CTA and TCA codons in 18 *E. coli* genes (10,333 codons in total) and, for contrast, in three particularly AT-rich genes of other origin. The codon usage was calculated in three reading frames, one coding RF-1 and two noncoding, RF-2 with one-base-shift of the frame and two-base-shifted RF-3. In RF-1, all three critical codons proved to be significantly under represented ($p < 0.01$ for most genes) with large negative normal deviation of observed number from that randomly expected at the given base composition. On the average, one such codon is detected per nearly 200 codons in class I genes, and per 100 codons in class II genes. In contrast, they are often excessive in RF-2 and RF-3 and characterized by large positive deviation from the expected value.

Solely, this difference does not necessarily indicate double-strand coding (Goldstein and Brutlag, 1989) because a nonrandom occurrence of codons in RF-1 is well known to be species-specific and depends on many factors including various underlying constraints on mono- and di-nucleotide frequencies (Nussinov, 1984; Filipski *et al.*, 1987; Ohno, 1988; etc.). Genes of aaRSs are not an exclusion. When their base composition is unbalanced toward a disproportionately increased content of A and T (shown bold in Table II), then TTA, TCA and CTA become present in RF-1 by either near expected number (TyrRS of *Bacillus caldotenax*) or even significantly abundant ($p < 0.05$) as in the case of critical TTA and TCA in LysRS gene of *Compylobacter jejuni* and mitochondrial PheRS gene of *Saccharomyces* *cerevisiae*. The same shifts of codon-usage were detected for most other AT-rich procaryotic and eucaryotic aaRS genes in RF-1, while RF-2 and RF-3 appeared to be evidently less sensitive (not shown). This suggests that codon usage was subsequently optimized to global changes in base composition, via pressure on 3rd codon positions, right in RF-1.

As emphasized above, common consensus sequences may be more relevant to test the original complementarity of the two aaRSs. Per all class I genes of *E. coli*, 3 critical codons in RF-1 were observed within the two conserved signature regions, all in the variable insert between P and HΦGH, and none in the KMSKS motif. In contrast, 21 critical codon were found in RF-2 and 25 in RF-3. For three signature motifs of *E. coli* class II enzymes this ratio was 4 (RF-1) : 33 (RF-2) : 41 (RF-3), again three of four critical codons in RF-1 were placed in the variable loop of the motif 2. In RF-1 these codons occupy non-homologous positions of the aligned sequences and consequently the consensus aaRSs do not contain the critical codons at all. In general, TTA, CTA and TCA tend to occur outside the catalytic moiety of aaRS genes. The most striking distribution was found in the 876-amino acid *E. coli* AlaRS from class II. Its N-terminal 461-amino acid part contains the motif 3 and the second fragment slightly resembling the motif 2 of the class II catalytic domain (Cusack *et al.*, 1991), while the following 415 amino acids may be deleted without having the least influence on specific aminoacylation (Schimmel *et al.*, 1993). Remarkably, all 8 critical codons of the gene are located within the 415-amino acid unnecessary domain. However, the antisense strand of the functional N-terminal half of AlaRS does not share any similarity with the two signature motifs of class I. It is not surprising inasmuch as of the three classII-defining signature motifs, the motif 3 is the only one which AlaRS confidently shares with other members of class II, while its highly diverged motifs 1 and 2 can be identified only due to more conservative secondary structural elements (Cusack *et al.*, 1991b; Ribas de Pouplana *et al.*, 1993; Musier-Forsyth and Schimmel, 1994; Shi *et al.*, 1994).

4. Discussion

In view of the hypothesis that minimalist precursors for both of the aaRS classes are primarily limited to their catalytic modules (Cusack *et al.*, 1991; Schimmel *et al.*, 1993; Moras, 1993), it is very remarkable that the inter-classes sequence complementarity described proved to be local, as well limited to their core signature motifs. Thus, the primary complementarity might envelop at least both signature motifs of class I pre-aaRSs and motif 1 with the proximal half of motif 2 of class II pre-aaRSs. Between motifs 1 and 2 of modern class II aaRSs is a variable insert of 35–80 residues (Eriani *et al.*, 1990) which roughly corresponds by length to the shortest connecting polypeptide found in *E. coli* CysRS (Hou *et al.*, 1991). In sum, it gives 100–150 residues as a near-minimum length of both of the hypothesized aaRS precursors (see also: Schimmel *et al.*, 1993). This also validates our jumbling tests, limited to the two signature regions with insertions of arbitrary structure between them. All the above means that nonconserved modules have been primarily shaped or even appeared later in evolution of aaRSs and then diverged, specifically for the two classes, their subclasses and so on up to individual synthetases.

