Evolution of Aminoacyl-tRNA Synthetases––Analysis of Unique Domain Architectures and Phylogenetic Trees Reveals a Complex History of Horizontal Gene Transfer Events

Yuri I. Wolf, L. Aravind, Nick V. Grishin, et al.

Genome Res. 1999 9: 689-710
Access the most recent version at doi:10.1101/gr.9.8.689

Evolution of Aminoacyl-tRNA Synthetases—Analysis of Unique Domain Architectures and Phylogenetic Trees Reveals a Complex History of Horizontal Gene Transfer Events

Yuri I. Wolf,[1,3] L. Aravind,[1,2] Nick V. Grishin,[1] and Eugene V. Koonin[1,4]

[1]*National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH), Bethesda Maryland 20894 USA;* [2]*Department of Biology, Texas A&M University, College Station, Texas 70843 USA*

Phylogenetic analysis of aminoacyl-tRNA synthetases (aaRSs) of all 20 specificities from completely sequenced bacterial, archaeal, and eukaryotic genomes reveals a complex evolutionary picture. Detailed examination of the domain architecture of aaRSs using sequence profile searches delineated a network of partially conserved domains that is even more elaborate than previously suspected. Several unexpected evolutionary connections were identified, including the apparent origin of the β-subunit of bacterial GlyRS from the HD superfamily of hydrolases, a domain shared by bacterial AspRS and the B subunit of archaeal glutamyl-tRNA amidotransferases, and another previously undetected domain that is conserved in a subset of ThrRS, guanosine polyphosphate hydrolases and synthetases, and a family of GTPases. Comparison of domain architectures and multiple alignments resulted in the delineation of synapomorphies—shared derived characters, such as extra domains or inserts—for most of the aaRSs specificities. These synapomorphies partition sets of aaRSs with the same specificity into two or more distinct and apparently monophyletic groups. In conjunction with cluster analysis and a modification of the midpoint-rooting procedure, this partitioning was used to infer the likely root position in phylogenetic trees. The topologies of the resulting rooted trees for most of the aaRSs specificities are compatible with the evolutionary "standard model" whereby the earliest radiation event separated bacteria from the common ancestor of archaea and eukaryotes as opposed to the two other possible evolutionary scenarios for the three major divisions of life. For almost all aaRSs specificities, however, this simple scheme is confounded by displacement of some of the bacterial aaRSs by their eukaryotic or, less frequently, archaeal counterparts. Displacement of ancestral eukaryotic aaRS genes by bacterial ones, presumably of mitochondrial origin, was observed for three aaRSs. In contrast, there was no convincing evidence of displacement of archaeal aaRSs by bacterial ones. Displacement of aaRS genes by eukaryotic counterparts is most common among parasitic and symbiotic bacteria, particularly the spirochaetes, in which 10 of the 19 aaRSs seem to have been displaced by the respective eukaryotic genes and two by the archaeal counterpart. Unlike the primary radiation events between the three main divisions of life, that were readily traceable through the phylogenetic analysis of aaRSs, no consistent large-scale bacterial phylogeny could be established. In part, this may be due to additional gene displacement events among bacterial lineages. Argument is presented that, although lineage-specific gene loss might have contributed to the evolution of some of the aaRSs, this is not a viable alternative to horizontal gene transfer as the principal evolutionary phenomenon in this gene class.

[Complete multiple alignments of all aaRSs from complete genomes as well as the alignments of conserved regions used for phylogenetic tree construction are available at ftp://ncbi.nlm.nih.gov/pub/koonin/aaRS]

Aminoacyl-tRNA synthetases (aaRSs) are key components of the protein translation machinery that catalyze two basic reactions: (1) activation of amino acids via the formation of aminoacyl adenylates and (2) linking the activated amino acid to the cognate tRNAs. The aaRSs generate AMP as the second end product of this reaction, which differentiates them from the majority of ATP-dependent enzymes that produce ADP. aaRSs specific for each of the 20 amino acids have been identified, and there are two structurally distinct and apparently unrelated classes of aaRS, each encompassing 10 specificities (Cusack et al. 1990, 1991; Eriani et al. 1990, 1995; Cusack 1995, 1997). The two classes have different modes of aminoacylation: aaRSs of class I

[3]*Permanent address: Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk 630090, Russia.*
[4]Corresponding author.
E-MAIL koonin@ncbi.nlm.nih.gov; FAX (301) 480-9241.

9:689–710 ©1999 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/99 $5.00; www.genome.org

aminoacylate the 2′OH of the cognate tRNA, whereas those that belong to class II aminoacylate 3′OH (with the exception of PheRS). aaRSs of each class contain a conserved core domain that is involved in ATP binding and hydrolysis and combines with additional domains that determine the specificity of interactions with the cognate amino acid and tRNA (Delarue and Moras 1993; Cusack 1995, 1997). The core domain of class I contains a parallel β-sheet, which resembles the nucleotide-binding Rossmann fold in its topology (Moras 1992). The class I core domain contains 2 conserved motifs, designated "HIGH" and "KMSKS" (after the characteristic amino acid signatures), that are directly involved in ATP binding (Eriani et al. 1990; Moras 1992; Arnez and Moras 1997). A specific structural similarity has been suggested to exist between the class I aaRSs core domain and a superfamily of nucleotidyltransferases that are typified by the bacterial cytidylyl transferase TagD and contain a conserved HIGH-like motif (Bork et al. 1995).

The class II core contains a mixed β-sheet similar to that found in biotin synthases (Artymiuk et al. 1994). This domain contains three loosely conserved motifs that participate in ATP binding; they are unrelated to the HIGH and KMSKS motifs (Eriani et al. 1990; Moras 1992; Arnez and Moras 1997). The extra domains of aaRSs are either inserted into loops within the core domain or appended to the amino and carboxyl termini of the core. These accessory domains show remarkable diversity, resulting in a complex, modular domain architecture, which is largely amino acid specific, although some domains are common in several aaRSs of different specificities (Delarue and Moras 1993).

Typically, aaRSs of the same specificity are highly conserved, whereas those with different specificities show only limited conservation, mostly confined to the core, ATP pyrophosphatase domain. There are only three apparent exceptions to this rule: (1) GlnRS, unlike other aaRSs, is present only in eukaryotes and γ-Proteobacteria and appears to be specifically related to a subset of GluRS (Freist et al. 1997a,b; Siatecka et al. 1998), (2) the same type of relationship has been described for AsnRS and AspRS (Shiba et al. 1998), and (3) there are two types of LysRS that belong to class I and class II, respectively, and appear to be unrelated to each other (Ibba et al. 1997a,b; Koonin and Aravind 1998; Siatecka et al. 1998). aaRSs for 17 amino acids appear to be universal, that is, they are encoded by all organisms for which genome sequences are available. The exceptions are GlnRS that, as already mentioned, is missing in most bacteria and archaea, AsnRS that is missing in most archaeal and several bacterial species, and CysRS that so far has not been identified in two archaea (Doolittle and Handy 1998; Koonin and Aravind 1998). The mechanism for postaminoacylation formation of Gln and Asn via transamidation of tRNAs charged with Glu and Asp, respectively, has been characterized (Curnow et al. 1996; Wilcox and Nirenberg 1968). The mechanism of cysteine incorporation into proteins in those archaea that lack CysRS remains a mystery. These exceptions notwithstanding, the ubiquity of aaRSs indicates that the two classes have evolved by serial duplication at a very early stage of evolution and had been already locked into the distinct specificities in the last common ancestor (LCA) of all extant life forms.

For several reasons, aaRSs appear to be an excellent test case for an analysis of the forces that shape gene and genome evolution on a large time scale.

1. The set of aaRSs is naturally defined by the 20 specificities required for protein synthesis.
2. aaRSs are ubiquitous (with the exceptions mentioned above) and essential, therefore a gene encoding an aaRS generally cannot be lost in evolution unless it is displaced by another gene that encodes a different form of aaRS of the same specificity.
3. aaRSs with the same specificity typically do not form paralogous families—only a few isolated duplications of this type have been noticed. This significantly reduces ambiguity in phylogenetic analysis.
4. As the specificities of at least 17 of the 20 aaRSs (GlnRS and possibly AsnRS and LysRS being the exceptions) apparently have been established in the LCA and have not changed ever since, it seems unlikely that the aaRS genes have undergone major changes in evolutionary rates.
5. Unlike, for example, ribosomal proteins, aaRSs typically are not involved in complex interactions with multiple protein partners. The only interactions that are essential for their function are those with amino acids, ATP, and the cognate tRNA (although exceptions are possible). Discrimination of cognate from noncognate tRNAs by aaRSs is a complex process, the details of which differ for different specificities, but, at least in some cases, aaRSs are compatible with tRNAs even from phylogenetically distant organisms (Bedouelle et al. 1993; Ripmaster et al. 1995; Shiba et al. 1997a,b; Soma and Himeno 1997, 1998; Lenhard et al. 1999). Accordingly, there is at least some potential for horizontal transfer of aaRS genes in evolution.
6. Given the variety of modular domain arrangements seen in aaRSs, phylogenetic analysis might shed light on the modes whereby such modules are acquired and exchanged during evolution.

aaRSs have been among the most popular objects of molecular phylogenetic studies, and several unusual evolutionary patterns have been observed when tree

topologies for aaRSs were compared to the topologies derived from the analysis of rRNAs and other molecules involved in translation (Nagel and Doolittle 1995; Brown and Doolittle 1997; Shiba et al. 1997a; Doolittle and Handy 1998). Phylogenetic analysis of aaRSs is becoming increasingly interesting with the growth of the collection of complete genome sequences that currently consists of >20 genomes of bacteria, archaea, and eukaryotes. Because each of the aaRSs is indispensable in the context of the modern-type translation system, this collection provides us with at least 17 sets of sequences of functionally equivalent aaRSs from all these diverse organisms. Although many sequences of aaRSs have been available for a long time, complete genomes are critical for conducting a convincing evolutionary analysis. Only from complete genome sequences, the full information on all aaRSs encoded by each species, including all possible paralogs, can be extracted. In a recent insightful overview, Doolittle and Handy (1998) note that the number of apparent evolutionary anomalies grows rapidly with the increase in genome sequence information, resulting in a highly complex picture.

Here, we describe an attempt of a comprehensive analysis of the evolutionary patterns for all 20 sets of aaRSs using, primarily, the available complete genome sequences. We pursued two principal goals: (1) Using the recently developed sensitive methods for sequence analysis, together with structural information, delineate as completely as possible the domain architecture of all aaRSs; (2) generate phylogenetic trees for aaRSs of all specificities using carefully constructed multiple alignments and, whenever feasible, infer the root position. The results of phylogenetic analysis of most of the aaRSs appear to be compatible with the "standard model" that postulates the original radiation of bacteria and archaea–eukaryotes, followed by the divergence of the latter two divisions. However, for at least 15 of the aaRS specificities, this straightforward scenario needs to be amended by including horizontal gene transfer events, in some cases multiple ones, between major phylogenetic lineages, as well as acquisition, loss, and exchange of accessory domains. Our general conclusion is that the available sequence information is sufficient for reconstructing the principal events in the evolution of most, if not all, of the aaRSs.

RESULTS AND DISCUSSION

Modular Domain Architectures of aaRSs—Previously Undetected Accessory Domains and New Occurrences of Known Domains

Careful examination of the multiple alignments of aaRSs of all 20 specificities shows that each of them, without exception, has a complex, modular architecture (Fig. 1). Furthermore, the accessory domains form a network that connects aaRS of different specificities. Many of these domains have been described in previous studies (Delarue and Moras 1993; Koonin et al. 1994; Simos et al. 1996; Aravind and Koonin 1999), but using iterative profile searches with PSI-BLAST, we identified several previously undetected domains as well as new occurrences of known domains. The extensive structural characterization of the aaRSs has produced representative structures for almost all of the domains seen in these proteins (Fig. 1).

Four domains are shared by aaRSs of class I and class II (Fig. 1A,B). These are (1) a predicted RNA-binding domain that is a distinct version of the OB-fold (EMAP domain) and is found in all archaeal and a subset of bacterial MetRS, some of the eukaryotic TyrRS (both of class I), and the β-subunit of PheRS (class II); (2) the "DALR" domain that is shared by seven aaRSs of class I and the β-subunit of bacterial GlyRS of class II (see below); (3) a small domain that is predicted to possess an α-helical, coiled-coil structure but nevertheless is highly specific to aaRSs, is readily detectable by iterative database searches without any false positives, is present in animal TrpRS, MetRS, and GlnRS (class I) and HisRS, ProRS, and GlyRS (class II), and has been shown to facilitate the formation of multi-aaRS complexes that have been isolated from animal cells as well as their interaction with tRNAs (Rho et al. 1998); and (4) a small carboxy-terminal domain (designated "C-V/I/G" in Fig. 1A,B) shared by ValRS, eukaryotic, and archaeal IleRS (class I) and archaeal and eukaryotic GlyRS (class II).

