FEBS Letters 470 (2000) 300–304 FEBS 23482

A gene fusion event in the evolution of aminoacyl-tRNA synthetases

Eric Berthonneau, Marc Mirande*

Laboratoire d'Enzymologie et Biochimie Structurales, C.N.R.S., 1 Avenue de la Terrasse, 91190 Gif-sur-Yvette, France

Received 14 February 2000

Edited by Lev Kisselev

Abstract The genes of glutamyl- and prolyl-tRNA synthetases (GluRS and ProRS) are organized differently in the three kingdoms of the tree of life. In bacteria and archaea, distinct genes encode the two proteins. In several organisms from the eukaryotic phylum of coelomate metazoans, the two polypeptides are carried by a single polypeptide chain to form a bifunctional protein. The linker region is made of imperfectly repeated units also recovered as singular or plural elements connected as N-terminal or C-terminal polypeptide extensions in various eukaryotic aminoacyl-tRNA synthetases. Phylogenetic analysis points to the monophyletic origin of this polypeptide motif appended to six different members of the synthetase family, belonging to either of the two classes of aminoacyl-tRNA synthetases. In particular, the monospecific GluRS and ProRS from *Caenorhabditis elegans*, an acoelomate metazoan, exhibit this recurrent motif as a C-terminal or N-terminal appendage, respectively. Our analysis of the extant motifs suggests a possible series of events responsible for a gene fusion that gave rise to the bifunctional glutamyl-prolyl-tRNA synthetase through recombination between genomic sequences encoding the repeated units.

Key words: Gene fusion; Paralogous gene; Molecular phylogeny; Aminoacyl-tRNA synthetase; Repeated motif

1. Introduction

Aminoacyl-tRNA synthetases are a family of 20 enzymes which are divided into two distinct classes of 10 enzymes each [1,2]. The class I and class II aminoacyl-tRNA synthetases have distinct architectures of their active sites: class I enzymes are built around a Rossmann fold and class II synthetases contain an anti-parallel β-fold. With the exception of lysyl-tRNA synthetase [3], this partition is frozen in the three kingdoms (the class to which a synthetase belongs is generally the same in bacteria, archaea and eukarya); it should have been established very early in evolution. Therefore, this ancient protein family is believed to harbor some remnants of the establishment of the genetic code. Accordingly, aminoacyl-tRNA synthetases have contributed a wealth of information about the origin of the genetic code [4–8], providing the rational to create an univocal relationship between nucleotide triplets (anticodons) and amino acids.

Phylogenetic analyses have invariably inferred a distinct ancestor for class I and class II synthetases. The consequence is that polypeptide chains for the two classes of enzymes share very little sequence similarity. Unexpectedly, glutamyl- (class I) and prolyl- (class II) tRNA synthetases have been shown to be carried by a single polypeptide chain in higher eukaryotes [9,10]. The N-terminal moiety of this large polypeptide specifies a glutamyl-tRNA synthetase (GluRS) and the C-terminal region contributes a prolyl-tRNA synthetase (ProRS). The two synthetase domains can be expressed separately as active enzymes [9,11–13]. Therefore, their association as a multifunctional protein is not a prerequisite for $tRNA^{Glu}$ or $tRNA^{Pro}$ aminoacylation and the two enzymes are unlikely to share functional domains. The existence of a bifunctional synthetase has suggested that a gene fusion might have arisen between distinct ancestral genes for GluRS and ProRSs [9]. Alternatively, glutamyl-prolyl-tRNA synthetase (GluProRS) might be an ancient feature in evolution, leading to separate genes in bacteria, archaea and lower eukaryotes.

The large polypeptide of GluProRS also displays an unprecedented feature for an aminoacyl-tRNA synthetase. The two synthetase domains are fused through a linker polypeptide made of imperfectly repeated units of about 50 amino acid residues. In human and fly GluProRS, the synthetase domains are separated by three and six repeated motifs, respectively. A single similar motif is also recovered as an N-terminal polypeptide extension in tryptophanyl-tRNA synthetase (TrpRS), glycyl-tRNA synthetase (GlyRS) and histidyl-tRNA synthetase (HisRS), or as a C-terminal extension in methionyl-tRNA synthetase (MetRS). Therefore, this repeated motif is a recurrent addition to both class I and class II aminoacyl-tRNA synthetases. Recent gene translocation events could be responsible for the spreading of this motif in class I and class II synthetases [14]. Several reports have suggested that this new protein motif has a general RNA binding activity [15–17].

