MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, Dec. 2003, p. 550–573 Vol. 67, No. 4
1092-2172/03/$08.00+0 DOI: 10.1128/MMBR.67.4.550–573.2003

On the Evolution of Structure in Aminoacyl-tRNA Synthetases

Patrick O'Donoghue and Zaida Luthey-Schulten*

Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

INTRODUCTION

Biological Background

According to the RNA world hypothesis, the modern biological world evolved from a form of life that was mostly RNA-based (for a review see reference 25). From this point of view, it is no surprise, therefore, that much of the translation machinery is composed of RNA molecules. As protein synthesis evolved and the resulting proteins themselves became more complex, they invaded functional niches, previously occupied by ribozymes, to enhance enzymatic activities.

The translation machinery is dedicated to interpreting the nucleic acid code in a two-part process. First, amino acids are covalently linked to their cognate tRNAs via an aminoacylation reaction catalyzed by a diverse group of proteins, the aminoacyl-tRNA synthetases (AARSs). At the ribosome the tRNA anticodon is matched to the mRNA codon, and the charged tRNA delivers the next residue of a nascent protein chain. Directed in vitro evolution experiments have shown that ribozymes can be constructed that aminoacylate tRNAs (37). It is possible that among the first proteins to take over ribozyme functions were the aminoacyl-tRNA synthetases. These ancient proteins are found in all extant organisms, and their inception likely predates the root of the universal phylogenetic tree (57, 62). The evolution of these proteins is of particular interest for understanding the evolution of translation and the transition from the RNA world to the modern form of life dominated by protein-enzymes and DNA genomes.

There are AARSs specific for each of the 20 standard amino acids. These enzymes are divided into two classes, class I and class II, which are unrelated in both sequence and structure (21, 29). The class I AARSs specify 11 amino acids, including Met, Val, Ile, Leu, Cys, Glu, Gln, Lys, Arg, Trp, and Tyr. The class II synthetases specify 10 amino acids, Ala, His, Pro, Thr, Ser, Gly, Phe, Asp, Asn, and Lys. The so-called class rule, stating that each amino acid is specified by a class I or class II synthetase but not both, was broken with the finding of a class I version of lysyl-tRNA synthetase (LysRS) in the archaeon *Methanococcus maripaludis* (31). The class I LysRS is found in most Archaea and in some Bacteria, while the class II version is found in all known eukaryotic genomes, the majority of Bacteria, and a small number of Archaea (3, 57). Both class I and class II LysRSs are found to coexist in two organisms of the archaeal genus *Methanosarcina*, *M. barkeri* (52) and *M. acetivorans* (24). The interesting evolutionary implications of the distribution of LysRSs are discussed below.

The AARSs are, in all known cases, multi-domain proteins. Within each class there is only one domain, referred to as the catalytic domain, that is conserved across all members of that class. Other domains are involved in anticodon binding, stabilization of the AARS-tRNA complex, and deacylating mischarged tRNAs. For a review and a clear presentation of the domain architectures of the AARSs see reference 62. We will focus on the structure of the catalytic domain as shown in Fig. 1.

The class I AARSs exhibit the basic Rossmann fold, which is

* Corresponding author. Mailing address: School of Chemical Sciences, University of Illinois, A544 CLSL, mc-712, 600 S Mathews Ave., Urbana, IL 61801. Phone: (217) 333-3518. Fax: (217) 244-3186. E-mail: schulten@scs.uiuc.edu.