All these domains have been described previously, but, with the exception of the EMAP domain that has been analyzed in considerable detail (Simos et al. 1996; Weiner and Maizels 1999), the present study expanded the range of aaRSs that contain each of them (Fig. 1). In particular, the domain that we designated DALR, after a characteristic pattern of amino acid residues that is conserved in many of the respective sequences, has been recognized in ArgRS (where it has been designated Add-2), MetRS, and the RS for the three aliphatic amino acids (Cavarelli et al. 1998) but, to our knowledge, not in CysRS, class I LysRS, or the β-subunit of the bacterial GlyRS. The detection of the DALR domain in these additional sets of aaRSs makes it the most widespread domain in aaRSs, after the two core domains. It is an α-helical domain with a unique architecture that has been implicated in anticodon binding (Brunie et al. 1990; Cavarelli et al. 1998). Furthermore, it has been shown that deletions in the carboxy-terminal portion of the β-subunit of *Escherichia coli* GlyRS affect tRNA recognition (Toth and Schimmel 1990), which appears to be compatible with an anticodon-binding function of the DALR domain. In this regard, the presence of the DALR domain in the class I

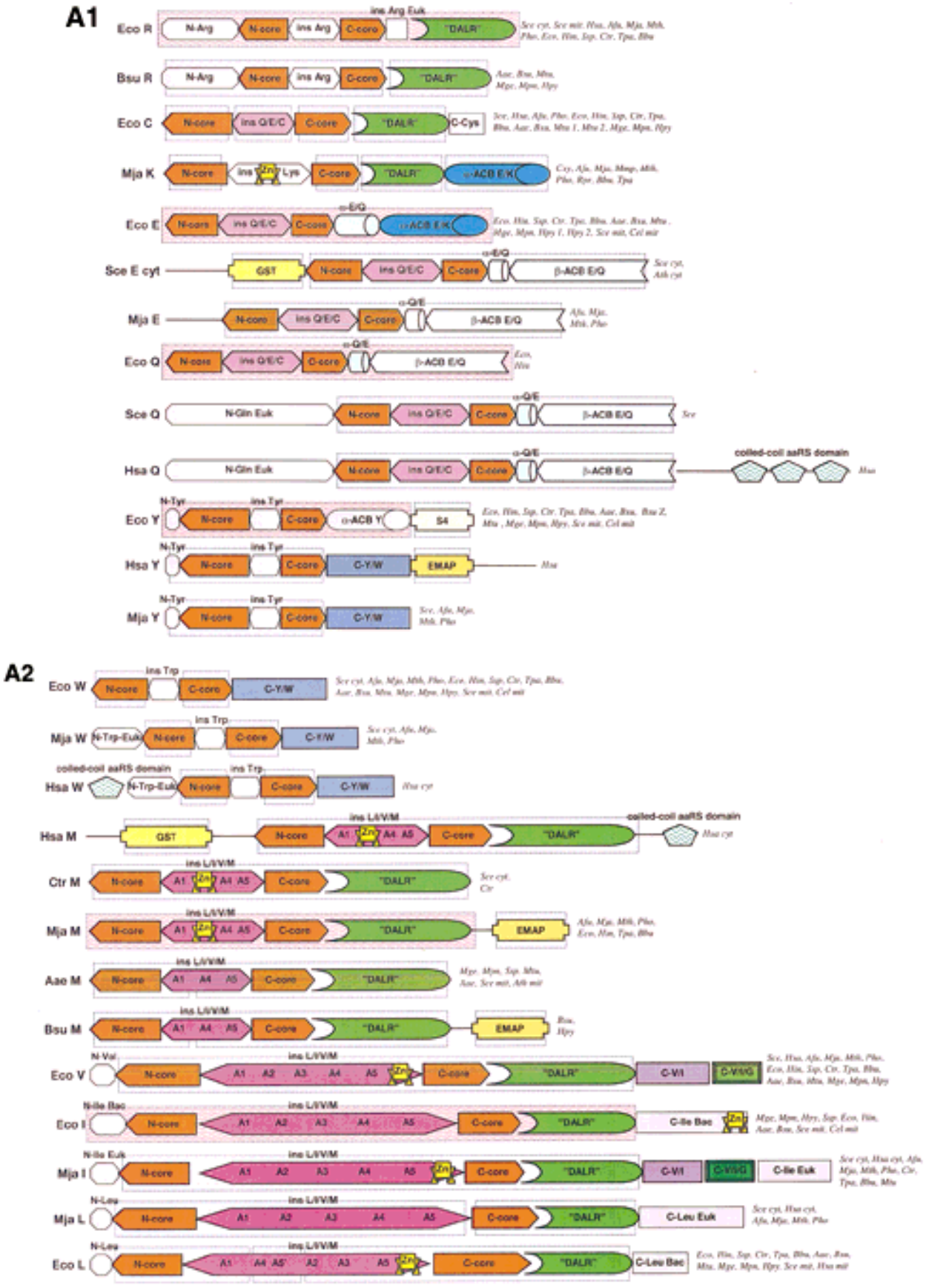

Figure 1 (*See p. 694 for legend.*)

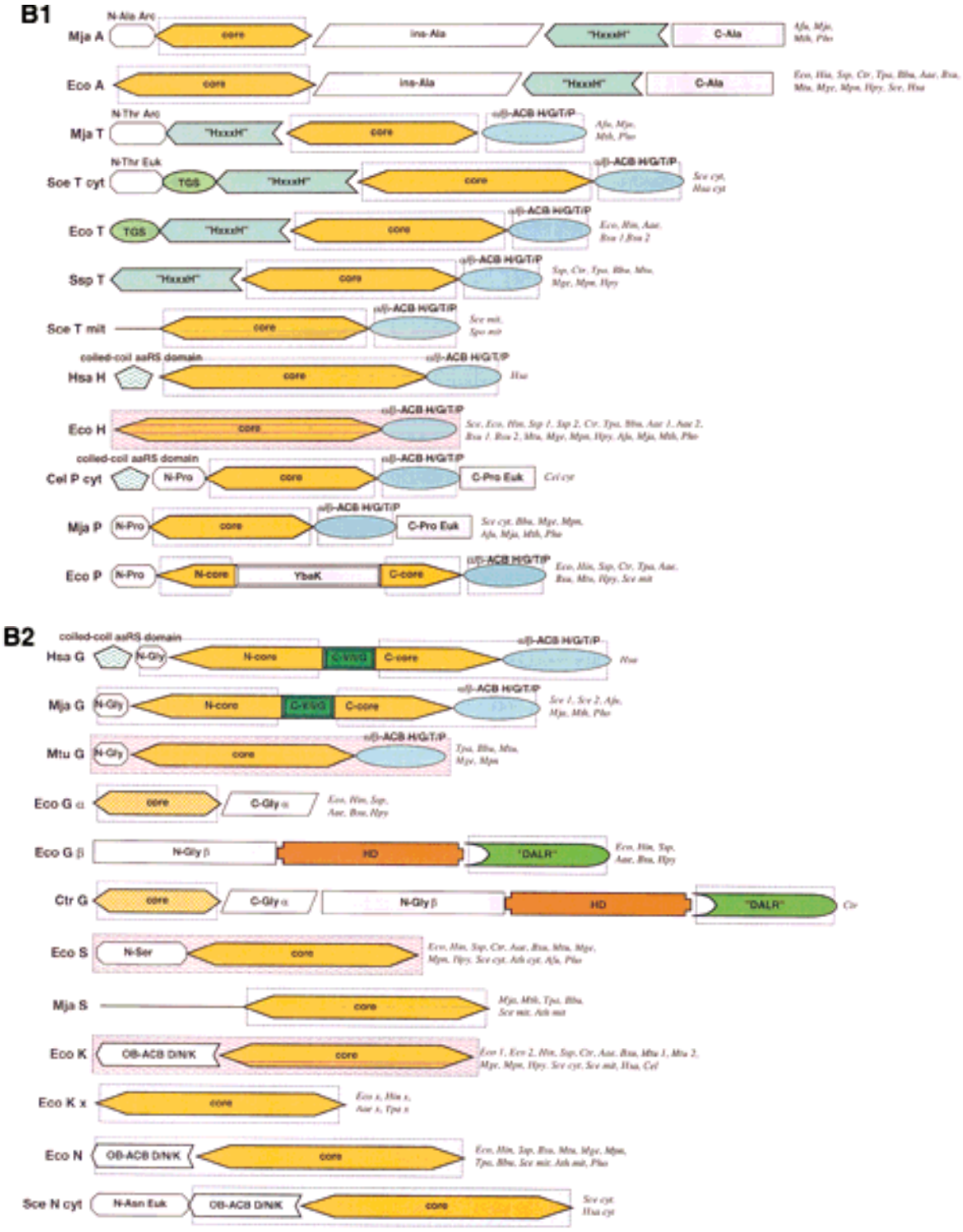

Figure 1 (*See p. 694 for legend.*)

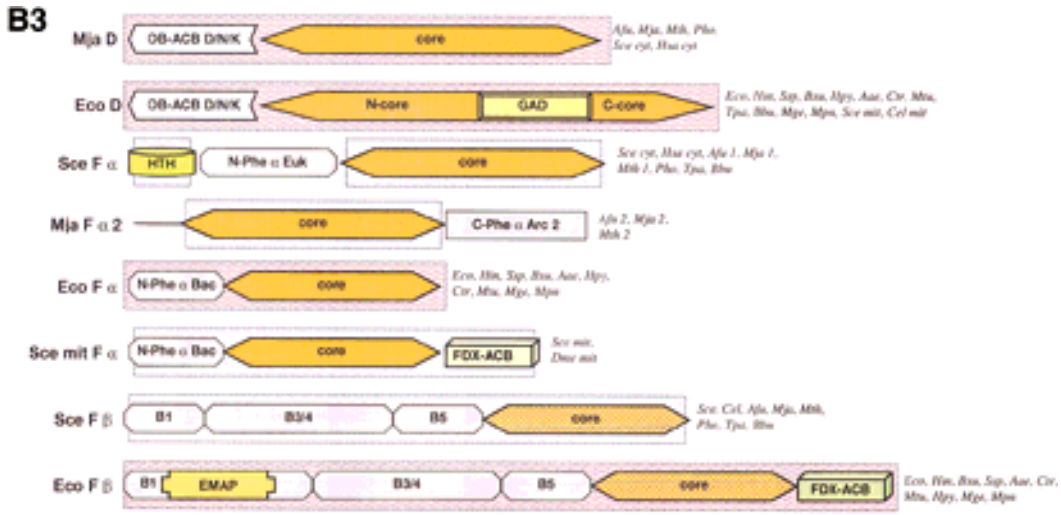

Figure 1 Domain architectures of aaRSs. (*A*) Class I aaRSs; (*B*) class II aaRSs. For each specificity, all detected distinct domain architectures are shown. The abbreviated names of the species names in which the given domain arrangement was observed are given to the right of each scheme. Domain name abbreviations: (A1–5) Distinct modules detectable in the large insert typical of the aaRS for aliphatic amino acids; (ACB) anticodon-binding domain; (FDX–ACB) ferredoxin-fold ACB; (GST) glutathione *S*-transferase; (ins) insert; (OB–ACB) OB-fold ACB (Zn) Zn ribbon motif; other domains are designated either after partially conserved sequence signatures (DALR and HxxxH) or after the aaRS specificities in which they are found (C-V/I, C-V/I/G). Pink background indicates domains for which three-dimensional structure is available; blue background shows domains for which the structure can be predicted on the basis of sequence similarity to structurally characterized domains. (Aae) *A. aeolicus*; (Afu) *A. fulgidus*; (Ath) *A. thaliana*; (Bbu) *B. burgdorferi*; (Bsu) *B. subtilis*; (Cel) *C. elegans*; (Csy) *C. symbiosum*; (Dme) *D. melanogaster*; (Hsa) *H. sapiens*; (Ctr) *C. trachomatis*; (Eco) *E. coli*; (Hin) *H. influenzae*; (Hpy) *H. pylori*; (Mja) *M. jannaschii*; (Mge) *M. genitalium*; (Mpn) *M. pneumoniae*; (Mth) *M. thermoautotrophicum*; (Mtu) *M. tuberculosis*; (Mpn) *M. pneumoniae*; (Pho) *P. horikoshii*; (Rpr) *R. prowazekii*; (Sce) *S. cerevisiae*; (Spo) *S. pombe*; (Ssp) *Synechocystis* sp. (Tpa) *T. pallidum*.

LysRS is especially interesting because this aaRS also contains an anticodon-binding domain shared with GluRS (Fig. 1A). The combination of these two domains may indicate a complex mode of anticodon binding by the LysRS.

Other connections between accessory domains are confined within class I or class II. In particular, there is a remarkable colinearity of the domain arrangements in class I aaRSs that are specific for aliphatic amino acids (Val, Ile, and Leu) and methionine. In addition to the carboxy-terminal DALR domain, these aaRSs share a large common insert in the core that contains five partially conserved motifs subject to deletion or rearrangement (Fig. 1A). Furthermore, all ValRS and subsets of aaRSs for each of the other three amino acids in this subclass of class I also contain an inserted Zn-ribbon module; a similar module is inserted also in class I LysRS (Fig. 1A). Another domain typical of class I is the insert shared by GluRS, GlnRS, and CysRS (Fig. 1A). In class II, the most common domains, after the core, are the α/β-structured anticodon-binding domain found in HisRS, ThrRS, and ProRS as well as eukaryotic

Figure 2 Previously undetected domain conservation in aaRS and proteins of other functions. (*A*) The GAD domain in bacterial AspRS and archaeal glutamyl-tRNA amidotrnasferases (GatB). (*Top* block of sequences) Archael GatB; (*bottom* block) bacterial AspRS. (*B*) The TGS domain in ThrRS, guanosine polyphosphatases (SpoT), OBG family GTPases, and uridine kinase (UDK) from Treponema. (*Top* block of sequences) ThrRS; (*bottom* block) other proteins containing the TGS domain. (*C*) The inactivated HD hydrolase domain in bacterial GlyRS β-subunit. (*Top* block of sequences) A selected subset of the HD superfamily hydrolases; (*bottom* block) β-subunits of bacterial GlyRS. (*D*) The winged helix–turn–helix domain in PheRS α-subunits from eukaryotes, archaea, and spirochaetes. (*Top* block of sequences) PheRS; (*bottom* block) a selected subset of other proteins containing the winged-HTH domain. The alignments were constructed on the basis of the PSI-BLAST results using the ClustalW program. The inclusion of each sequence in the alignments was statistically supported by PSI-BLAST results, with an e-value of at least 0.01. The *left* column includes the protein (gene) names, the abbreviated species name, and the gene identification (GI) nos. (following the underscore). A consensus derived using a 90% or an 85% cutoff is shown underneath the alignment, and the respective alignment columns are highlighted; (b) a big residue (E,K,R,I,L,M,F,Y,W); (h) a hydrophobic residue (A,C,F,I,L,M,V,W,Y); (s) a small residue (A,C,S,T,D,N,V,G,P); (u) a tiny residue (G,A,S); (o) a hydroxy residue (S,T); (p) a polar residue (D,E,H,K,N,Q,R,S,T); (c) a charged residue (K,R,D,E,H); (+) a positively charged residue (K,R,H). In *A,B,* and *D,* the numbers indicate the positions of the first and last residue of the aligned region in the respective protein sequence. The alignment in *C* consists of conserved blocks separated by variable spacers; the lengths of the spacers and the distances between the protein termini and the aligned regions are indicated by numbers. The secondary structure elements predicted using the PHD program, with the multiple alignment as the input (Rost and Sander 1994), is shown above the alignment in *A, B,* and *D*. [E(e)] indicates extended conformation (β-strand); [H(h)] α-helix; uppercase letters indicate the predictions made with a high level of confidence. In *C,* the line above the alignment indicates the predicted catalytic residues of the HD superfamily hydrolases (Aravind and Koonin 1998) that are replaced in the β-subunits of GlyRS. (Aae) *A. aeolicus*; (Af) *A. fulgidus*; (At) *A. thaliana*; (Bb) *B. burgdorferi*; (Bs) *B. subtilis*; (Ce) *C. elegans*; (Ct) *C. trachomatis*; (Dm) *D. melanogaster*; (Ec) *E. coli*; (Hi) *H. influenzae*; (Hs) *H. sapiens*; (Mj) *M. jannaschii*; (Mta) *M. thermoautotrophicum*; (Mtu) *M. tuberculosis*; (Ph) *P. horikoshii*; (Sp) *S. pombe*; (Sso) *S. solfataricus*; (Ssp) *Sy.* sp.; (Tm) *Th. maritima*; (Tp) *T. pallidum*; (VV) vaccinia virus.

and archaeal GlyRS and the OB-fold anticodon-binding domain present in AspRS, AsnRS, and LysRS (Fig. 1B). The other accessory domains are found in aaRSs of one or two specificities.