In this paper, we report the isolation of the cDNA for the linker domain of hamster GluProRS and the phylogenetic analysis of this protein motif. Our analysis of the extant motifs reveals a monophyletic origin for the motif and suggests that GluProRS arose by recombination between homologous DNA sequences encoding conserved motifs. Putative remnants of the pre-fusion state have been identified in the monofunctional genes for GluRS and ProRS from the nematode *Caenorhabditis elegans*.

2. Materials and methods

2.1. Cloning cDNA encoding the repeated units of hamster EPRS

Total human RNA isolated from HeLa cells was used for cDNA synthesis. PCR was performed to isolate the cDNA for the repeated units of human EPRS using primers RS1 (5′-TCTGTGTCACTTATGAGCAC-3′), RS2 (5′-TTGATCACAAAGCCAGTGCT-3′) and RS3 (5′-GCCGACACAGGCTTATACTC-3′) deduced from the published sequence [10]. The amplified cDNA was used as a probe to screen a cDNA library constructed in the Uni-ZAP XR vector (Stratagene) starting from poly(A)$^+$ mRNA isolated from exponentially growing CHO cells [18]. The nucleotide sequence of the cDNA insert was determined on both strands by the chain-termination method [19].

*Corresponding author. Fax: (33)-1-69 82 31 29.
E-mail: marc.mirande@lebs.cnrs-gif.fr; http://www.lebs.cnrs-gif.fr/

PII: S0014-5793(00)01343-0

2.2. Protein and DNA sequence analyses

Computational searches were performed at the NCBI using the BLAST network service [20]. Amino acid sequences were aligned and analyzed using the Clustal X [21] and PAUP 4.0.0b2 [22] packages. For phylogenetic analyses, the positions with gaps were excluded. Maximum-parsimony analyses were conducted by using the random heuristic search method. The neighbor-joining method, based on the distance matrix calculated between all pairs from the sequence alignment, was also used for tree reconstruction. In both cases, the confidence limits of branch points were estimated from 1000 bootstrap replicates.

3. Results

The hamster GluProRS contains three imperfectly repeated units in the linker region between the N-terminal GluRS domain and the C-terminal ProRS domain. Each repeat contains 50 amino acid residues (Fig. 1). Repeats 1 and 2, and repeats 2 and 3 are separated by 23 and 27 amino acid residues, respectively. The primary structures of these inter-repeat segments are not conserved, but their global amino acid compositions are strikingly similar. Two thirds of the residues are Ala, Ser or Pro.

The multifunctional GluProRS from hamster has therefore a domain structure analogous to that of the human protein. By contrast, the *Drosophila* enzyme displays six repeated motifs in the linker region. Since the human, hamster and fly repeats share more than 50% identical residues, they most likely arose from a single ancestral motif. To search for the origin of this motif, and to investigate the putative gene fusion events that led to the emergence of the unique multifunctional aminoacyl-tRNA synthetase, we examined the occurrence of this conserved motif throughout the various data bases.

The first noticeable result is the finding that all the proteins that have been recovered by using this motif as a bait are aminoacyl-tRNA synthetases (Fig. 1). Therefore, the functional role of this motif is likely to be related to the function of these enzymes, suggesting that the repeats play a role in the aminoacylation reaction or in the interaction of these enzymes with other cellular components involved in protein biosynthesis.

Aminoacyl-tRNA synthetases that were recovered are either class I (GluRS, MetRS or TrpRS) or class II (ProRS, GlyRS or HisRS) synthetases, and are either monofunctional (GluRS, ProRS, MetRS, TrpRS, GlyRS and HisRS) or multifunctional enzymes (GluProRS). The conserved motif is present either as a singular unit (ProRS, MetRS, TrpRS, GlyRS and HisRS) or as a plural element (GluRS and GluProRS), and is appended to these aminoacyl-tRNA synthetases either as an N-terminal extension (ProRS, TrpRS, GlyRS and HisRS), a C-terminal extension (GluRS and MetRS) or as an internal linker (GluProRS). GluRS and ProRS from the nematode *C. elegans* deserve special mention. In *C. elegans*, these two synthetases are monofunctional enzymes, and GluRS from this organism is the only known example of a monofunctional synthetase contributing several tandemly repeated motifs.