The original double-strand coding readily explains the failure of all previous attempts to find any parallel sequence homology between the two synthetases. To be more exact, it is possible that the 'parallel' homology was looked for in the 'wrong' places. Indeed, when genes are abundant in palindromes one may expect to meet homologous peptide oligomers on both ('sense' and 'antisense') strands of DNA (Ohno, 1991; Ohno and Yomo, 1991). Moreover, retention of some secondary structures like β-barrel and $\alpha\beta\alpha$ motifs can occur in the antisense proteins (Zull and Smith, 1990). However, the hallmark of two proteins encoded by complementary strands in the same reading frame is a special type of symmetry such that (i) in general, any proximal segment near the amino end of one protein has to be complementary to the 'mirror' distal segment located near the carboxyl end of the other (and *vice versa*) (Ohno, 1991; Ohno and Yomo, 1991; Zull and Smith, 1990; Borst *et al.*, 1985), and (ii) in particular, if these segments are derived from a palindrome, they could save the original similarity (Ohno, 1991; Ohno and Yomo, 1991). While applied to the origin of the two aaRS classes, the second condition requires a detailed investigation, the first one is certainly the case of this enzyme's earliest history. Indeed, because of the intensive 'adaptive radiation' of connective polypeptides in synthetases of both classes, the distance between their signature sequences is highly variable. Nevertheless, the order which they follow along the sequence remains unchanged. This order is such that in all class I synthetases the HϕGH motif precedes the KMSKS and in all class II synthetases motif 1, a presumable complementary image of KMSKS, precedes motif 2 which resembles a complementary counterpart of the HϕGH motif. This reciprocal mapping argues in favor of the complementary origin of the two aaRS classes. Particularly favoring this conclusion is the localization of nonconserved anticodon-binding domains: in all class I aaRSs this domain occurs at their C-terminus, while again with eloquent symmetry, in all class II aaRSs at their N-terminus.

Other puzzling features of the two aaRSs also become more transparent in view of their suggested complementary origin, including:

1) All 20 synthetases are equally distributed amongst the two classes, exactly 10 : 10, that clearly is not expected for their independent origins but seems very natural if they were originally encoded by complementary strands of the same nucleic acids.

2) The two alternative synthetases approach the cognate tRNAs from opposite sides thus structurally motivating the two modes of class-specific hydroxyl group recognition (Ruff *et al.*, 1991; Moras, 1993). Moreover, this duality could be necessitated by a preexisting RNA world 'with its intrinsic ambiguity of the attachment at the $2'$- or $3'$-OH' (Ruff *et al.*, 1991).

3) In tune with the above is the notable observation that 3-D models of the class I Gln RS and class II AspRS complexed with cognate tRNAs look like mirror images of one another, the same being generally expected for other pairs of such complexes (Ruff *et al.*, 1991; Moras, 1993; Carter, 1993).

We suppose that the above intriguing rationalization may descend from the common complementary root of the two synthetases. Moreover, tRNAs with complementary anticodons also might originate concertedly as pairs of (+) and (−) sequences (Rodin *et al.*, 1993a,b). Even better average complementarity within the anticodon stem and loop was shown by the pairs of consensus tRNAs with quasi-complementary anticodons, including those with G versus U at their 2nd sites (*ibid*). If so, of 20 amino acids, the only one, Gly, appears to be activated by the aaRS from the same class II as *all* its complementary partners, Ser, Pro, Thr and Ala. Remarkably, the idea that the first 'ribozymic' t-RNA synthetases were likely those specific for basic amino acids (Weiner and Maizels, 1987) appears to be supported by the double concerted origins of tRNAs with complementary anticodons and the two classes of aaRSs. Indeed, 9 of 10 anticodons for the main basic amino acids, Arg, His, and Lys, have a complementary partner which is recognized by aaRS from the opposite class. Thus, rather well consistent are both the hypotheses of concerted origin, for tRNAs and for their cognate aaRSs. What is more, the strikingly rational two modes of class-specific tRNA hydroxyl group recognition by aaRSs(Ruff *et al.*, 1991; Moras, 1993) could be 'inborn' unavoidable rationality since literally *the mirror enzymes 'have to' recognize the opposite sides of mirror substrates.*