Several aaRSs share conserved domains with other classes of proteins, both involved in translation and performing very different functions (cf. Koonin et al. 1994; Simos et al. 1996,1998; Tas and Murray 1996; Markus et al. 1998; Aravind and Koonin 1999). In the course of the present study, we identified six additional domains that are shared by aaRSs and proteins of other functional classes. For each of these cases, the evolutionary relationship between the respective domains in the aaRSs and other proteins was supported by a sta-

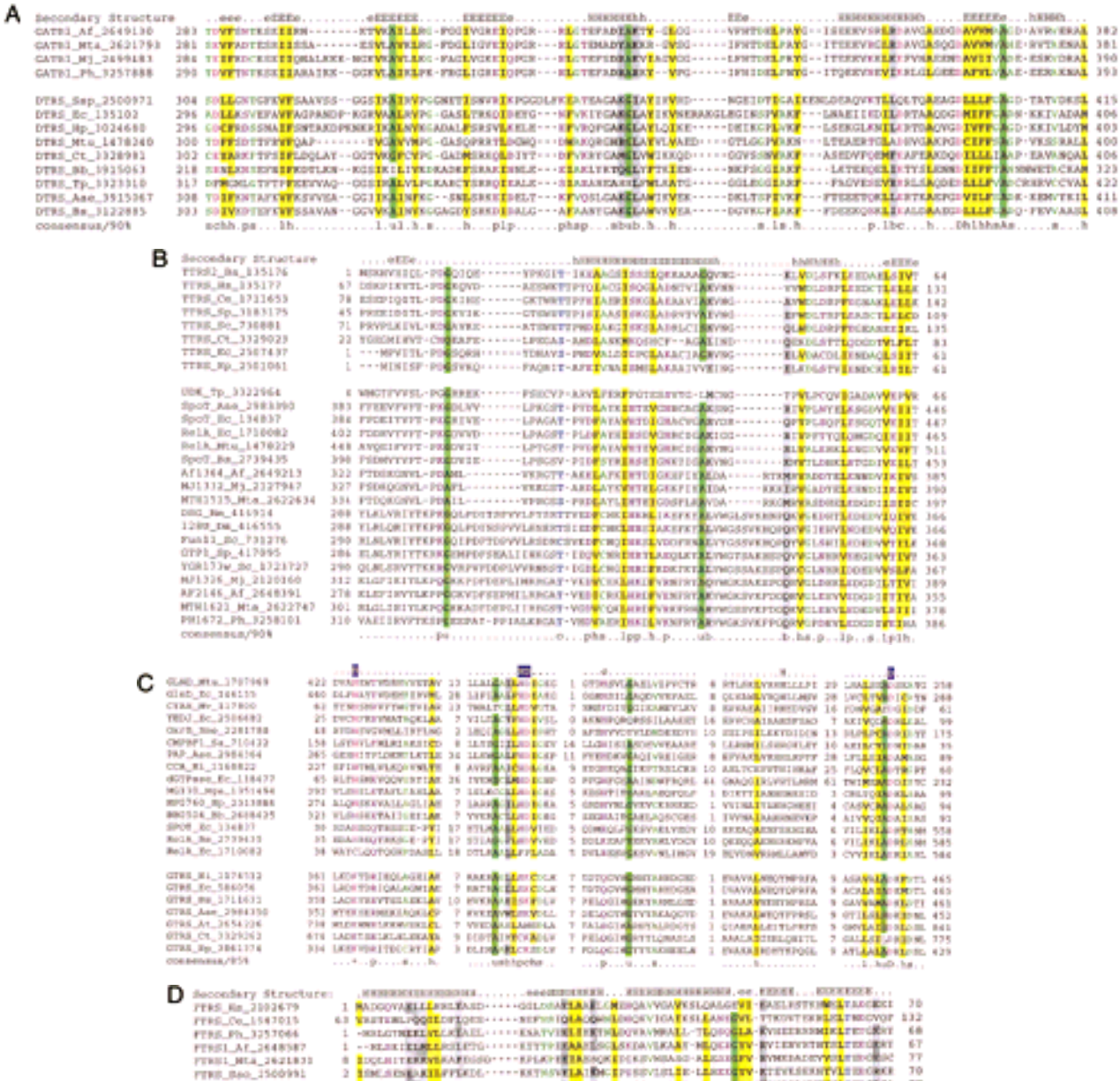

Figure 2 (*See facing page for legend.*)

Table 1. Phyletic Distribution of AspRS, AsnRS, GatB, and the GAD Domain

Species range	AspRS		AsnRS	GatB	
	with GAD domain	without GAD domain		with GAD domain	without GAD domain
Archaea *M. jannaschii, M. thermoautorphicum, A. fulgidus*	−	+	−	+	+
Archaea *P. horikoshii*	−	+	+	+	−
Bacteria *Proteobacteria, Cyanobacteria,* Spirochetes, Gram-positive bacteria	+	−	+	−	+
Bacteria *Mycobacteria, Helicobacter, Chlamydia, Aquifex*	+	−	−	−	+

tistically significant sequence similarity ($e < 0.01$ or better) as computed using the PSI-BLAST program.

The first unexpected finding involves a domain that is inserted in the core of bacterial AspRS and in the B subunit (GatB) of archaeal Glu-tRNA^{Gln} amidotransferases [hereinafter GAD domain, after GatB-AaRS-for-Asp (D)] (Figs. 1B and 2A). In archaeal GatB proteins, the GAD domain also forms an insert that is readily detectable by comparison with the bacterial counterparts (data not shown). The GAD domain contains ~120 amino acid residues and, as seen in the X-ray structure of the *Thermus thermophilus* AspRS, consists of an antiparallel β-sheet flanked by α-helices (Delarue et al. 1994) and resembles a circularly permuted ferredoxin-like fold (data not shown). GAD domain has been tentatively implicated in the stabilization of the interaction of the bacterial AspRS with the cognate tRNA (Delarue et al. 1994). This is generally compatible with the fact that GatB does not possess the transamidase activity (which resides in GatA) and is expected to be involved in tRNA recognition, although it may also be responsible for the ATPase activity of the complex (Curnow et al. 1997). The presence of the GAD domain in two different proteins that recognize tRNAs for acidic amino acids (Asp and Glu) suggests a specific role of this domain in the recognition of, and possibly discrimination between, these tRNAs. Given the presence of two versions of GatB—one with and one without the GAD domain—in most archaea (Table 1) and its presence in bacterial AspRS (but not GluRS), a simple hypothesis could be that GAD domain is responsible for the specific recognition of tRNA^{Asp} and its discrimination from tRNA^{Asn}. This, however, is hardly compatible with the presence of the GAD-containing GatB protein in the archaeon *Pyrococcus horikoshii* that also encodes an AsnRS (Table 1). Thus, GAD domain might recognize tRNA^{Glu} in archaea and tRNA^{Asp} in bacteria.

The second previously undetected domain is shared by eukaryotic and some of the bacterial ThrRSs, a distinct family of GTPases (the Obg family), and guanosine polyphosphate hydrolase (SpoT) and synthetase (RelA), which are involved in stringent response in bacteria (Cashel et al. 1996). We named it the TGS domain, after ThrRS, GTPase, and SpoT (Figs. 1B and 2B). Interestingly, TGS domain was detected also at the amino terminus of the uridine kinase from the spirochaete *Treponema pallidum* (but not any other organism, including the related spirochaete *Borrelia burgdorferi*) where it precedes the "HxxxH" domain, in an arrangement similar to that seen in ThrRS (Fig. 1B; see below). TGS is a small domain that consists of ~50 amino acid residues and is predicted to possess a predominantly β-sheet structure; this is one of the few domains in the aaRSs for which no structure has been determined so far (Fig. 1B). There is no direct information on the functions of the TGS domain, but its presence in two types of regulatory proteins (the GTPases and guanosine polyphosphate phosphohydrolases/synthetases) suggests a ligand (most likely nucleotide)-binding, regulatory role.

We observed that the β-subunit of bacterial GlyRS contains a domain that showed a distant but statistically significant similarity to the recently described HD superfamily of hydrolases [Figs. 1B and 2C; (Aravind and Koonin 1998)]. The principal predicted catalytic residues of the HD hydrolases, namely the histidine–aspartate doublet that is the namesake of the superfamily, are missing in GlyRS-β, although a carboxy-terminal aspartate also implicated in catalysis is conserved; this resembles the conservation pattern seen in the guanosine polyphosphate synthetases (RelA) [Fig. 2C (Aravind and Koonin 1998)]. The function of the HD domain in the β-subunit of GlyRS remains uncertain; it has been reported that the amino-terminal one-half of the β-subunit, along with the α-subunit, is required for the glycyl-adenylate formation (Toth and Schimmel 1990). An interesting aspect of these observations is that they make the β-subunit of the bacterial GlyRS the only aaRS subunit that does not contain the core domain of either class I or class II (Fig. 1A,B).

ThrRs and AlaRS share a domain that is typified by

the presence of two conserved histidines separated by three amino acid residues and was accordingly designated the HxxxH domain (Fig. 1B; data not shown). The HxxxH consists of ~120–140 amino acids and is predicted to possess a mixed β/α structure; along with the TGS domain, this is one of the remaining structurally uncharacterized domains in aaRSs (Fig. 1B). In addition to the aaRSs of two specificities, the HxxxH domain was found in four uncharacterized gene product (three from the archaeon *P. horikoshii* and one yeast) that contain additional sequences similar to those seen in AlaRS and seem to have evolved from the latter by gene truncation. A version of the HxxxH with a disrupted motif was detected in the ThrRS from Mycoplasma as well as in the uridine kinase from *T. pallidum* (see above). Finally, a fragment of the HxxxH domain is fused to the *Pseudomonas syringae* CmaT protein that, interestingly, is involved in nonribosomal peptide synthesis (Ullrich and Bender 1994). An HxxxH signature is generally typical of metal-dependent hydrolases, for example Zn-dependent proteases. The presence of a domain containing this motif in aaRSs may suggest a functionally important hydrolytic activity, for example, hydrolysis of mischarged aminoacyl-tRNAs.

We identified a winged helix-turn-helix (HTH) domain at the amino termini of the PheRS α-subunit from eukaryotes and archaea (including the crenarchaeon *Sulfolobus solfataricus* but with a highly modified version in *Methanococcus jannaschii*) and the spirochaetes (Figs. 1B and 2D). This domain is specifically related to the similar nucleic-acid-binding domains from double-stranded (ds)RNA adenosine deaminases, ribosomal protein S10, and the poxvirus dsRNA-binding protein E3L (Fig. 2D). The structure of the adenosine deaminase has been recently determined, and thus the winged-HTH structure of the amino-terminal domain of this protein, which has been shown to bind Z-DNA, was demonstrated experimentally (Schade et al. 1999). Given that some of the proteins containing this domain, for example, S10 and E3L, bind RNA, particularly dsRNA, this might be the likely function of the winged-HTH domain in PheRS. Specifically, it seems possible that this domain contributes to an unusual, for aaRSs, mode of tRNA binding via a stem.

Finally, we observed that a domain inserted in the core of the bacterial ProRS (Fig. 1B) is also represented by a family of small proteins found in several bacterial species (typified by the *E. coli* YbaK protein and accordingly designated "YbaK domain"; data not shown). The structure and function of this domain remain to be determined.

Taken together, all these observations reinforce the notion that aaRSs are prone to recruiting domains from other types of proteins and hence acquire additional functional capabilities. The readily recognizable domain recruitments in aaRSs typically are lineage specific and can be mapped to very different, ancient or relatively recent, stages of evolution. For example, given their near ubiquity in bacteria, the TGS domain, the S4 domain, and the GAD domain most likely became fused to the ThrRS, TyrRS, and AspRS, respectively, early in bacterial evolution (see also below). Other apparently ancient domain recruitments in aaRSs include the EMAP domain in MetRS and the winged-HTH domain in PheRS. The latter, for example, must have been present at the amino terminus of the PheRS α-subunit already in the common ancestor of archaea and eukaryotes. In contrast, the glutathione *S*-transferase domain and the small, coiled-coil interaction module appear to be relatively recent acquisitions because they are present in aaRSs of different specificities but exclusively within the animal lineage (Fig. 1).