From the sequence alignment shown in Fig. 1, it is already clearly apparent that the 34 motifs distribute into two families according to the length of the conserved motif. GluRS, ProRS, GluProRS and TrpRS display a 50 amino acids long motif, whereas MetRS, GlyRS and HisRS have a shorter motif, with a deletion of the 14 C-terminal amino acids.

To find out the possible origin of this motif, the 34 repeats

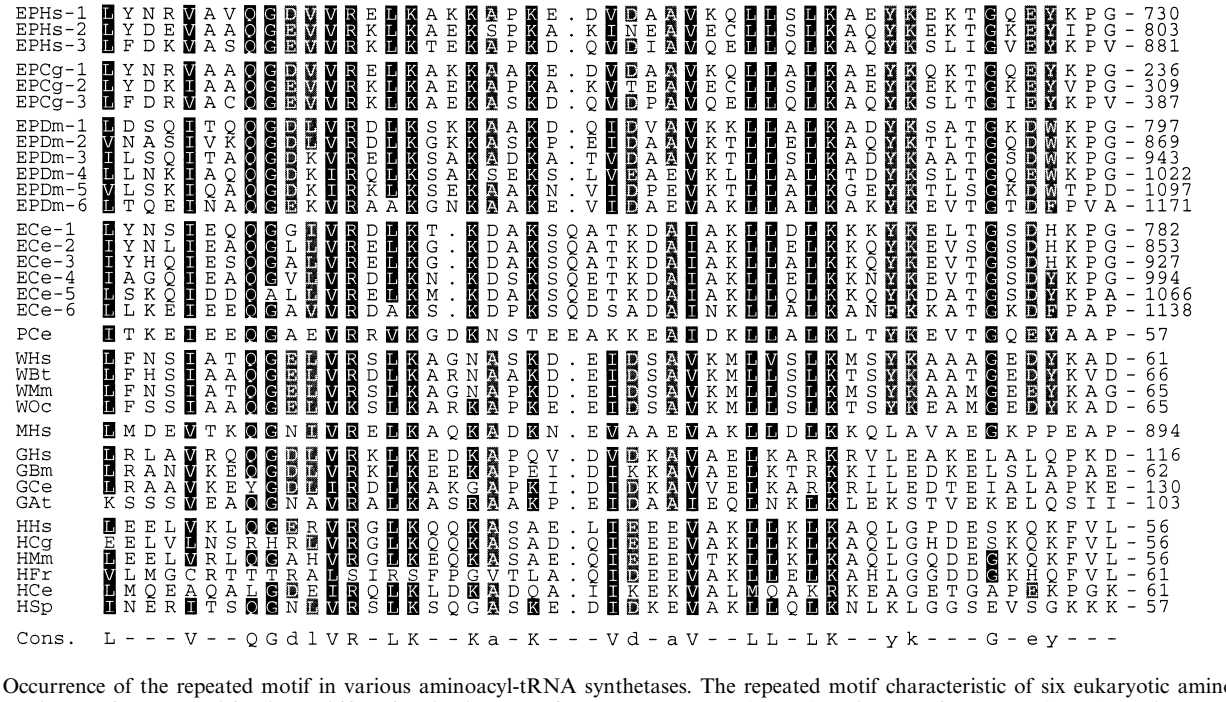

Fig. 1. Occurrence of the repeated motif in various aminoacyl-tRNA synthetases. The repeated motif characteristic of six eukaryotic aminoacyl-tRNA synthetases is recovered in the multifunctional GluProRS from *Homo sapiens* (EP-Hs) and *Cricetulus griseus* (EPCg) as three repeated units, in GluProRS from *Drosophila melanogaster* (EPDm) and GluRS from *C. elegans* (ECe) as six repeated units, in ProRS from *C. elegans* (PCe), in TrpRS from *H. sapiens* (WHs), *Bos taurus* (WBt), *Mus musculus* (WMm) and *Oryctolagus cuniculus* (WOc), in MetRS from *H. sapiens* (MHs), in GlyRS from *H. sapiens* (GHs), *Bombyx mori* (GBm), *C. elegans* (GCe) and *A. thaliana* (GAt), in HisRS from *H. sapiens* (HHs), *C. griseus* (HCg), *M. musculus* (HMm), *Fugu rubripes* (HFr), *C. elegans* (HCe) and *S. pombe* (HSp). Numbering of the repeated units refers to their position in the repeated domains. Conserved residues are indicated in black (75% of conservation) or gray (60% of conservation). The consensus sequence is shown at the bottom of the figure.