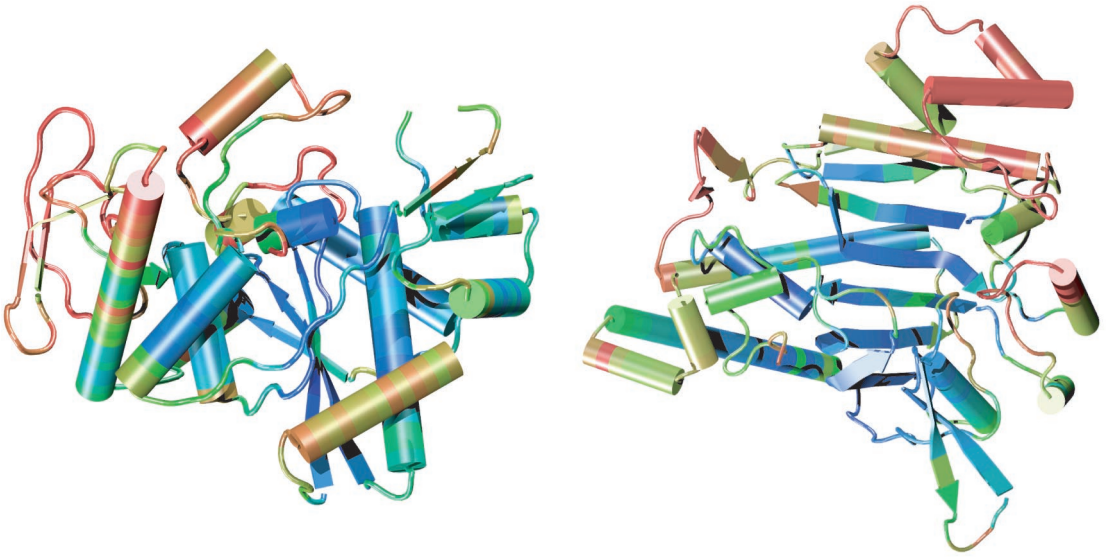

FIG. 1. Structural folds of the class I and class II lysyl-tRNA synthetases color coded by structural conservation across the non-redundant set. Regions of high structural conservation are shaded blue and low conservation, red. The class I synthetases have a Rossmann fold, and the class II synthetases have a novel α + β fold seen in few other proteins. (All structures drawn with VMD [28a].)

a three-layer α/β/α topology with an inner core of approximately five parallel beta sheets. The HIGH, for His-Ile-Gly-His, and the KMSKS, for Lys-Met-Ser-Lys-Ser, consensus motifs define two regions of sequence conservation for all class I AARSs (20, 27). The class II synthetases exhibit a unique fold found only in the class II synthetases themselves and in biotin synthetase holoenzyme (40). This fold is of the mixed α + β fold class and is typified by a central core of antiparallel β-strands flanked by α-helices. There are three short conserved sequence motifs in the class II synthetases (21). Most of the class II synthetases form homodimers (32), and much of motif 1 is involved in these dimer contacts. Motifs 2 and 3 form components of the active site.

Although the AARSs of different classes are not related by divergent evolution, they are clearly the result of a functional evolutionary convergence, as they carry out the same basic biochemical function. In many organisms all of the amino acids are attached to their cognate tRNAs through a direct mechanism (see Fig. 2) (32). As an example, we show the acylation reaction for glutamate. Although we only depict the overall reaction, this reaction is catalyzed in two steps. Initially the amino acid, Glu, and a molecule of ATP are bound to the synthetase, GluRS, active site, where they are covalently linked by ester bond formation between the α-carboxylate of the amino acid and the α-phosphate of the ATP. The products of this reaction are pyrophosphate and an enzyme-bound aminoacyl-adenylate. In the second step the amino acid is transfered to the terminal 2′- (typical for class II) or 3′-hydroxyl (typical for class I) of the acceptor stem of the cognate tRNA, $tRNA^{Glu}$, to form the aminoacylated tRNA $Glu\text{-}tRNA^{Glu}$.

The first alternative pathway for aminoacylation of a tRNA was described over 30 years ago (56), but the discovery that the