Other domains that we now consider integral parts of aaRSs, such as those involved in anticodon binding (e.g., the DALR domain), might have evolved in the same fashion very early in evolution, but the sources are not readily identifiable anymore. "Horizontal evolution" of aaRSs, that is transfer of domains between aaRSs of different specificities, has been discussed (Delarue and Moras 1993). It does seem likely that the observed mosaic of domains in part has been generated by recombination between aaRS genes themselves, as opposed to independent acquisition of domains. The presence of the DALR domain that generally is typical of class I aaRSs in the β-subunit of GlyRS (see above) may be indicative of this type of an evolutionary event; this mode of dissemination also seems likely for the EMAP domain (Fig. 1A,B).

Reconstructing the Evolution of aaRSs

Phylogenetic Trees

We used the multiple alignments of the conserved portions of the aaRSs of all 20 specificities to generate distance matrices and construct phylogenetic trees using the neighbor-joining and Fitch-Margoliash methods. For each of these methods, 1000 bootstrap replications were performed, to evaluate the reliability of the results, and the consensus topology was derived. The consensus topologies for the neighbor-joining and Fitch-Margoliash methods were then combined (see Materials and Methods for details) to produce the final trees shown in Figure 3.

This procedure produces unrooted trees. Given that most aaRS specificities already have evolved in the LCA (see above), aaRS trees, in principle, can be rooted by jointly analyzing enzymes of two or more specificities. In practice, however, this approach cannot be used consistently because only for a few pairs of aaRSs of different specificities, can sufficiently long align-

ments suitable for tree construction be produced. Such conserved groups of aaRSs of different specificities include (1) different aliphatic amino acids; (2) tyrosine and tryptophan, and (3) lysine, aspartate, and asparagine. Construction of trees rooted by paralogy for the former two groups has been described (Brown and Doolittle 1995; Brown et al. 1997; Hashimoto et al. 1998). We used this approach only for the latter group, to resolve the complex evolutionary picture seen in AspRS and AsnRS (see below). Generally, the root position for aaRSs of all specificities was tentatively inferred using a combination of three criteria: (1) clustering by sequence similarity, (2) determination of the tree midpoint using the modified procedure described under Material and Methods, and (3) examination of unique features of domain architecture (synapomorphies). The first method is expected to give the correct rooting only under the molecular clock model, whereas the second method will infer the root correctly under a relaxed molecular clock assumption whereby the variation of the evolutionary rates along all tree branches is considered random (see Materials and Methods). Although caution is due in the interpretation of the results obtained under any version of molecular clock, the relaxed assumption seems likely to hold for the conserved portions of the aaRSs whose basic function remained unchanged throughout their evolutionary history. The approach based on synapomorphies, which is at least partially independent of, and complementary to, the other two methods, appears to be particularly valuable, and we discuss it in more detail.

Likely Synapomorphies in aaRSs and Their Use as Phylogenetic Markers

For many gene families, analysis of shared derived features (characters) of proteins, or *synapomorphies*, can be used as an important complement to the traditional, alignment-based phylogenetic tree analysis (e.g., Makarova et al. 1999). Primarily, such features are manifest as unique domain arrangements. Synapomorphies can be used to define monophyletic groups and may be helpful in establishing the root position because the root cannot lie within a monophyletic group

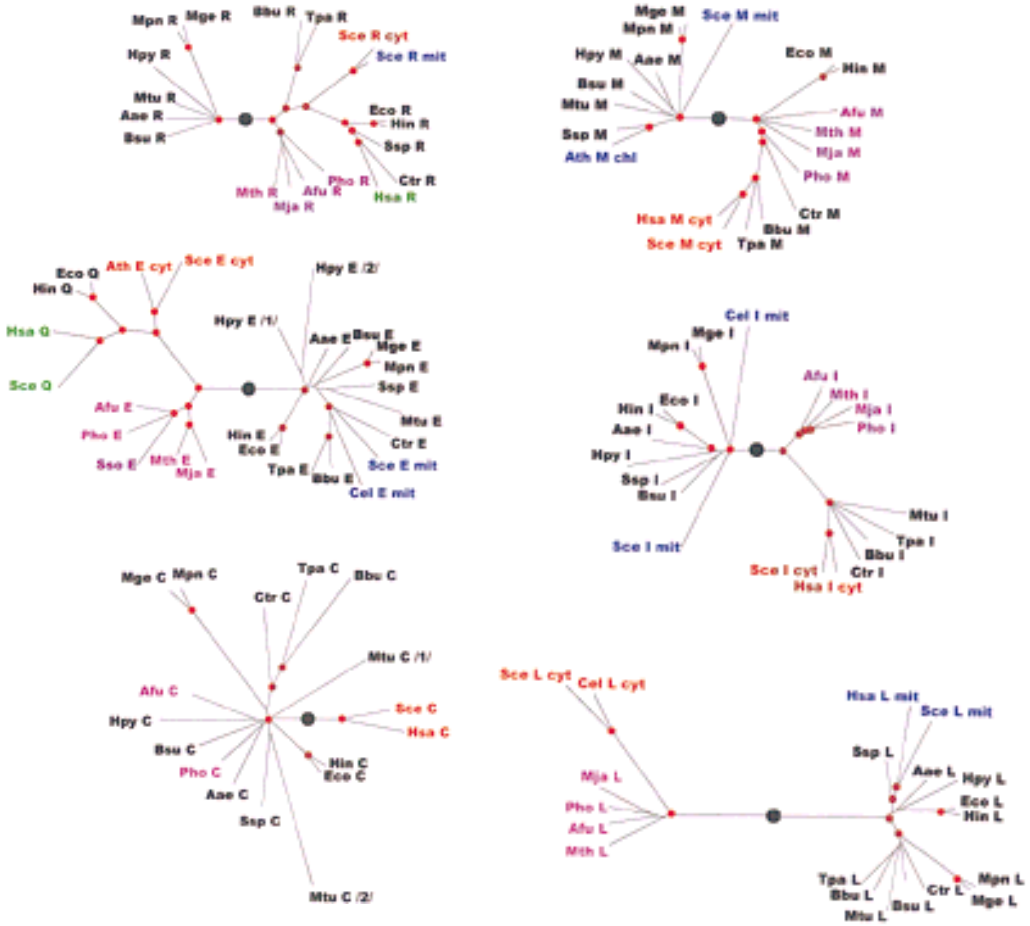

Figure 3 (*See p. 700 for legend.*)

defined by a synapomorphy. Using synapomorphies as phylogenetic markers requires distinguishing them, first, from primitive features that were already present in the common ancestor of the analyzed protein family, although they might have been lost in some lineages, and second, from independently acquired features. In cases when there are two distinct domain architectures within an aaRS specificity, one of these is likely to be a primitive feature and the other one a derived feature (synapomorphy). Deciding which is which is not straightforward and can be confidently done only when conserved features of domain architecture are seen in different aaRS specificities as discussed below (this is analogous to tree rooting by paralogy).

We attempted to systematically delineate the likely synapomorphies in aaRSs, to partition the aaRSs of the same specificity into likely monophyletic groups. For this purpose, domain architectures of aaRSs with the same and different specificities were compared in conjunction with clustering by sequence similarity and tree analysis using the Fitch-Margoliash and neighbor-joining methods.

With the exception of ValRS and CysRS, all ubiquitous aaRSs have more than one distinct domain architecture (Fig. 1A,B). Such distinctions do not exist in class I LysRS and in the bacterial-type GlyRS either, but these have limited phyletic distribution (Fig. 1A,B; see discussion below). The complete conservation of the elaborate domain architecture of ValRS, which consists of seven distinct domains, including the core (Fig. 1A), in all studied life forms seems unexpected given the diversity of domain organizations seen in the other aaRSs (see also discussion below).

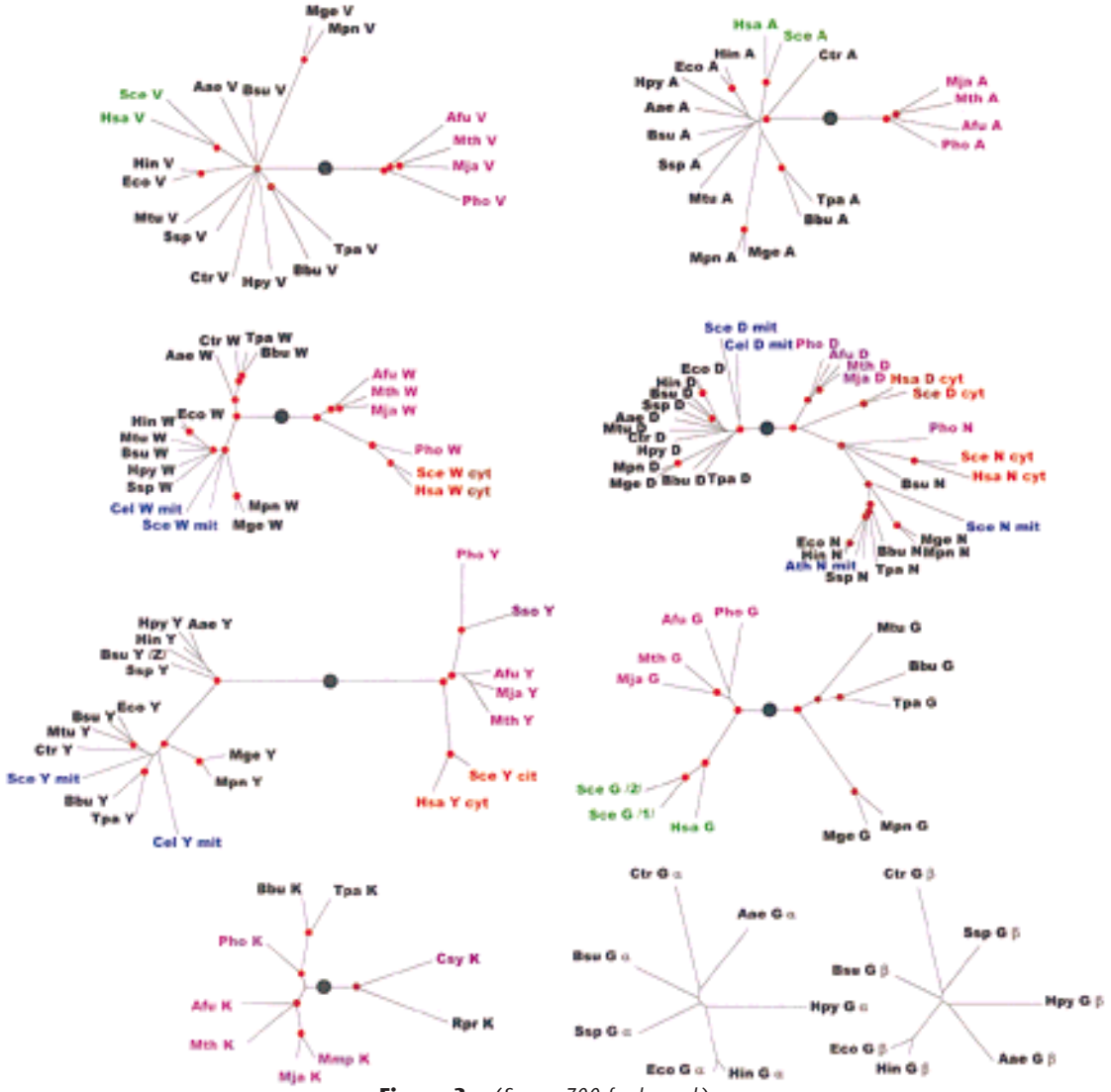

Figure 3 (*See p. 700 for legend.*)

For some of the aaRSs, synapomorphies appear unambiguous and allow us to easily detect distinct lines of descent. The most obvious of these are the two types of LysRS (see above) and GlyRS. As indicated above, the class I LysRS found in euryarchaea, spirochaetes, and rickettsia is unrelated to the type II enzyme present in eukaryotes, the rest of the bacteria, and the crenarchaea. The majority of bacteria possess a GlyRS that consists of two unrelated subunits (see also the discussion of the domain architecture of the β-subunit above) and is distinct from the enzyme found in eukaryotes, archaea, and a small subset of bacteria (Freist et al. 1996; Fig. 1B). The α-subunit of the bacterial GlyRS contains a modified class II core domain and is no more similar to the eukaryotic–archaeal GlyRS than it is to other class II aaRSs. Thus, there is no indication that these two types of GlyRS have a common origin.