listed in Fig. 1 were used to build a phylogenetic tree. The availability of sequence data for the aminoacyl-tRNA synthetases that do or do not possess this motif in different species along the eukaryotic branch of the universal tree of life is listed in Table 1. Among the six aminoacyl-tRNA synthetases listed above, none of them from the budding yeast *Saccharomyces cerevisiae* has such an appended motif, and only HisRS from the fission yeast *Schizosaccharomyces pombe* has one repeat. We could not find any significant similarity between the amino acid sequence of the repeats and any other protein deduced from the genomic sequences of the various bacteria and archaea known to date. Therefore, according to the universal tree inferred from sequence comparisons of ribosomal RNA genes [23], the motifs appended to HisRS from *S. pombe* and/or GlyRS from the plant *Arabidopsis thaliana* are likely to correspond to the first occurrence of this motif in evolution. Conversely, HisRS from *A. thaliana* and GlyRS from *S. pombe* have no repeat motif. In *C. elegans*, that motif is already spread to ProRS and GluRS, in addition to HisRS and GlyRS, but is absent form TrpRS and MetRS. In vertebrates, the six synthetases possess this motif. The evolutionary relationships between the repeated motifs were appraised by using the 34 sequences available to construct a phylogenetic tree by the maximum-parsimony and neighbor-joining methods (Fig. 2). The parsimony and distance approaches gave essentially similar relationships between the repeats. The tight grouping of the different branches of the tree strongly suggests that the different motifs share a common ancestor. In addition, the relative clustering of the motifs appended to a synthetase of particular specificity involves that once acquired by a member of this family, the motifs have been vertically transferred in evolution.

Noteworthy, the unique N-terminal motif appended to the monofunctional ProRS from *C. elegans* (R-PRS) is clustered with the sixth motif from the six repeated units located at the C-terminus of the monofunctional GluRS from the same species (R6-ERS) (Fig. 2). The particular relationship between these two repeats was also assessed by the comparison of the nucleotide sequences encoding the conserved motifs of ProRS and GluRS from *C. elegans*. As shown in Fig. 3, the DNA sequences encoding the N-terminal motif of ProRS and the sixth C-terminal motif of GluRS are both interrupted by an intron that splits the coding sequences precisely at the same place. Moreover, the 150 nucleotide coding sequences are very

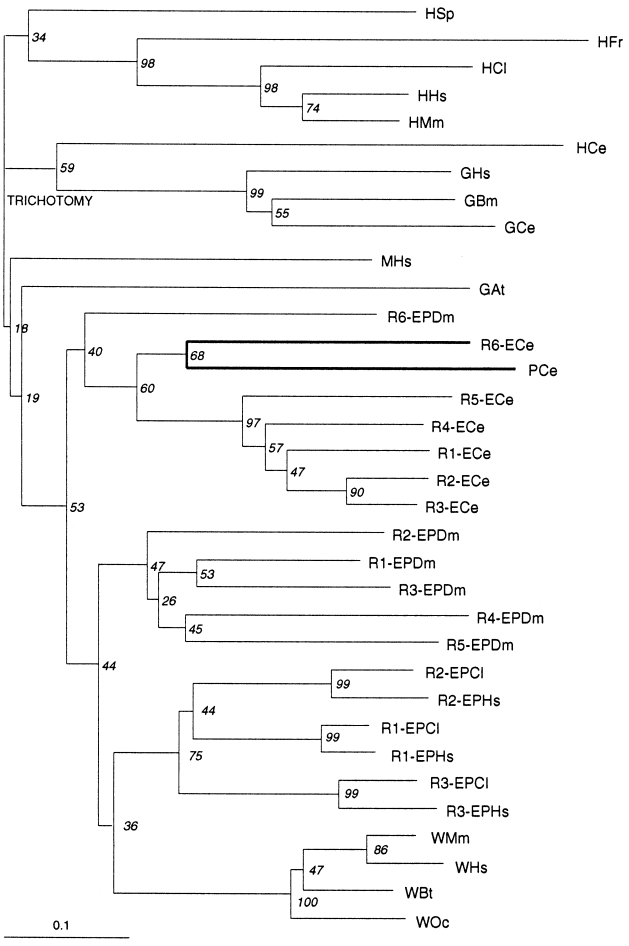

Fig. 2. Monophyletic origin of the repeated units appended to class I and class II aminoacyl-tRNA synthetases. Essentially similar unrooted trees were constructed with maximum-parsimony and neighbor-joining (shown here) methods. The numbers at the nodes of the branches represent frequencies of occurrence from 1000 bootstrap replicates. Abbreviations for synthetases and numbering of the repeats are as in Fig. 1. The scale bar corresponds to 0.1 amino acid replacement per site.