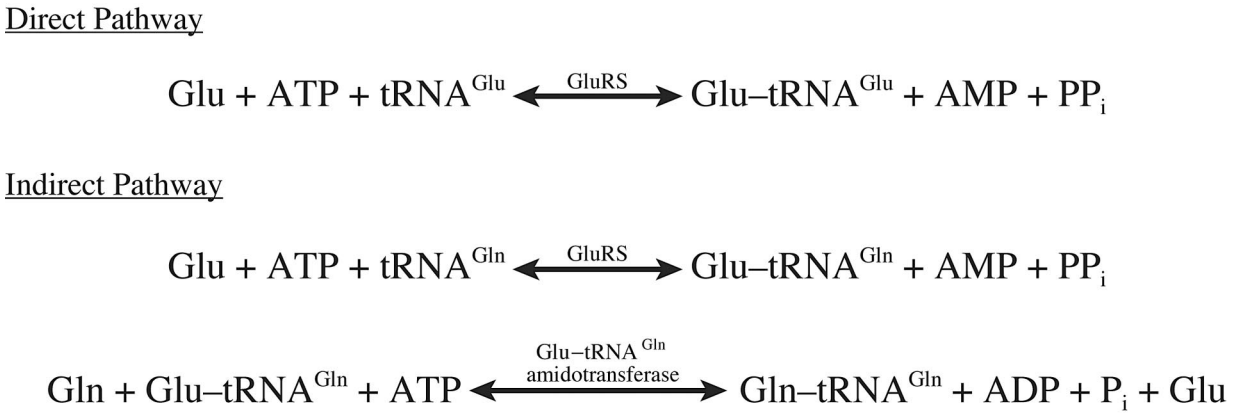

FIG. 2. Mechanisms for the direct and indirect formation of aminoacyl-tRNAs.

genome of *Methanococcus jannaschii* encoded only 16 of the 20 expected AARSs (12) dismantled the notion that all organisms require a full complement of AARSs. One of the four apparently missing enzymes was the class I LysRS, discussed above. Two of the remaining missing enzymes, asparagine and glutamine, were accounted for by the existence of an indirect mechanism for charging the tRNA (see Fig. 2). In the indirect pathway, $tRNA^{Gln}$ is first misacylated with Glu by a nondiscriminating GluRS (29). Glu-$tRNA^{Gln}$ amidotransferase converts the misacylated Glu to Gln, and the result is a correctly charged tRNA. GlnRS is absent from all known Archaea and most Bacteria, and these organisms exclusively utilize the indirect pathway.

In some Bacteria and most of the Archaea, AsnRS is not present, and $tRNA^{Asn}$ is aminoacylated via an analogous indirect pathway (4, 54). The genomes of both *Deinococcus radiodurans* and *Thermus thermophilus* encode a nondiscriminating AspRS and the standard AsnRS, so these organisms are able to correctly aminoacylate $tRNA^{Asn}$ through the direct and indirect pathways (5, 15, 39). In these organisms, the indirect pathway is the only metabolic route for asparagine biosynthesis (39). Similar indirect pathways also exist for incorporation of formylmethionine (45) and selenocysteine, the 21st amino acid (38). Pyrrolysine (Pyl), the 22nd amino acid, is proposed to be charged to its cognate tRNA via a similar indirect pathway. A class II LysRS that is not specifically related to other characterized class II LysRSs is responsible for charging the cognate tRNA for pyrrolysine with lysine. Putatively, other unknown enzymes are responsible for the conversion of the charged lysine to pyrrolysine to form Pyl-$tRNA^{Pyl}$ (52).

The last standard AARS missing from the genome of *M. jannaschii* is CysRS, yet *M. jannaschii* is able to synthesize Cys-$tRNA^{Cys}$ (53). CysRS is also absent from *Methanobacterium thermoautotrophicus* and *Methanopyrus kandleri* (34). Although the complete mechanism of Cys-$tRNA^{Cys}$ formation is still unknown, Söll and colleagues have shown that in *M. jannaschii* ProRS is able to bind cysteine as a substrate and acylate $tRNA^{Pro}$ to form Cys-$tRNA^{Pro}$ (4, 53). They have further shown that Cys-$tRNA^{Pro}$ can be formed with varying efficiency in vitro by ProRS enzymes from all three domains of life, including organisms that posses the standard CysRS (1). Structural analysis of ProRS cocrystallized with Cys-adenylate and Pro-adenylate analogs has shown that in *M. jannaschii* and *M. thermoautotrophicus* the ProRS active site does not have the ability to discriminate between proline and cysteine (35). Taken together, these structural and biochemical studies provide convincing proof that ProRS is capable of and indeed does catalyze the formation of Cys-$tRNA^{Pro}$. It remains unknown, however, how Cys-$tRNA^{Cys}$ is formed in organisms lacking CysRS or if Cys-$tRNA^{Pro}$ is involved in the formation of the essential Cys-$tRNA^{Cys}$. For a review, see reference 34.