These two exceptional cases apart, the synapomorphies seen in IleRS are most striking. Here, the distinction between the eukaryotic, archaeal, and a small subset of bacterial enzymes, on one hand, and the rest of the bacterial enzymes, on the other hand, involves

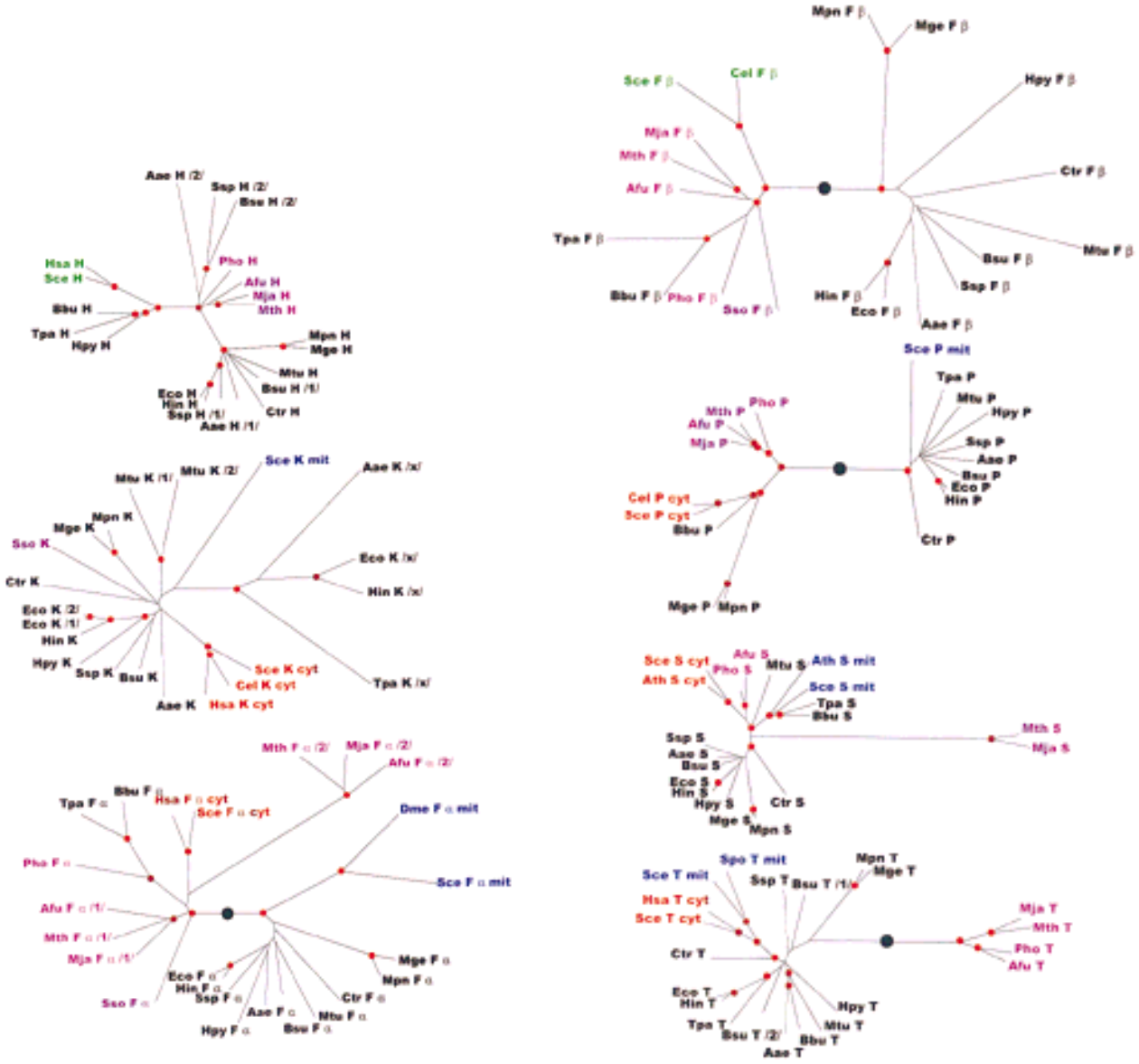

Figure 3 Phylogenetic trees for aaRSs. The trees are shown as unrooted, but a large black circle denotes the likely root position inferred by a combination of synapomorphy analysis, clustering by sequence similarity and midpoint rooting (see Table 2). Red circles mark statistically supported nodes (> 70% bootstrap support for both the neighbor-joining and the Fitch–Margoliash methods). Three-letter species labels are as in Fig. 1. Labels for archaeal proteins are shown in magenta; eukaryotic cytoplasmic in red; eukaryotic organellar in blue; eukaryotic without indication of origin in green; bacterial in black. Additional abbreviation: (Sso) *S. solfataricus*.

four distinct domains (modules). One of these (the Zn-ribbon) is located differently in the two sets of IleRSs, two others, namely the C-V/I domain and the C-V/I/G domain, are present only in the eukaryotic–archaeal subset, and finally, the more specific carboxy-terminal domains are conserved within each set but not between them (Fig. 1A). Notably, in this case, the arrangement of three of these domains, namely the Zn-ribbon and the C-V/I and C-V/I/G domains, is exactly the same in the eukaryotic–archaeal IleRS and in the ValRS (Fig. 1A). Thus, the ancestral domain architecture can be inferred with considerable confidence, and we are in a position to conclude that bacterial IleRSs have lost the C-V/I and C-V/I/G domains, whereas the Zn-ribbon has relocated in the course of bacterial evolution. In the same vein, a comparison of the two distinct domain architectures of the LeuRSs with those of the ValRSs and IleRSs suggests that the presence of the Zn-ribbon in the bacterial as opposed to archaeal–eukaryotic LeuRS is an ancestral feature, whereas the rearrangement of the modules in the large insert of bacterial LeuRS is derived (Fig. 1A). Another convincing synapomorphy is seen in the bacterial TyrRSs that possess a conserved arrangement of two domains (the α-helical anticodon-binding domain and the S4 domain) that are missing in the eukaryotic–archaeal set (Fig. 1A). Just as in the case of IleRS, it is possible to infer the ancestral state "by paralogy" because TrpRS shares a carboxy-terminal domain (C-Y/W) with the archaeal–eukaryotic but not bacterial TyrRS (Fig. 1A). More tentatively, it can be hypothesized that certain more complex domain architectures are more likely to be derived states than simpler ones, particularly when inserts in the core domain are involved. This is the case for ProRS, AspRS, and eukaryotic-type GlyRS (Fig. 1; Table 2).

The analysis of other aaRSs illustrates the distinction between those features of domain organization that are likely to be bona fide synapomorphies and those that do not seem to qualify. Consider, for example, MetRS, for which five distinct domain arrangements are discernible (Fig. 1A). The EMAP domain, which is present in the archaeal MetRS and those from several diverse groups of bacteria but not in other bacteria or eukaryotes (Fig. 1A), does not seem to be a useful marker for large-scale phylogenetic analysis. Its distribution does not at all reflect clustering of MetRS by sequence similarity (data not shown) or the topology of the trees constructed using the neighbor-joining and Fitch-Margoliash methods (Fig. 3). Amidst the bacteria, those MetRSs that contain EMAP domain do not form a compact group (Fig. 3). Thus, the phyletic distribution of the EMAP domain may be explained by lineage-specific losses, independent acquisitions, or, most likely, a combination thereof. The mobility of this domain is underscored by the fact that it has been detected also in subsets of TyrRS and PheRS (Fig. 1A,B). Among the other domains found in MetRS, the GST domain and the carboxy-terminal coiled-coil domain (Fig. 1A) could be valid phylogenetic markers, but these would be useful only to examine the evolution within the eukaryotic crown group. In contrast, the Zn-ribbon module that is inserted in the middle of the domain typical of the aliphatic aaRSs in the archaeal, eukaryotic, and some of the bacterial MetRSs (Fig. 1A) is a likely synapomorphy. The distribution of this motif correlates with the clustering by sequence similarity (data not shown) and with the monophyletic groups that are apparent from tree analysis (Fig. 3).

Altogether, unique features of domain architecture that are likely synapomorphies and thus may be valid phylogenetic markers, allowing us to establish or corroborate the primary evolutionary partitioning in the given set of aaRSs, were found for 14 of the 20 specificities (Fig. 1A,B; Table 2). Thus, at least in the case of aaRSs, comparative analysis of domain architectures is a major source of evolutionary information that must be carefully reconciled with other lines of evidence, to produce credible evolutionary scenarios.

Evolutionary Scenarios for aaRS

The most notable outcome of this analysis seems to be that, with only a few exceptions, the trees produced by the described procedures are readily interpretable in terms of relatively simple evolutionary scenarios (Fig. 3; Table 2). The most contentious issue in any phylogenetic analysis is the root position. The apparent synapomorphies do not directly indicate the root position; they only outline monophyletic groups. However, the correlation between partitioning produced by comparison of domain architectures, the results of clustering by sequence similarity, and modified midpoint rooting that was observed for the majority of the aaRS typically allows one to locate the root with considerable confidence (Fig. 3; Table 2).

We examined the trees and the domain architectures of the aaRSs with regard to the evolutionary relationships between the three major divisions of life—bacteria, archaea, and eukaryotes. Assuming the monophyly of each of these divisions and (for the moment) ignoring the possibility of horizontal gene transfer, three topologies of a rooted tree are possible: (1) B|A,E; (2) E|A,B, and (3) A|B,E (A = archaea, B = bacteria, E = eukaryotes; the vertical line indicates the root position). The predominant pattern in the aaRS phylogenies that is seen in 12 of the 20 specificities involves partitioning into two major groups, one of which includes archaea, eukaryotes, and a subset of bacteria, and the second one the rest (typically, the majority) of bacteria (Fig. 3; Table 2). The most likely position of the root typically is between these groups. Thus, these 12 phylogenies are best compatible with model 1,

Table 2. Evolutionary Scenarios for aaRSs

Specificity	Inferred evolutionary events (standard model is assumed unless indicated otherwise)	Support for root placement (see Fig. 1 for likely synapomorphies manifest in distinct domain architectures)[a]
Class I		
Tyrosine	horizontal transfer with displacement from Gram-positive bacteria to *E. coli*	synapomorphies in the carboxy-terminal region (ACB, S4 domain in bacteria); clustering; MMPR; compatible with rooting by paralogy (Brown et al. 1997).
Tryptophan	horizontal transfer from eukaryotes to the archaeal lineage including *P. horikoshii*	clustering; MMPR; compatible with rooting by paralogy (Brown et al. 1997).
Leucine	pure standard model, no interdivision horizontal transfer	synapomorphies in the large insert (A1–A5) typical of aliphatic aaRS (Zn-ribbon, module shuffling) and in the carboxy-terminal domain (see text); clustering; MMPR; compatible with rooting by paralogy (Brown and Doolittle 1995; Hashimoto et al. 1998).
Valine	displacement of ancestral eukaryotic enzyme by a bacterial, probably mitochondrial one (a single enzyme for the cytosol and the organelles in eukaryotes)	clustering; MMPR; rooting by paralogy; compatible with rooting by paralogy (Brown and Doolittle 1995; Hashimoto et al. 1998).
Isoleucine	horizontal transfer from eukaryotes to spirochaetes, *Chlamydia,* and Mycobacteria	multiple synapomorphies in the domain architecture (see text), clustering; MMPR; compatible with rooting by paralogy (Brown and Doolittle 1995; Hashimoto et al. 1998).
Methionine	(1) horizontal transfer from eukaryotes to spirochaetes and *Chlamydia;* (2) horizontal transfer from archaea(?) to Proteobacteria; (3) horizontal transfer from eukaryotes or bacteria (e.g., spirochaetes) to the archaeal lineage leading to *P. horikoshii.*	apparent synapomorphies in the large insert (A1–A5) typical of aliphatic aaRS (Zn-ribbon in archaea, eukaryotes and a subset of bacteria), clustering; MMPR.
Arginine	independent horizontal transfer from eukaryotes to (1) spirochaetes; (2) Proteobacteria, Cyanobacteria, *Chlamydia*	apparent synapomorphy—the small inserted domain between the core and DALR domains in archaea, eukaryotes, and a subset of bacteria; MMPR; clustering suggested a different root position, on the branch between yeast and the spirochaetes
Glutamate/glutamine	duplication of GluRS in eukaryotes, followed by switch of specificity to glutamine in one of the copies; horizontal transfer of GlnRS from eukaryotes to Proteobacteria; horizontal transfer of mitochondrial GluRS from eukaryotes to spirochaetes and Chlamydiae	apparent synapomorphies—unrelated ACB domains in (1) archaeal and eukaryotic GluRS and in all GlnRS in the core insertion domain (bulging α-helices); and (2) bacterial GluRS; clustering, MMPR.
Cysteine	uncertain; the standard model does not directly apply; possible ancient gene loss in archaea, followed by horizontal transfer from bacteria to some of the archaeal species	clustering; MMPR suggests several possible rootings; the root position remains uncertain.
Lysine-I	no eukaryotic representatives; independent horizontal transfer from archaea to (1) spirochaetes; (2) Rickettsia (from the *Cenarchaeum* lineage)	clustering, MMPR
Class II		
Alanine	displacement of ancestral eukaryotic enzyme by a bacterial, probably mitochondrial one (a single enzyme for the cytosol and the organelles in eukaryotes)	apparent synapomorphy—extra amino-terminal domain in archaeal AlaRs; clustering, MMPR
Threonine	duplication of mitochondrial ThrRS in eukaryotes, with displacement of the ancestral eukaryotic form	clustering, MMPR
Proline	horizontal transfer from eukaryotes to *Borrelia* (a spirochaete) and mycoplasmas	apparent synapomorphies—YbaK domain inserted within the core in bacterial ProRS and an extra carboxy-terminal domain in eukaryotes and bacteria; clustering, MMPR
Histidine	(1) horizontal transfer from eukaryotes to spirochaetes and *Helicobacter;* (2) horizontal transfer from archaea to *Aquifex, Synechocystis, Bacillus* without displacement of the ancestral bacterial form	root position not defined

Table 2. (*Continued*)

Specificity	Inferred evolutionary events (standard model is assumed unless indicated otherwise)	Support for root placement (see Fig. 1 for likely synapomorphies manifest in distinct domain architectures)[a]
Aspartate/asparagine	duplication of AspRS in eukaryotes, followed by switch of specificity to asparagine in one of the copies; ancient horizontal transfer of AsnRS from eukaryotes to bacteria	synapomorphy—insertion of GAD domain in the core of bacterial AspRS; clustering, paralogous routing using LysRS. MMPR suggest a root position between AspRS and AsnRS
Serine	complex picture. Anomalous rapid evolution in one of the archaeal lineages; horizontal transfer from eukaryotes or archaea to Mycobacteria; possible displacement of mitochondrial genes by ancestral ones in eukaryotes; horizontal transfer, apparently of eukaryotic mitochondrial genes, to spirochaetes	root position not defined
Phenylalanine α	horizontal transfer to spirochaetes, probably from archaea	apparent synapomorphies in the amino-terminal region—HTH domain and an additional conserved domain in archaea, eukaryotes, and spirochaetes, a distinct amino-terminal domain in bacteria; clustering, MMPR
Phenylalanine α–β	horizontal transfer to spirochaetes, probably from archaea	apparent synapomorphies in domain architecture—EMAP domain and FDX–ACB domain in bacteria; clustering, MMPR
Lysine-II	no archaeal representatives; horizontal exchange of type X *LysRS* gene between Proteobacteria, Aquifex, Treponema	root position not defined
Glycine-1	Early horizontal transfer from eukaryotes, archaea or a common ancestor thereof to spirochetes, Mycobacteria, Mycoplasmas; alternatively, an ancestral form displaced in most bacteria (see text)	apparent synapomorphy—insert in the core domain in archaea and eukaryotes; clustering, MMPR
Glycine-2 α–β	bacterial-only; phylogeny uncertain	root position not defined

[a](ACB) Anticodon-binding domain; (FDX–ACB) ferredoxin-fold ACB; (MMPR) modified midpoint rooting.

which has been aptly designated the standard model by Doolittle and Handy (1998). The conclusion that the standard model is best compatible with the data rests on some semblance of "relaxed molecular clock" being valid. Most of the aaRS trees contain a long branch separating archaea–eukaryotes (in many cases, with the admixture of several bacterial species) and the bulk of bacteria; this is where the procedures we used typically place the root, and this is supported by the analysis of likely synapomorphies (Fig. 3; Table 2). However, should this long branch correspond instead to a systematic, major increase in the rate of aaRS evolution at the base of the bacterial trunk, it would be impossible to rule out topologies 2 and 3. In that case, the observed distinctions in domain architectures would be interpreted to indicate that the archaeal–eukaryotic architecture is the primitive state, whereas the bacterial architecture is the derived state (synapomorphy).