Table 1
Occurrence of repeated units in aminoacyl-tRNA synthetases from various origins

	Class II aaRS			Class I aaRS		
	HisRS	GlyRS	ProRS	GluRS	TrpRS	MetRS
Protist[a]	●[g]	●	●	●	−	●
Budding yeast[b]	−	−	−	−	−	−
Fission yeast[c]	+	−	−	−	−	−
Plant[d]	−	+	●	−	−	−
Nematode[e]	+	+	+	+	−	−
Vertebrate[f]	+	+	+	+	+	+

[a]From the amitochondrial protist *Encephalitozoon cuniculi*.
[b]*Saccharomyces cerevisiae*.
[c]*Schizosaccharomyces pombe*.
[d]From *Arabidopsis thaliana* or *Oriza sativa* (rice).
[e]*Caenorhabditis elegans*.
[f]Data are from *Fugu rubripes*, *Drosophila melanogaster*, *Bombyx mori*, *Mus musculus*, *Cricetulus griseus*, *Oryctolagus cuniculus*, *Bos taurus* or *Homo sapiens*.
[g]● indicates that the sequence is yet unknown. − and + indicate that the sequence of a particular synthetase is known and does (+) or does not (−) display the motif.

Fig. 3. Homologous sequences encoding repeated units in *C. elegans*. The nucleotide sequences encoding the sixth repeated unit of GluRS and the single N-terminal repeat of ProRS from the nematode *C. elegans* are aligned from the ATG initiation codon of ProRS to the TAA stop codon of GluRS, indicated in bold-faced type. The sequences encoding the repeats are split in two parts by an intron. Black boxes indicate conserved residues, gray boxes transitions.

similar; they share 94 identical nucleotide residues and 25 transitions (C–T and A–G permutations). Long patches of similarity are recovered both upstream and downstream of the intervening sequences (Fig. 3). As compared to the 37% nucleotide changes recovered for the coding sequences of R-PRS and R6-ERS, R-PRS displays 49%, 44%, 43%, 45% and 47% nucleotide changes with the DNA sequences encoding the fifth, fourth, third, second and first motif of the *Caenorhabditis* GluRS. This particular relationship between the two motifs R-PRS and R6-ERS suggests a possible series of events that could have led to the emergence of a multisynthetase protein.

The marked similarity between the two coding sequences R-PRS and R6-ERS suggests a common origin. This conserved region of PRS and ERS could have been the site of a gene fusion event leading to a bifunctional EPRS protein. Alternatively, this region of high similarity could represent the result of a gene duplication event. The latter possibility would imply that the fusion state was the ancestral state of ERS and PRS, with subsequent separation of the fused genes. However, several lines of evidence suggest that the fusion event is of late origin. First, the presence of an EPRS protein is restricted to organisms belonging to the coelomate branch of metazoan phylogeny, from arthropods to mammals. It is of prime significance to remark that the nematode *C. elegans*, that does not possess a bifunctional EPRS, is a member of the pseudocoelomate subdivision of metazoan, a phylum that immediately predates the emergence of coelomates. Second, no remnant of a putative fusion state can be uncovered from the genome sequences of the eubacteria and archaea known to date. Especially, as stated above, the earliest appearance of a repeat in evolution is observed in *S. pombe* and *A. thaliana*. Therefore, the present day EPRS is most likely the result of a late fusion event that occurred at the early stages of appearance of the coelomate phylum after PRS and ERS independently acquired identical sequences encoding a single (PRS) or a sixth repeated unit (ERS). The close similarity observed between R-PRS and R6-ERS from *C. elegans* can be viewed as a landmark of the pre-fusion state.