Evolutionary Theory of the Universal Phylogenetic Tree

The universal phylogenetic tree, resulting from the molecular phylogeny of the rRNA sequences (22, 58), presents a framework for understanding the evolutionary history of each genetic element in the biosphere. In order to examine the complex evolutionary path of the aminoacyl-tRNA syntheta-

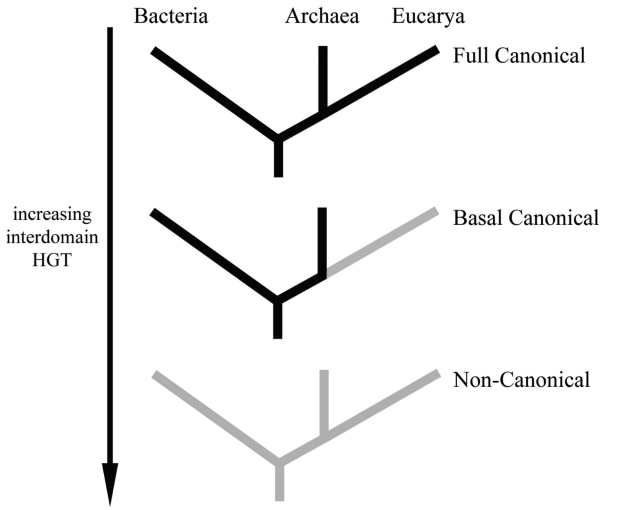

FIG. 3. Effect of horizontal gene transfer between organisms in different domains of life on the canonical pattern observed in molecular phylogenetic trees. Increasing interdomain HGT erases the historical trace which is evident in the universal phylogenetic tree (60).

ses, we must first present a summary of the evolutionary theory of Woese and colleagues (57, 60).

Vertical and horizontal gene flow are the molecular processes that shape the evolutionary course. Vertical gene transfer is the process of transmitting genes from parent to offspring and the molecular divergence resulting from mutation and gene duplication which ultimately leads to organismal divergence and speciation. This type of gene flow is responsible, in large part, for the phylogenetic distribution that is referred to as the universal phylogenetic tree or the canonical evolutionary pattern. At the most basic level, this pattern shows the ancient split between the Bacteria and the Archaea and also depicts the evolution of the Eucarya from the archaeal branch at a later time (Fig. 3, top). Molecules that strictly adhere to this pattern, referred to as the full canonical pattern, are the result of an evolution that was dominated, after the organismal domains began to diverge, by vertical gene flow.

Horizontal gene transfer (HGT) is the second and seemingly more extraordinary type of gene flow. This is the transfer of genetic elements between different species. The effect can be local, e.g., HGT between two closely related bacteria, *Bacillus stearothermophilus* and *Staphylococcus aureus*, and in some cases undetectable by molecular sequence analysis. The effect of HGT, however, can be much more dramatic. Horizontal transfer from a bacterium to a eukaryote, say very early on in eukaryotic evolution, can partially erase the historical trace so evident in the universal tree. Namely, this type of event can produce the basal canonical pattern in which there is only an obvious distinction between the archaeal and bacterial versions of a given molecule (see Fig. 3, middle).

The operational definition of an archaeal versus a bacterial version of a molecule has been described (57). The authors clearly state that "for these two organismal domains, the interdomain differences between the archaeal and bacterial proteins must far outweigh any intra-domain differences: the two must appear to differ in genre." In keeping with this terminology, we refer to bacterial and archaeal versions of a given