The idea of two synthetases sharing a common complementary origin is also derived from some general theoretical premises. Dictated by the problem of an error catastrophe (Eigen and Schuster, 1979), especially serious for replication of originating single-stranded RNA molecules (Goldberg and Wittes, 1966; Eigen and Schuster, 1979; Reanney, 1987), synchronous usage of both relatively short (+) and (−) sequences is thought to have been a key characteristic of primordial self-replicating systems. Accordingly, there are some evidences of the genetic code optimization to the double-strand coding (Konecny *et al.*, 1993). Only when the code has become basically established and, as a consequence, the error threshold has been overcome, can the direct coevolution of complementary proto-genes by replication be left for their more independent evolution by duplication followed by mutational decay of the original extensive complementarity except for conservative signature motifs in catalytic modules. However, molecular evolution is an opportunistic process and still tests both 'sense' and 'antisense' strands (Adelman *et al.*, 1987; Lasar *et al.*, 1989; Goldgaber, 1991; Hewinson *et al.*, 1991; Fukuchi and Otsuka, 1992). Furthermore, it is just this coevolution of minimalist pre-aaRSs with minimalist pre-tRNAs which could provide the gradual shaping of the genetic code (Schimmel *et al.*, 1993; Moras, 1993). Therefore, as a very tempting opportunity we view their combined phylogenetic study in context of originating tRNA synthetases made of RNAs (Gilbert, 1986; Darnell and Doolittle, 1986; Weiner and Maizels, 1987; Szathmary, 1993) and the coevolutionary theory of the origin of the genetic code (Wong, 1975, 1980, 1981). This study is currently underway (Rodin *et al.*, in preparation). Also, the two alternative catalytic domains of aaRSs are now open for direct experimental testing of the conjecture (Blalock and Smith, 1984; Borst *et al.*, 1985; Blalock, 1990; see also: Brentani, 1988, 1990) that many of polypeptides encoded by complementary nucleic acid strands must show overall a mutual binding affinity.

Acknowledgements

We thank G. Cocho and W.-H. Li for stimulating discussions and an anonymous referee for critical comments and useful suggestions. Special thanks are due to A. Rodin for help in jumbling tests and E. Szathmary for sending his paper before publication. This work was supported in part by CONACYT of Mexico and by the programs of the Russian Academy of Sciences 'Geninform', 'Human Genome' and 'Frontiers in Genetics'.