4. Discussion

The discovery of a bifunctional aminoacyl-tRNA synthetase in higher eukaryotes raises the question of the origin of this enzyme. Our results suggest that the repeated units that link the two synthetase domains were involved in a gene fusion mechanism that occurred at the early stage of coelomate evolution. According to the phylogenetic analysis of the extant motifs presented in this study, we can tentatively propose a credible evolutionary scenario to account for the emergence

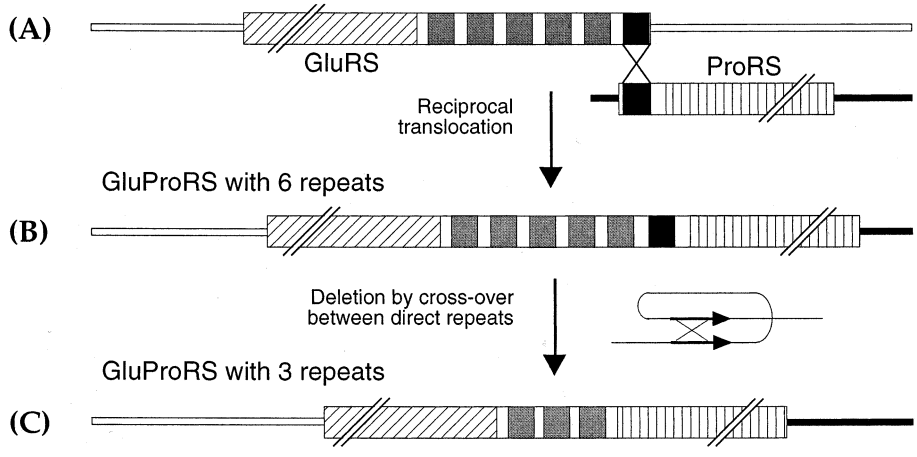

Fig. 4. Hypothetical scenario of gene fusion to yield a bifunctional aminoacyl-tRNA synthetase. (A) Two independent loci carrying a GluRS and a ProRS of the *C. elegans* type are schematized. The repeated units are indicated by gray or black boxes. The original GluRS locus (open line) is indicated with six repeated units encoded at the 3′-extremity of the gene, the ProRS locus (filled line) is shown with a single repeat at the 5′-end of its gene. The homologous nucleotide sequences are represented by black boxes. To seek for clarity, introns have been removed. (B) Reciprocal translocation through homologous sequences encoding the repeated units indicated in black led to a bifunctional synthetase with a linker region made of six repeated units, as is the case in the *Drosophila* protein. (C) A single or several cross-over events might have occurred through the directly repeated sequences encoding the six motifs to give GluProRS polypeptides with three repeated units, as found in human and hamster.

of a multifunctional enzyme in the family of aminoacyl-tRNA synthetases (Fig. 4).

The preliminary event that took place a long time ago in the eukaryotic phylum before EPRS formed concerns the capture of the same polypeptide extension by several aminoacyl-tRNA synthetases. The data from Table 1 and Fig. 2 suggest that it has first been acquired by a class II synthetase HisRS or GlyRS, and then captured by ProRS and GluRS, followed by TrpRS and MetRS. The implicit assumption is that the capture of this domain contributed a selective advantage. We recently showed that this motif corresponds to a novel general RNA binding domain with a helix-turn-helix conformation [17]. Other putative RNA binding domains appended to different eukaryotic aminoacyl-tRNA synthetases have been characterized in yeast GlnRS [24,25], in the yeast protein Arc1p that associates to MetRS and GluRS [26], in the mammalian p43 protein that associates to the multisynthetase complex [18]. These non-specific RNA binding domains were ascribed to *cis*- or *trans*-acting cofactors that enhance the rate for association of tRNA to the synthetase.

As observed in the present day synthetases from *C. elegans*, strikingly similar motifs have been captured at the C-terminus of a monofunctional GluRS and at the N-terminus of a monofunctional ProRS. We propose that GluRS and ProRS of the organism from which pseudocoelomates and coelomates diverged independently acquired homologous nucleotide sequences encoding repeated units (Fig. 4A). These homologous DNA segments would have been the target of a reciprocal translocation event leading to fusion of the GluRS and ProRS genes to generate a bifunctional polypeptide (Fig. 4B). Modern EPRSs possess three (human and hamster) or six (fly) repeated units in their linker region. If an EPRS with six units predated the species with three repeats, it is conceivable that deletion of genetic material might have occurred via a cross-over mechanism involving adjacent units (Fig. 4C).