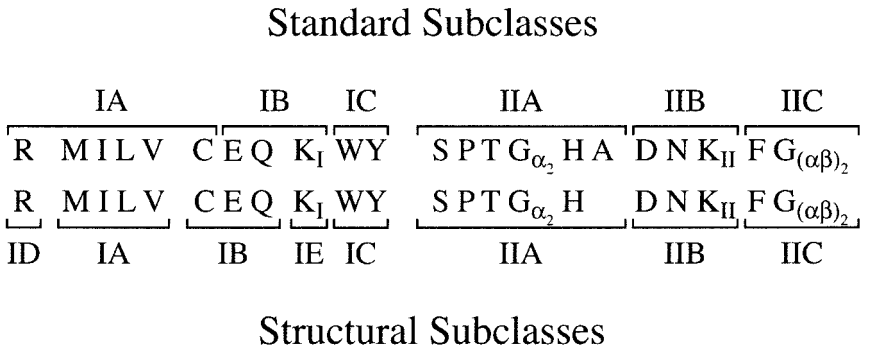

FIG. 4. Division of the aminoacyl-tRNA synthetases by class and subclass. The standard subclasses are based on data from references 17 and 48. Structural subclass definitions come from this study.

molecule as being of the bacterial genre or the archaeal genre. The effect of HGT can be yet more extreme. Widespread HGT at a sufficiently late time, after the three primary domains of life have emerged, can completely erase the historical trace. The hallmark of this type of horizontal transfer is a completely noncanonical pattern (Fig. 3, bottom).

Vertical and horizontal gene flows contribute to genetic diversity and novelty in different ways. Vertical flow is a slow process, the so-called "descent with modification," which over long evolutionary times leads to large magnitudes of divergence. Horizontal transfer, however, is a mechanism for rapid mixing of genetic material. This property of HGT is best understood through the following thought experiment. Consider two genes (or genetic elements) that are orthologs, i.e., homologous genes in different species. Gene Ab is the bacterial version, and gene Ae is the eukaryotic version. The Archaea and Bacteria diverged approximately 3.5 billion years ago, so with this time scale in mind, imagine that Ae and Ab have diverged from a common ancestor at time t_0 = 3.5 billion years ago. At a more recent time, say t_1 = 500 million years ago, Ab is horizontally transfered to a eukaryote that has gene Ae. After some short evolutionary interval, gene Ae is displaced by gene Ab in this eukaryote via vertical gene transfer. At time t_1, Ae and Ab are quite divergent. Both genes have experienced very different processes of vertical descent over a 3-billion-year period. The end result is that the eukaryote in possession of Ab has now obtained a gene, essentially instantaneously, that would have required the processes of vertical gene flow at least 3 billion years to create. In this way HGT is able to rapidly introduce genetic novelty.

Sequence-Based Evolutionary Analysis of the AARSs

Sequence-based evolutionary analysis of the AARSs has been the subject of intense study. The evolution of these proteins has been documented at approximately three levels of divergence: specificity, subclass, and class. The specificity level includes studies that focus on the evolution of individual AARSs specific for a particular amino acid. The synthetases of each class have been roughly grouped into the standard subclasses (see Fig. 4). The subclass level examines the series of gene duplications that gave rise to a subset of the AARSs of each class that are thought to be more closely related to each other than to other AARSs of the same class. The most divergent level involves the gene duplication events that gave rise to all of the synthetases in each class separately.

The amount and quality of the investigations into the AARSs decrease as the level of divergence examined increases. The most accurate work involves the specificity level, and to some extent slightly beyond this level (detailed below). Sequence comparison between AARSs of different specificities typically occupies a range of sequence identity values that fall below the twilight zone threshold, less than 25% sequence identity (8). Sequence comparison quickly becomes unreliable at this and lower levels of sequence identity. In this regime of similarity, it becomes difficult to distinguish between correctly aligned homologous sequences and unrelated sequences or random alignments. We will review the developments in the phylogenetic analysis of the AARSs, beginning with the specificity level and moving hierarchically toward the class level.

The most recent and insightful analysis of the synthetases at the specificity level is found in the paper of Woese et al. (57). In a contemporaneous paper, Wolf et al. (62) also examined this level of AARS evolution in detail. Although the evolutionary analysis of the catalytic domain presented (57) is more complete and extensive, the paper by Wolf et al. delivers an invaluable discussion of the accessory domains (anticodon binding and other domains) common to subsets of the synthetases within and in some cases across both classes. For the most part, the results of these papers are consistent with regard to the catalytic domain, so we will summarize the results of Woese and colleagues.