References

Adelman, J.P., Bond, C.T., Douglas, J. and Herbert, E.: 1987, *Science* **235**, 1514.
Blalock, J.E.: 1990, *Trends Biotechnol.* **8**, 140.
Blalock, J.E. and Smith, E.M.: 1984, *Biochem. Biophys. Res. Commun.* **121**, 203.
Borst, K.L., Smith, E.M. and Blalock, J.E.: 1985, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1372.
Brentani, R.R.: 1988, *J. Theor. Biol.* **135**, 495.
Brentani, R.R.: 1990, *J. Mol. Evol.* **31**, 239.
Burbaum, J.J. and Schimmel, P.: 1992, *Protein Sci.* **1**, 575.
Carter, C.W.: 1993, *Annu. Rev. Biochem.* **62**, 715.
Cusack, S., Bertnet-Colominas, C., Hartlein, M., Nassar, N. and Leberman, R.: 1991a, *Nature* **347**, 249.
Cusack, S., Hartlein, M. and Leberman, R.: 1991b, *Nucleic Acids Res.* **19**, 3489.
Darnell, J.E. and Doolittle, W.F.: 1986, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1271.
Doolittle, R.F.: 1987, *Of URFS and ORFS*, University Science Books, Mill Valley.
Eigen, M. and Schuster, P.: 1979, *Hypercycle: A Principle of Natural Self-Organization*, Springer-Verlag, Heidelberg.
Eriani, G., Delarue, M., Poch, O., Gangloff, J. and Moras, D.: 1990, *Nature* **347**, 203.
Eriani, G., Dirheimer, G. and Gangloff, J.: 1991, *Nucleic Acids Res.* **19**, 265.
Filipski, J., Salinas, J. and Rodiger, F.: 1987, *DNA* **6**, 109.
Fukuchi, S. and Otsuka, J.: 1992, *J. Theor. Biol.* **158**, 271.
Gilbert, W.: 1986, *Nature* **319**, 818.
Goldberg, A.L. and Wittes, R.E.: 1966, *Science* **153**, 420.
Goldgaber, A.L.: 1991, *Nature* **351**, 106.
Goldstein, A. and Brutlag, D.L.: 1989, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 42.
Hewinson, R.G., Lowings, J.P., Dawson, M.D. and Woodward. M.J.: 1991, *Nature* **352**, 291.
Hou, Y.-M., Shiba, K., Mottes, C. and Schimmel, P.: 1991, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 976.
Konecny, J., Eckert, M., Schoniger, M. and Hofacker, G.L.: 1993, *J. Mol. Evol.* **36**, 407.
Lasar, M.A., Hodin, R.A., Darling, D.S. and Chin, W.W.: 1989, *Mol. Cell. Biol.* **9**, 1128.
Moras, D.: 1993, in *The Robert A. Welch Foundation 37th Conference on Chemical Research. 40 Years of the DNA Double Helix*, Houston.

Musier-Forsyth, K. and Schimmel. P.: 1994, *Biochemistry* **33**, 773.

Nagel, G.M. and Doolittle, R.F.: 1991, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8121.

Nussinov, R.: 1984, *Nucleic Acids Res.* **12**, 1749.

Ohno, S.: 1988, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 9630.

Ohno, S.: 1991, in *Evolution of Life: Fossils, Molecules and Culture*, Osawa, S. and Honjo, T. (eds.), Springer, Tokyo, 97.

Ohno, S. and Yomo, T.: 1991, *Electrophoresis* **12**, 103.

Perona, J.J., Rould, M.A., Steitz, T.A., Risler, J.-L., Zelwer, C. and Brunie, S.: 1991, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 2903.

Reanney, D.C.: 1987, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 751.

Ribas de Pouplana, L., Buechter, D., Davis, M. and Schimmel, P.: 1993, *Protein Sci.* **2**, 2259.

Rodin, S., Ohno, S. and Rodin, A.: 1993 a, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4723.

Rodin, S., Ohno, S., and Rodin, A.: 1993 b, *Origins Life Evol. Biosphere* **23**, 393.

Rould, M.A., Perona, J.J., Soll, D. and Steitz, T.A.: 1989, *Science* **246**, 1135.

Ruff, M., Kishnaswamy, S., Boeglin, M., Poterzman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J.C. and Moras, D.: 1991, *Science* **252**, 1682.

Schimmel, P., Gieget, R., Moras, D. and Yokoyama, S.: 1993, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 8763.

Smith, R.F. and Smith, T.F.: 1990, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 118.

Shi, J.-P., Musier-Forsyth, K. and Schimmel, P.: 1994, *Biochemistry* **33**, 5312.

Szathmary, E.: 1993, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 9916.

Weiner, A. and Maizels, N.: 1987, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7383.

Wong, J. T-F.: 1975, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1909.

Wong, J. T-F.: 1980, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 1083.

Wong, J. T-F.: 1981, *Trends Biochem. Sci.* **6**, 33.

Yarus, M.: 1988, *Science* **240**, 1751.

Zharkikh, A.A., Rzhetsky, A.Y., Morozov, P., Sitnikova, T. and Krushkal, J.S.: 1991, *Gene* **101**, 251.

Zull, J.E. and Smith, S.K.: 1990, *Trends Biochem. Sci.* **15**, 257.