Examination of the aaRS trees lends no support to evolutionary schemes that postulate the origin of eukaryotes from a particular subdivision of archaea, such as the eocyte hypothesis (Rivera and Lake 1992) or the hydrogen hypothesis (Martin and Müller 1998). No specific association was seen between eukaryotic aaRSs and those from Crenarchaeota as suggested by the first of these schemes or those from methanogens as implied by the second one. It should be further noticed that the standard model is supported by the results of rooting by paralogy where such are available, namely for aliphatic amino acid aaRSs (Brown and Doolittle 1995; Hashimoto et al. 1998), tyrosine and tryptophan (Brown et al. 1997), and aspartate–asparagine and lysine (see below).

The standard model, however, requires major amendments to account for the topology of the aaRS trees. There are only three trees, curiously all from class I, that conform to this model precisely, namely those for LeuRS, TyrRS, and TrpRS (this does not rule out interesting and unusual events in the evolution of these aaRSs; see Table 2; discussion below). The remaining trees fall into two categories: (1) those in which eukaryotic aaRSs cluster with the bacterial ones, to the exclusion of the archaea, namely ValRS, AlaRS, and ThrRS, and (2) those in which varying subsets of bacterial aaRSs invade the eukaryotic–archaeal cluster—all the rest except for LysRS, CysRS, and HisRS (Fig. 3). Class I LysRS is seen only in archaea and a small subset of bacteria (Fig. 3), so, by definition, the standard model (or any alternative model) does not apply. Class II LysRS so far was found in only one archaeal species, the crenarchaeon *S. solfataricus*; the *Sulfolobus* LysRS clearly groups with the bacterial subtree (Fig. 3). CysRS so far has been detected in only two

archaeal species (see above). The CysRS tree is poorly resolved, and there are no synapomorphies to complement it, but both clustering and modified midpoint rooting methods suggest a root position between the eukaryotic branch and the rest of the tree that includes the two archaeal CysRSs along with the bacterial ones (Fig. 3). The HisRS tree does not show a clear separation of the eukaryotic–archaeal and bacterial clusters but rather contains a trifurcation in which archaea and eukaryotes are equidistant from each other and from the bulk of bacteria; thus, this tree is not incompatible with the standard model although it lends no direct support to it (Fig. 3).

The eukaryotic–bacterial affinity seen in three aaRSs is readily explained by displacement of the original eukaryotic gene by a cognate bacterial version, in all likelihood, the mitochondrial gene transferred to the nuclear genome (although, just as for the standard model and the alternative models discussed above, the possibility of a major acceleration of evolution in the archaeal lineage cannot be formally ruled out as the underlying basis of the observed topology). In one case, that of ThrRS, this has been preceded by a duplication of the mitochondrial gene (Fig. 3). Conversely, displacement of the mitochondrial enzyme by the ancestral eukaryotic one seems to have occurred in the evolution of HisRS and SerRS, in the latter case following a duplication (Fig. 3).

Examination of the emerging complete evolutionary picture (Fig. 3; Tables 2 and 3) seems to suggest horizontal gene transfer, rather than lineage-specific gene loss, as the principal explanation for the anomalies in the evolution of aaRSs. Given that clustering of a subset of bacteria with eukaryotes (and/or archaea) is observed for the majority of the aaRSs, the lineage-specific gene loss theory would imply that the LCA contained diverged duplicates of many, if not all, of the aaRS genes, which have been differentially lost in different lineages during subsequent evolution. One would assume, however, that these diverged duplicates of aaRSs have been fixed by selection in the lineage leading to the LCA as a result of the adaptation of the two versions to distinct functional niches. Should that be the case, there would be selective pressure to maintain both versions, along with likely advantages of shedding one, and we would be sure to expect relics of the original duplication persisting in at least some species and some aaRS specificities. Such traces, however, are conspicuously missing.

It is instructive to consider two cases that, at a superficial glance, might have been considered evidence in support of the primordial duplication theory. The first one involves the two unrelated types of LysRS, one of which belongs to class I and the other one to class II. It has been noted that if an organism was found that encoded both types, this would support the differential loss theory (Doolittle and Handy 1998). A genome of such an organism is available—the spirochaete *T. pallidum*. A closer analysis shows, however, that *T. pallidum* encodes a distinct type of class II LysRS—the small form comprised of the core domain alone—that is present, in addition to the typical bacterial LysRS, in γ-proteobacteria, and *Aquifex* (Figs. 1A and 3). Under the differential loss theory, one would be forced to conclude that the LCA encoded three LysRS—the class I enzyme and two distinct forms of the class II enzyme. Evolution of the truncated version in one of the bacterial lineages, with its subsequent

Table 3. Apparent Horizontal Transfer of Eukaryotic and Archaeal aaRS Genes into Different Bacterial Lineages

Bacterial group	Horizontally transferred *aaRS* genes	Source
Spirochetes	Pro (*Borrelia* only), Ile, Met, Arg, His, Asn	eukaryotic
	Ser, Glu	eukaryotic mitochondrial
	Lys-I, Phe α–β	archaeal
	Gly-1	eukaryotic or archaeal
	Lys-II (*Treponema* only)	bacterial (γ-Proteobacteria?)
Chlamydia	Ile, Met, Arg, Asn??	eukaryotic
	Glu	eukaryotic mitochondrial
Mycobacteria	Ile, Asn??	eukaryotic
	Ser, Gly(1)	eukaryotic or archaeal
Mycoplasmas	Pro, Asn	eukaryotic
	Gly(1)	eukaryotic or archaeal
γ-Proteobacteria	Asn, Gln	eukaryotic
	Met	archaeal(?)
Helicobacter	His	eukaryotic
Cyanobacteria	Arg, Asn	eukaryotic
	His (2nd form, no displacement)	archaeal
Bacillus	Asn	eukaryotic
	His (2nd form, no displacement)	archaeal
Aquifex	His (2nd form, no displacement)	archaeal

dissemination by horizontal transfer, seems to be a more realistic explanation for the observed phyletic distribution of LysRS.

The presence of two versions of HisRS in *Aquifex*, *Synechocystis*, and *Bacillus* also might appear to support the differential gene loss theory. The HisRS tree is more difficult to interpret than those for other aaRSs because of the uncertainty of the root position (Fig. 3). Nevertheless, assuming that the standard model still applies, the differential loss scenario predicts that whereas one of the HisRS in *Aquifex*, *Synechocystis*, and *Bacillus* should be a typical bacterial form, the other one—inherited directly from the LCA—should be equidistant from the archaeal and eukaryotic orthologs. In reality, however, reliable clustering of this second form with the archaeal HisRS was observed (Fig. 3), which again makes horizontal transfer, in this case from an archaeal source, the most likely explanation.

An equally strong argument for the major role of horizontal gene transfer, as opposed to differential loss, in aaRS evolution is the nonrandomness of the set of bacterial species that invade the archaeal–eukaryotic part of the phylogenetic trees for the aaRSs (Fig. 3; Table 3). The main contribution to this invasion is from bacterial groups that include intimate parasites and symbionts of eukaryotes, particularly spirochaetes and chlamydiae (Table 3). The differential gene loss theory offers no explanation why these groups of bacteria should have lost the aaRS versions retained by the majority of bacteria. In contrast, it is obvious that, compared with other bacteria, these organisms had a greater opportunity to acquire eukaryotic genes because of their long-term and intimate contact with the eukaryotic hosts.

Thus, multiple horizontal gene transfers, typically resulting in the displacement of the original aaRS genes in the recipient lineage, seem to have made the principal contributions to the deviations of the phylogenetic trees for the aaRSs from the standard model. It must be emphasized, however, that should this model prove to be wrong (see above), this would not affect the conclusion that these horizontal transfer events occurred in the course of evolution of aaRSs. Examination of the tree topologies in Figure 3 makes it clear that should the root for most of the trees lie, for example, on the branch connecting eukaryotes and archaea (model 2 above), the statistically supported clustering of subsets of bacterial aaRSs with eukaryotes still remains to be accounted for, horizontal gene transfer being the most likely explanation.

Most of the tree topologies are readily explained by a small number of horizontal transfer events; three trees, namely, MetRS, ArgRS, and HisRS, present a complex but seemingly interpretable picture, and two—SerRS and CysRS—are hard to interpret (Table 2). It has been suggested that the anomalies observed in some of the aaRS trees, particularly for IleRS, can be explained by just one horizontal gene transfer from eukaryotes, with subsequent dissemination among bacteria (Shiba et al. 1997; Brown et al. 1998; Doolittle and Handy 1998). This is a plausible idea that is compatible with the reliable clustering of all bacterial species that are suspected to have acquired the respective eukaryotic gene in the IleRS tree and in the HisRS tree (Fig. 3). Furthermore, the bacterial groups that appear to be most prone to horizontal transfer from eukaryotes—the spirochaetes and chlamydiae—also form clusters in the CysRS and TrpRS trees, which suggests gene exchange between them, although in these cases, horizontal transfer from eukaryotes is not suspected (Fig. 3). The dissemination of the eukaryotic-type IleRS on plasmids, which renders bacteria resistant to the antibiotic mupirocin, explains not only the mechanism of, but also the likely selective pressure behind at least some of these interbacterial horizontal gene transfers (Hodgson et al. 1994; Sassanfar et al. 1996; Brown et al. 1998). The topologies of the trees for MetRS, ArgRS, and Asp–AsnRS, however, are not readily compatible with this possibility and, rather, suggest multiple transfers of eukaryotic genes into different bacteria (Fig. 3; Table 2).

Gene transfer from archaea to bacteria has been much less prominent, at least amidst the bacterial taxa that have been sampled so far (Table 2). The apparent transfer of HisRS from archaea to bacteria has already been discussed. Other events of this type involve class I LysRS, PheRS, and possibly MetRS (Table 2). Given its ubiquity in archaea, sporadic presence in bacteria, and apparent absence in eukaryotes, it seems most likely that class I LysRS evolved in archaea and has been horizontally transferred to bacteria. Notably, the tree topology, which is strongly supported by bootstrap analysis, suggests two independent transfer events—from euryarchaea to the spirochaetes and from Crenarchaea to Rickettsiae (Fig. 3). The notable aspect of the evolutionary scenario for PheRS is that in most bacteria, including the spirochaetes, the genes for α and β subunits form an operon, whereas in the archaea, which apparently donated both genes to the spirochaetes, they are not adjacent. It appears likely that the operon organization is ancestral, and an archaeon containing this operon might be found eventually.

Other types of interdivision transfer appear to be rare. Horizontal gene transfer from bacteria to archaea seems to be a distinct possibility only for CysRS, which is present in two of the four completely sequenced archaeal genomes, AsnRS, so far identified only in *P. horikoshii*, and class II LysRS that might have been acquired in the *Sulfolobus* lineage. In addition, the TrpRS of *P. horikoshii* apparently has been acquired from eukaryotes (Fig. 3). This limited extent of aaRS gene exchange between bacteria and archaea appears rather unex-

pected, given the prominence of horizontal transfer of aaRS genes from eukaryotes to bacteria, and also the apparently considerable exchange of other genes between archaea and bacteria (Koonin et al. 1997; Aravind and Koonin 1998; Makarova et al. 1999). The most straightforward explanation, which allows direct experimental verification, is that bacterial aaRSs are generally poorly compatible with archaeal tRNAs. Of course, a cautionary note regarding the small available sampling of complete archaeal genomes, which all come from thermophilic species, also applies to this direction of horizontal gene transfer.

Unlike the relationship between the three primary divisions of life that could be resolved for the majority of aaRSs in support of a modified standard model, no consistent, large-scale bacterial phylogeny emerged from the aaRS trees. This is in itself not surprising because inconsistent and unreliable tree topologies have been observed frequently for different bacterial genes. In the majority of the aaRS trees, the "true bacterial" part (i.e., those bacterial aaRSs that have not been horizontally transferred from eukaryotes or archaea as discussed above) shows, more or less, a star topology, with no or little statistical support for any particular relationship between the major lineages (Fig. 3). The trees for TyrRS, TrpRS, and LeuRS are exceptional in that strongly supported—but different in each case—partitioning of the bacteria into two clusters is observed (Fig. 3). In the rest of the trees, the only clusters that are consistently seen are the terminal branches, namely, the two species of γ-proteobacteria (*E. coli* and *Haemophilus influenzae*), spirochaetes (*B. burgdorferi* and *T. pallidum*), and mycoplasmas (*Mycoplasma genitalium* and *Mycoplasma pneumoniae*). Interestingly, however, even these relatively close affinities are violated in some of the aaRS trees. Thus, *E. coli* and *H. influenzae* behave differently in the TyrRS tree, whereas the two spirochaetes show different affinities in the ProRS and ThrRS trees, in addition to the aforementioned presence of the truncated form of Class II LysRS in *Treponema* but not in *Borrelia* (Fig. 3). In each of these cases, the members of the respective pair of related species cluster, with a good bootstrap support, with other, distantly related bacteria or with eukaryotes. Thus, in the ThrRS tree, the *Treponema* protein clusters with γ-proteobacteria, whereas the one from *Borrelia* clusters with *Aquifex* and *Mycobacterium* (Fig. 3). In the ProRS tree, there is strongly supported clustering of *Borrelia* with eukaryotes and *Treponema* with other bacteria (Fig. 3); the latter association is corroborated by the insertion of the YbaK domain that is a hallmark of bacterial ProRS and is present in *Treponema* but not in *Borrelia*. Horizontal transfer of the eukaryotic ProRS gene into the *Borrelia* lineage subsequent to its divergence from *Treponema*, followed by the elimination of the typical bacterial gene, might explain it.