Phylogenetic analyses have shown that horizontal gene transfer events may have contributed an essential process in the evolution of aminoacyl-tRNA synthetases [8,27,28]. Our results show that gene fusion is also an important mechanism in the evolution of aminoacyl-tRNA synthetases. Because the emergence of EPRS is a late event in evolution, we could recognize remnants of the pre-fusion state in *C. elegans* and identify possible molecular features involved in the fusion event. Aminoacyl-tRNA synthetases are ancient proteins composed of a large variety of structural and functional domains [8,29]. It is generally believed that specialization of these enzymes for a given amino acid has been accomplished via the insertion of extra polypeptide segments to the core catalytic domains representing ancestral class I and class II synthetases. It is conceivable that the early addition of new domains involved gene fusion mechanisms similar to those described in this study.

References

[1] Eriani, G., Delarue, M., Poch, O., Gangloff, J. and Moras, D. (1990) Nature 347, 203–206.
[2] Cusack, S., Berthet-Colominas, C., Härtlein, M., Nassar, N. and Leberman, R. (1990) Nature 347, 249–255.
[3] Ibba, M. et al. (1997) Science 278, 1119–1122.
[4] Brown, J.R. and Doolittle, W.F. (1995) Proc. Natl. Acad. Sci. USA 92, 2441–2445.
[5] Wetzel, R. (1995) J. Mol. Evol. 40, 545–550.
[6] Siatecka, M., Rozek, M., Barciszewski, J. and Mirande, M. (1998) Eur. J. Biochem. 256, 80–87.
[7] Ribas de Pouplana, L., Turner, R.J., Steer, B.A. and Schimmel, P. (1998) Proc. Natl. Acad. Sci. USA 95, 11295–11300.
[8] Wolf, Y.I., Aravind, L., Grishin, N.V. and Koonin, E.V. (1999) Genome Res. 9, 689–710.
[9] Cerini, C., Kerjan, P., Astier, M., Gratecos, D., Mirande, M. and Semeriva, M. (1991) EMBO J. 10, 4267–4277.
[10] Fett, R. and Knippers, R. (1991) J. Biol. Chem. 266, 1448–1455.
[11] Kerjan, P., Triconnet, M. and Waller, J.P. (1992) Biochimie 74, 195–205.
[12] Ting, S.M., Bogner, P. and Dignam, J.D. (1992) J. Biol. Chem. 267, 17701–17709.
[13] Stehlin, C., Burke, B., Yang, F., Liu, H.J., Shiba, K. and Musier-Forsyth, K. (1998) Biochemistry 37, 8605–8613.
[14] Brenner, S. and Corrochano, L.M. (1996) Proc. Natl. Acad. Sci. USA 93, 8485–8489.
[15] Wu, H., Nada, S. and Dignam, J.D. (1995) Biochemistry 34, 16327–16336.
[16] Rho, S.B., Lee, J.S., Jeong, E.J., Kim, K.S., Kim, Y.G. and Kim, S. (1998) J. Biol. Chem. 273, 11267–11273.
[17] Cahuzac, B., Berthonneau, E., Birlirakis, N., Guittet, E. and Mirande, M. (2000) EMBO J. 19, 445–452.
[18] Quevillon, S., Agou, F., Robinson, J.C. and Mirande, M. (1997) J. Biol. Chem. 272, 32573–32579.
[19] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[20] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[21] Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G. and Gibson, T.J. (1998) Trends Biochem. Sci. 23, 403–405.
[22] Swofford, D.L. (1998) Sinauer Associates, Sunderland, MA.
[23] Woese, C.R., Kandler, O. and Wheelis, M.L. (1990) Proc. Natl. Acad. Sci. USA 87, 4576–4579.
[24] Whelihan, E.F. and Schimmel, P. (1997) EMBO J. 16, 2968–2974.
[25] Wang, C.C. and Schimmel, P. (1999) J. Biol. Chem. 274, 16508–16512.
[26] Simos, G., Sauer, A., Fasiolo, F. and Hurt, E.C. (1998) Mol. Cell 1, 235–242.
[27] Lamour, V., Quevillon, S., Diriong, S., Nguyen, V.C., Lipinski, M. and Mirande, M. (1994) Proc. Natl. Acad. Sci. USA 91, 8670–8674.
[28] Shiba, K., Motegi, H. and Schimmel, P. (1997) Trends Biochem. Sci. 22, 453–457.
[29] Delarue, M. and Moras, D. (1993) Bioessays 15, 675–687.