Other unexpected, but statistically supported, bacterial clusters are seen in the trees for AspRS (*Bacillus–Synechocystis*) and HisRS (spirochaetes–*Helicobacter*); clustering of spirochaetes with chlamydiae, mentioned above, belongs in the same category. These observations seem to indicate horizontal transfer of at least some of the aaRS genes between distant bacterial species. Additional, more ancient gene transfer events might be obscured by the star topology.

Our phylogenetic analysis may clarify the evolutionary scenario for AspRS and AsnRS. Eukaryotes encode both a cytoplasmic and a mitochondrial aaRS for each of these amino acids; archaea typically lack AsnRS (so far the only exception is *P. horikoshii*) and incorporate asparagine into proteins via the transamidation route, whereas the majority of bacteria encode AsnRS (Curnow et al. 1996; Shiba et al. 1998). The insertion of the GAD domain identifies bacterial aaRSs as a likely monophyletic group. Clustering by sequence similarity suggested the root position between this group and the rest of the AspRSs together with AsnRS. The midpoint procedure, however, placed the root between all AspRSs and AsnRSs. To resolve the ambiguity, we aligned the sequence of class II LysRS with those of AspRS and AsnRS and rooted the tree using the LysRS as an outgroup. Under this approach, the root was confidently placed between bacterial AspRS and the rest of the AsxRSs (Fig. 3; data not shown). Thus, the most likely scenario is that AsnRS originally evolved by duplication of eukaryotic AspRS, which was followed by horizontal transfer into bacteria, perhaps with subsequent dissemination among bacterial species, and at least one archaeon (Fig. 3; Table 2). This scenario is similar to that for GluRS and GlnRS (Siatecka et al. 1998; Fig. 3) but different from the one recently proposed for AsnRS, which postulated its origin by duplication of the AspRS gene early in bacterial evolution (Shiba et al. 1998). It remains unclear why the topologies of the AspRS and AsnRS trees observed in our analysis and in that of Shiba and coworkers (1998) were different; differences in the alignments are likely to contribute.

The case of the eukaryotic-type GlyRS is particularly interesting. Here, both the analysis of domain architectures (a unique insert in the core of the eukaryotic and archaeal proteins) clustering and modified midpoint rooting procedures suggested the likely position of the root between archaea–eukaryotes and bacteria (Fig. 3). However, only a minority of bacterial species possess this form of GlyRS. It appears that either a horizontal transfer of the archaeal–eukaryotic GlyRS to bacteria occurred very early during evolution or this is the ancestral GlyRS that has been displaced by a newly evolved form in the majority of bacteria.

Evolutionary scenarios for CysRS and SerRS remain uncertain. There is a notable correlation between the

absence of CysRS and the presence of an unusual, highly diverged SerRS in some of the archaea, namely the methanogens *M. jannaschii* and *Methanobacterium thermoautotrophicum* (Fig. 3). The properties of this unique SerRS and the pathway of cysteine incorporation into proteins in these archaea remain to be investigated experimentally. The SerRS from the other two archaeal species reliably cluster with the eukaryotes and a small subset of bacteria (Fig. 3). A recent study by others also reported this dramatic difference between the two types of archaeal SerRS as well as clustering of the SerRS from the methanogens with those from Gram-positive bacteria and Cyanobacteria (Lenhard et al. 1999); our analysis failed to provide support for the latter grouping.

Conclusions

Comparison of the complete sets of aaRSs from diverse species of bacteria, archaea, and eukaryotes reveals a number of unique domain architectures. Despite numerous structural studies on aaRSs, several previously undetected domains could be identified using improved methods of sequence analysis. The exact functions of these domains and the mode of their interaction with the aaRS core remain to be determined by combination of structural and biochemical analyses. Some of the distinct domain arrangements appear to be synapomorphies, that is, they define monophyletic groups within a given aaRS specificity.

Combined with traditional phylogenetic trees, analysis of these synapomorphies suggests relatively simple evolutionary scenarios for most of the aaRSs. All these scenarios are based on the standard model of evolution for the translation system, which postulates an original radiation of bacteria and the common ancestor of archaea and eukaryotes. This standard model is compatible with the results of the phylogenetic analysis of aaRSs, both qualitatively—at the level of synapomorphies—and quantitatively—in terms of the statistically supported topology of phylogenetic trees. However, alternative models for the relationships between bacteria, archaea, and eukaryotes cannot be ruled out if a major, systematic increase in the evolutionary rates at the base of the bacterial subtrees is postulated.

Regardless of the exact model of relationships between bacteria, archaea, and eukaryotes, phylogenetic analysis makes it clear that evolution of aaRSs involved a variety of horizontal gene transfers. The principal types of such events are transfer of eukaryotic aaRS genes into bacteria, resulting in the displacement of the respective ancestral bacterial genes, and displacement of original eukaryotic genes by mitochondrial genes transferred to the nuclear genome. Instances of likely horizontal transfer of aaRS genes from archaea to bacteria also were detected but are less common. There were no clear indications of horizontal transfer of aaRS genes from bacteria to archaea, although two likely cases of a eukaryotic gene being acquired by an archaeon were detected. In addition, for several aaRSs, there were strong indications of gene transfer between major bacterial lineages, and it appears that other events of this type might be obscured by the star topology of the bacterial trees.

The influx of eukaryotic aaRS genes into the bacterial world has been nonrandom. The fraction of transferred eukaryotic genes is the greatest in bacterial groups that consist predominantly or exclusively of parasites or symbionts, particularly the spirochaetes. Thus, horizontal gene transfer seems to have been a major force in the evolution of aaRSs, but some routes have been strongly favored, (e.g. from eukaryotes to spirochaetes), whereas others might have been (nearly) prohibited (from bacteria to archaea). Further genome sequencing, for example, of nonthermophilic and particularly symbiotic archaea, should be revealing in terms of the nature of these preferences and restrictions. It hopefully will become clear which of them simply correlate with the intensity of contact between two particular taxa and which stem from intrinsic features of the translation system, such as compatibility (or lack thereof) between aaRSs and the cognate tRNAs.

METHODS

Databases and the aaRS Sequence Set

The databases used in this study were the nonredundant database (NR) at the NCBI (NIH, Bethesda, MD) and a collection of aaRS sequences from completely sequenced genomes. The latter were initially extracted from the Genomes division of the Entrez system (http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html) using the available genome annotation. Additionally, all the protein sequences from complete genomes were searched (see below) using the *E. coli* aaRS sequences as queries, to detect any aaRS homologs that might have been misannotated. The aaRS sequence set used in this analysis included the entire complement of aaRSs from 12 complete bacterial genomes (*E. coli*, *H. influenzae*, *Helicobacter pylori*, *M. genitalium*, *M. pneumoniae*, *Bacillus subtilis*, *Chlamydia trachomatis*, *B. burgdorferi*, *T. pallidum*, *Mycobacterium tuberculosis*, *Synechocystis sp.*, *Aquifex aeolicus*), four archaeal genomes (*M. jannaschii*, *M. thermoautotrophicum*, *Archaeoglobus fulgidus*, *P. horikoshii*), and one eukaryotic genome, that of the yeast *Saccharomyces cerevisiae*; in addition, all the available aaRS sequences from crenarchaeota (four from *S. solfataricus* and one from *Cenarchaeum symbiosum*) were included. An attempt to use the full aaRS complement from the other eukaryotic genome that recently has been (nearly) completed, that of the nematode *Caenorhabditis elegans* (The C. elegans Sequencing Consortium 1998), was unsuccessful because several nematode aaRSs were found to miss large regions conserved in other species, apparently as a result of exon misassembly. Therefore, human aaRS sequences were included in the set for the analysis whenever available; otherwise, the sequences from *C. elegans*, *Arabidopsis thaliana*, or *Drosophila*

melanogaster were used as the second representative of the eukaryotes.

Sequence Alignment and Database Searches

Multiple alignments of the aaRS sequences were initially constructed using the progressive alignment program ALITRE (Seledtsov et al. 1995). The alignments were then manually adjusted on the basis of the results of iterative PSI-BLAST searches (Altschul et al. 1997; see below) and the boundaries of domains and secondary structure elements that were extracted from the aaRS structures present in the Protein Data Bank (PDB) (Bernstein et al. 1977). All the sequences of aaRS domains other than the class I and class II cores were cut out of the alignments and used as queries for iterative database search with the PSI-BLAST program (Altschul et al. 1997; Altschul and Koonin 1998). Briefly, this program constructs a position-dependent weight matrix (profile) from multiple alignments of BLAST hits that have an associated expectation value (*e*-value) above a certain cutoff and iterates the database search using this evolving profile as the new query. The statistical significance of the PSI-BLAST hits is assessed on the basis of the extreme value distribution statistics that originally had been developed by Karlin and Altschul for local alignments without gaps (Karlin and Altschul 1990; Karlin et al. 1991) and subsequently modified for gapped alignments (Altschul and Gish 1996; Altschul et al. 1997). There is no analytical proof of the applicability of the Karlin–Altschul statistics for searches using profiles as queries, but extensive computer simulations have shown a near-perfect fit of the score distribution obtained in such searches to the extreme value distribution. Accordingly, *e*-values reported by PSI-BLAST for each retrieved sequence in the iteration when its alignment with the query scores above the cutoff for the first time appear to be accurate estimates of the statistical significance; once a sequence is included in the profile, *e*-values reported for it (and its closely related homologs) at subsequent iterations become inflated and do not accurately represent the statistical significance (Altschul et al. 1997; Altschul and Koonin 1998). In this analysis, only *e*-values recorded for the first appearance of the given sequence above the cutoff were used to assess the statistical significance of database hits. Normally, the PSI-BLAST program was run to convergence, with the *e*-value of 0.01 used as the cutoff. The searches were normally run without filtering for regions of low compositional complexity, to avoid loss of information. However, in cases when apparent false positives caused by low complexity were noticed upon examination of the search results, these regions in the query sequence were masked using the SEG program with appropriately chosen parameters (Wootton 1994; Wootton and Federhen 1996).

Phylogenetic Trees

For the purpose of phylogenetic tree construction, large inserts and ambiguously aligned regions were removed from the aaRS alignments. Phylogenetic trees were constructed using the PHYLIP package programs (Felsenstein 1996). First, 1000 bootstrap replications were obtained from each alignment using the SEQBOOT program. Distance matrices were computed using the PROTDIST program with the "Dayhoff PAM distance" option. Each set of 1000 distance matrices was analyzed using the Fitch–Margoliash (Fitch and Margoliash 1967; FITCH program) and neighbor-joining (Saitou and Nei 1987; NEIGHBOR program) tree-building methods. Consensus trees and bootstrap support for each method were separately computed using the CONSENSE program. A global consensus topology was manually derived by collapsing internal branches that resulted in different branching orders in the trees produced by the Fitch–Margoliash and neighbor-joining methods. Nodes that were strongly supported by bootstrap analysis (>70%) under both of these methods were considered reliable. Branch lengths for the derived consensus topology were computed using the Fitch–Margoliash method (FITCH program).

The most likely root position in the consensus trees was determined using a least-squares modification of the midpoint rooting procedure (N.V. Grishin, unpubl.). Let us consider an unrooted tree with n leaves S_i. The point A on the tree for which

$$M(A) = \sum_{i=1}^{n} \left(d(A, S_i) - \frac{1}{n} \sum_{j=1}^{n} d(A, S_j) \right)^2 = \min,$$

in which $d(A, S_i)$ is the length of the path over the tree branches from point A to a leaf S_i, approximates the root position under the relaxed molecular clock assumption. Under the "precise molecular clock" and no errors in tree branch lengths, the path lengths from the root to each leaf are equal. Hence, $M(A)$ is zero if A is the root. When branch lengths are known with random errors resulting from finite lengths of sequences used for distance estimation, as they are in any real example, $M(A)$ is greater than zero if A is the root but is expected to be small. In this case, the point A which minimizes $M(A)$ is a least-squares root estimate. Alternatively, let "molecular clock" be "relaxed" to some extent. Then, the lengths of the paths from the leaves to the root are not equal. Each path length can be treated as a sum of a "true" path length, which is the same for all leaves, and a "random" component, namely, a random variable with a zero mean. If the variance of this random variable is sufficiently small, these deviations from the "clock" can be treated as random errors, and the total error of a branch length results from both the finite sequence lengths and the "relaxation" of the molecular clock. Thus, the point A that minimizes $M(A)$ is again a least-squares estimate of the root.

This method is related to the midpoint rooting, which places the root in the middle of the longest path between two leaves. The least-squares version is, however, more efficient because it uses information about all the sequences in the tree and not only about the most divergent pair as does the classical midpoint rooting. As a result, the least squares method is less sensitive to "outliers", that is, the longest terminal branches that are the result of molecular clock violation. The performance of the method improves with the increase of the number of leaves in the tree, which is equivalent to the increase of the statistical sample for estimation. However, it should be kept in mind, that if the leaves in a tree cannot be treated as a random sample in the sense of their distance from the root, or the variance of the random component is too large, the root will not be placed correctly. To our knowledge, no other method, except rooting by a paralog, works in these situations.

Clustering of proteins by sequence similarity was performed using the GROUPER program of the SEALS package (Walker and Koonin 1997) that uses a single-linkage clustering procedure; a series of cutoff values (in terms of alignment score) was used to select the value that partitioned the given protein set into two subsets. The results of the modified midpoint rooting (MMPR) and the clustering results were used to infer the root position, in conjunction with the analysis of likely synapomorphies—shared derived features of domain architecture (see Results and Discussion).

Other Procedures

Protein secondary structure prediction was carried out using the PHD program, with multiple sequence alignments used as the input (Rost and Sander 1994). Nonglobular protein domains were predicted using the SEG program with the set of parameters optimized for this task (window length, 45; trigger complexity, 3.4; extension complexity, 3.75) (Wootton 1994; Wootton and Federhen 1996). Coiled-coil domains were predicted using the COILS2 program (Lupas 1996). PDB files were viewed and manipulated using the InsightII program (Biosym). Sequence retrieval and large-scale analysis were handled with the programs of the SEALS package (Walker and Koonin 1997).

ACKNOWLEDGMENTS

We thank Joe Felsenstein for a helpful discussion of the modified midpoint rooting procedure.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

NOTE ADDED IN PROOF

While this manuscript was being processed for publication, the crystal structure of the threonyl-tRNA synthetase-tRNA (Thr) complex from *E. coli* was published [Sankaranarayanan, R., A.C. Dock-Bregeon, P. Romby, J. Caillet, M. Springer, B. Rees, C. Ehresmann, B. Ehresmann, and D. Moras. 1999. The structure of threonyl-tRNA synthetase-tRNA (Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell* **97:** 371–381]. This provides the prototype structures for the TGS domain and the HxxxH domain defined here. Sankaranarayanan and coworkers show that the HxxxH (the N2 domain in their notation) makes minor groove contacts with the tRNA acceptor stem and also notice the presence of a counterpart of this domain in AlaRS.

REFERENCES

Altschul, S.F. and W. Gish. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.

Altschul, S.F. and E.V. Koonin. 1998. PSI-BLAST—a tool for making discoveries in sequence databases. *Trends Biochem. Sci.* **23:** 444–447.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Aravind, L. and E.V. Koonin. 1998. The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **23:** 469–472.

———. 1999. Novel predicted RNA-binding domains associated with the translation machinery. *J. Mol. Evol.* **48:** 291–302.

Arnez, J.G. and D. Moras. 1997. Structural and functional considerations of the aminoacylation reaction. *Trends Biochem. Sci.* **22:** 211–216.

Artymiuk, P.J., D.W. Rice, A.R. Poirrette, and P. Willet. 1994. A tale of two synthetases. *Nat. Struct. Biol.* **1:** 758–760.

Bedouelle, H., V. Guez-Ivanier, and R. Nageotte. 1993. Discrimination between transfer-RNAs by tyrosyl-tRNA synthetase. *Biochimie* **75:** 1099–1108.

Bernstein, F.C., T.F. Koetzle, G.J. Williams, E.E. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112:** 535–542.

Bork, P., L. Holm, E.V. Koonin, and C. Sander. 1995. The cytidylyltransferase superfamily: Identification of the nucleotide-binding site and fold prediction. *Proteins* **22:** 259–266.

Brown, J.R. and W.F. Doolittle. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci.* **92:** 2441–2445.

———. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61:** 456–502.

Brown, J.R., F.T. Robb, R. Weiss, and W.F. Doolittle. 1997. Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. *J. Mol. Evol.* **45:** 9–16.

Brown, J.R., J. Zhang, and J.E. Hodgson. 1998. A bacterial antibiotic resistance gene with eukaryotic origins. *Curr. Biol.* **8:** R365–367.

Brunie, S., C. Zelwer, and J.L. Risler. 1990. Crystallographic study at 2.5 A resolution of the interaction of methionyl-tRNA synthetase from *Escherichia coli* with ATP. *J. Mol. Biol.* **216:** 411–424.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science* **282:** 2012–2018.

Cashel, M., D.R. Gentry, V.J. Hernandez, and D. Vinella. 1996. The stringent response. In *Escherichia coli and Salmonella. Cellular and molecular biology*, ed. F.C. Neidhardt, R.C. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, M. Schaechter, and H.E. Umbarger. pp. 1458–1496. ASM Press, Washington, DC.

Cavarelli, J., B. Delagoutte, G. Eriani, J. Gangloff, and D. Moras. 1998. L-arginine recognition by yeast arginyl-tRNA synthetase. *EMBO J.* **17:** 5438–5448.

Curnow, A.W., M. Ibba, and D. Soll. 1996. tRNA-dependent asparagine formation. *Nature* **382:** 589–590.

Curnow, A.W., K. Hong, R. Yuan, S. Kim, O. Martins, W. Winkler, T.M. Henkin, and D. Soll. 1997. Glu-tRNAGln amidotransferase: A novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl. Acad. Sci.* **94:** 11819–11826.

Cusack, S. 1995. Eleven down and nine to go. *Nat. Struct. Biol.* **2:** 824–831.

———. 1997. Aminoacyl-tRNA synthetases. *Curr. Opin. Struct. Biol.* **7:** 881–889.

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

Cusack, S., M. Hartlein, and R. Leberman. 1991. Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases. *Nucleic Acids Res.* **19:** 3489–3498.

Delarue, M. and D. Moras. 1993. The aminoacyl-tRNA synthetase family: Modules at work. *BioEssays* **15:** 675–687.

Delarue, M., A. Poterszman, S. Nikonov, M. Garber, D. Moras, and J.C. Thierry. 1994. Crystal structure of a prokaryotic aspartyl tRNA-synthetase. *EMBO J.* **13:** 3219–3229.

Doolittle, R.F. and J. Handy. 1998. Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.* **8:** 630–636.

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

Eriani, G., J. Cavarelli, F. Martin, L. Ador, B. Rees, J.C. Thierry, J. Gangloff, and D. Moras. 1995. The class II aminoacyl-tRNA synthetases and their active site: Evolutionary conservation of an ATP binding site. *J. Mol. Evol.* **40:** 499–508.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266:** 418–427.

Fitch, W.M. and E. Margoliash. 1967. Construction of phylogenetic trees. *Science* **155:** 279–284.

Freist, W., D.T. Logan, and D.H. Gauss. 1996. Glycyl-tRNA synthetase. *Biol. Chem. Hoppe Seyler* **377:** 343–356.

Freist, W., D.H. Gauss, M. Ibba, and D. Soll. 1997a. Glutaminyl-tRNA synthetase. *Biol. Chem.* **378:** 1103–1117.

Freist, W., D.H. Gauss, D. Soll, and J. Lapointe. 1997b. Glutamyl-tRNA synthetase. *Biol. Chem.* **378:** 1313–1329.

Goldgur, Y., L. Mosyak, L. Reshetnikova, V. Ankilova, O. Lavrik, S.

Khodyreva, and M. Safro. 1997. The crystal structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus* complexed with cognate tRNAPhe. *Structure* **5:** 59–68.
Hashimoto, T., L.B. Sanchez, T. Shirakura, M. Muller, and M. Hasegawa. 1998. Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc. Natl. Acad. Sci.* **95:** 6860–6865.
Hodgson, J.E., S.P. Curnock, K.G. Dyke, R. Morris, D.R. Sylvester, and M.S. Gross. 1994. Molecular characterization of the gene encoding high-level mupirocin resistance in *Staphylococcus aureus* J2870. *Antimicrob. Agents Chemother.* **38:** 1205–1208.
Ibba, M., J.L. Bono, P.A. Rosa, and D. Soll. 1997a. Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci.* **94:** 14383–14388.
Ibba, M., S. Morgan, A.W. Curnow, D.R. Pridmore, U.C. Vothknecht, W. Gardner, W. Lin, C.R. Woese, and D. Soll. 1997b. A euryarchaeal lysyl-tRNA synthetase: Resemblance to class I synthetases. *Science* **278:** 1119–1122.
Karlin, S. and S.F. Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.* **87:** 2264–2268.
Karlin, S., P. Bucher, V. Brendel, and S.F. Altschul. 1991. Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.* **20:** 175–203.
Koonin, E.V. and L. Aravind. 1998. Genomics: Re-evaluation of translation machinery evolution. *Curr. Biol.* **8:** R266–269.
Koonin, E.V., A.R. Mushegian, R.L. Tatusov, S.F. Altschul, S.H. Bryant, P. Bork, and A. Valencia. 1994. Eukaryotic translation elongation factor 1 gamma contains a glutathione transferase domain—study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Sci.* **3:** 2045–2054.
Lenhard, B., O. Orellana, M. Ibba, and I. Weygand-Durasevic. 1999. tRNA recognition and evolution of determinants in seryl-tRNA synthesis. *Nucleic Acids Res.* **27:** 721–729.
Lupas, A. 1996. Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266:** 513–525.
Makarova, K.S., L. Aravind, M.Y. Galperin, N.V. Grishin, R.L. Tatusov, Y.I. Wolf, and E.V. Koonin. 1999. Comparative genomics of the archaea (Euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* **9:** 608–628.
Markus, M.A., R.B. Gerstner, D.E. Draper, and D.A. Torchia. 1998. The solution structure of ribosomal protein S4 delta41 reveals two subdomains and a positively charged surface that may interact with RNA. *EMBO J.* **17:** 4559–4571.
Martin, W. and M. Müller. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* **392:** 37–41.
Moras, D. 1992. Structural and functional relationships between aminoacyl-tRNA synthetases. *Trends Biochem. Sci.* **17:** 159–164.
Mosyak, L., L. Reshetnikova, Y. Goldgur, M. Delarue, and M.G. Safro. 1995. Structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus*. *Nat. Struct. Biol.* **2:** 537–547.
Nagel, G.M. and R.F. Doolittle. 1995. Phylogenetic analysis of the aminoacyl-tRNA synthetases. *J. Mol. Evol.* **40:** 487–498.
Quevillon, S., F. Agou, J.C. Robinson, and M. Mirande. 1997. The p43 component of the mammalian multi-synthetase complex is likely to be the precursor of the endothelial monocyte-activating polypeptide II cytokine. *J. Biol. Chem.* **272:** 32573–32579.
RajBhandary, U.L. 1997. Once there were twenty. *Proc. Natl. Acad. Sci.* **94:** 11761–11763.
Rho, S.B., J.S. Lee, E.J. Jeong, K.S. Kim, Y.G. Kim, and S. Kim. 1998. A multifunctional repeated motif is present in human bifunctional tRNA synthetase. *J. Biol. Chem.* **273:** 11267–11273.
Ripmaster, T.L., K. Shiba, and P. Schimmel. 1995. Wide cross-species aminoacyl-tRNA synthetase replacement in vivo: Yeast cytoplasmic alanine enzyme replaced by human polymyositis serum antigen. *Proc. Natl. Acad. Sci.* **92:** 4932–4936.
Rivera, M.C. and J.A. Lake. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257:** 74–76.
Rost, B. and C. Sander. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19:** 55–72.
Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.
Sassanfar, M., J.E. Kranz, P. Gallant, P. Schimmel, and K. Shiba. 1996. A eubacterial Mycobacterium tuberculosis tRNA synthetase is eukaryote-like and resistant to a eubacterial-specific antisynthetase drug. *Biochemistry* **35:** 9995–10003.
Schade, M., C.J. Turner, K. Lowenhaupt, A. Rich, and A. Herbert. 1999. Structure-function analysis of the Z-DNA-binding domain Zalpha of dsRNA adenosine deaminase type I reveals similarity to the (alpha + beta) family of helix-turn-helix proteins. *EMBO J.* **18:** 470–479.
Seledtsov, I.A., I. Vul'f Iu, and K.S. Makarova. 1995. Multiple alignment of biopolymer sequences, based on the search for statistically significant common segments. *Mol. Biol. (Moscow)* **29:** 1023–1039.
Shiba, K., H. Motegi, and P. Schimmel. 1997a. Maintaining genetic code through adaptations of tRNA synthetases to taxonomic domains. *Trends Biochem. Sci.* **22:** 453–457.
Shiba, K., T. Stello, H. Motegi, T. Noda, K. Musier-Forsyth, and P. Schimmel. 1997b. Human lysyl-tRNA synthetase accepts nucleotide 73 variants and rescues Escherichia coli double-defective mutant. *J. Biol. Chem.* **272:** 22809–22816.
Shiba, K., H. Motegi, M. Yoshida, and T. Noda. 1998. Human asparaginyl-tRNA synthetase: Molecular cloning and the inference of the evolutionary history of Asx-tRNA synthetase family. *Nucleic Acids Res.* **26:** 5045–5051.
Siatecka, M., M. Rozek, J. Barciszewski, and M. Mirande. 1998. Modular evolution of the Glx-tRNA synthetase family—rooting of the evolutionary tree between the bacteria and archaea/eukarya branches. *Eur. J. Biochem.* **256:** 80–87.
Simos, G., A. Segref, F. Fasiolo, K. Hellmuth, A. Shevchenko, M. Mann, and E.C. Hurt. 1996. The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl- and glutamyl-tRNA synthetases. *EMBO J.* **15:** 5437–5448.
Simos, G., A. Sauer, F. Fasiolo, and E.C. Hurt. 1998. A conserved domain within Arc1p delivers tRNA to aminoacyl-tRNA synthetases. *Mol. Cell.* **1:** 235–242.
Soma, A. and H. Himeno. 1997. Recognition system of class II tRNA in Escherichia coli and yeast. *Nucleic Acids Symp. Ser.* **37:** 295–296.
———. 1998. Cross-species aminoacylation of tRNA with a long variable arm between *Escherichia coli* and *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **26:** 4374–4381.
Tas, M.P. and J.C. Murray. 1996. Endothelial-monocyte-activating polypeptide II. *Int. J. Biochem. Cell. Biol.* **28:** 837–841.
Toth, M.J. and P. Schimmel. 1990. Deletions in the large (beta) subunit of a hetero-oligomeric aminoacyl- tRNA synthetase. *J. Biol. Chem.* **265:** 1000–1004.
Ullrich, M. and C.L. Bender. 1994. The biosynthetic gene cluster for coronamic acid, an ethylcyclopropyl amino acid, contains genes homologous to amino acid-activating enzymes and thioesterases. *J. Bacteriol.* **176:** 7574–7586.
Walker, D.R. and E.V. Koonin. 1997. SEALS: A system for easy analysis of lots of sequences. *Intell. Syst. Mol. Biol.* **5:** 333–339.
Weiner, A.M. and N. Maizels. 1999. A deadly double life. *Science* **284:** 63–64.
Wilcox, M. and M. Nirenberg. 1968. Transfer RNA as a cofactor coupling amino acid synthesis with that of protein. *Proc. Natl. Acad. Sci.* **61:** 229–236.
Wootton, J.C. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.* **18:** 269–285.
Wootton, J.C. and S. Federhen. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266:** 554–571.

Received February 24, 1999; accepted in revised form May 27, 1999.