Phylogenetic analysis of each of the AARSs separately shows the distribution of each AARS across the three domains of life. The pattern of this distribution can be categorized with respect to the universal phylogenetic tree and classified as the full canonical, basal canonical, or noncanonical pattern. In Fig. 5 we summarize the previous results (57) and group the synthetases according to which pattern they exhibit. Class I synthetases that conform to the full canonical pattern are specific for W, Y, E, L, and I, while class II synthetases specific for H, D, F, and P also show the full canonical distribution. The basal canonical pattern is observed for the class I synthetases that specify M, V, and R. Class II synthetases specific for A and T follow the basal pattern. Woese et al. reported that the class I LysRS shows a noncanonical pattern, but a more recent study

	Class I	Class II
Full Canonical	W Y L I E	F H P D
Basal Canonical	R M V K_I	T A
Non-Canonical	C Q	S G_{α_2} K_{II} N $G_{(\alpha\beta)_2}$

FIG. 5. Phylogenetic pattern exhibited by the aminoacyl-tRNA synthetases according to reference 57.

that included more sequences concluded that this synthetase exhibits the basal pattern (3).

For each of the above-mentioned synthetases, those that fit the canonical or basal canonical pattern, all exhibit exceptions to these patterns as a result of various levels of horizontal gene transfer. LeuRS conforms completely to the canonical pattern without exception, while for ProRS a number of bacteria have received the archaeal type via HGT. Of the synthetases that exhibit the basal canonical pattern, HGT from Bacteria to the eukaryotes accounts for the majority of the deviations from the full canonical pattern. This is the case for ThrRS, ValRS, and AlaRS. Some bacterial MetRSs are distinctly of the archaeal genre, having received the gene via HGT from Archaea. Furthermore the eukaryotic MetRSs are of the archaeal genre, but they fail to form a distinct grouping that is required for the full canonical pattern to hold.

In some instances, sequence analysis has been able to go beyond constructing phylogenies for the specificities alone and show some of the gene duplication events that led to the formation of the AARS specificities. Due to the close homology between AspRS and AsnRS and a similarly close homology between GluRS and GlnRS, Woese et al. (57) were able to complete the evolutionary path that gave rise to AsnRS and GlnRS. GlnRS evolved from the eukaryotic GluRS through gene duplication and was later horizontally transfered to a limited number of Bacteria. AsnRS evolved somewhat earlier as the result of a gene duplication, prior to the advent of the Eucarya, from the ancestral archaeal type AspRS. In these cases, the differences between the archaeal genre and the bacterial genre for AspRS (GluRS) are greater than the differences between the archaeal type AspRS (GluRS) and AsnRS (GlnRS). The sequence-based analysis of Brown and Doolittle depicts the duplication events that gave rise to LeuRS, IleRS, and ValRS (10) and demonstrates that these synthetases group monophyletically with respect to amino acid specificity. Sequence comparison also shows that TrpRS and TyrRS group monophyletically (9).

Moving to a greater level of divergence, Schimmel and coworkers used structure to guide the alignment of subclass IIA AARSs (see Fig. 4) (46). With the structural alignment of ProRS, SerRS, ThrRS, GlyRS, HisRS, and AspRS as a template, they aligned the most conserved sequence fragments (motifs 1, 2, and 3; see above) of all of the available sequences for these synthetases. AspRS was used to root the class IIA phylogeny. Their results indicated that SerRS, ProRS, and ThrRS form a distinct supercluster but that the exact relationships between these three synthetases could not be reliably determined. GlyRS and HisRS were each reported to form separate groups. Although this phylogeny was guided by alignment of structurally equivalent regions across the subclass IIA, the resulting phylogeny was constructed by the maximum parsimony method, which is strictly sequence dependent.

In an earlier study, Nagel and Doolittle employed multiple alignments of highly conserved sequence fragments of the AARSs to construct a phylogeny, with a sequence-based similarity measure, at the class level (41). Although only a limited set of AARS sequences, including no archaeal examples, was available, this early work put forth suggestions that remain accurate in light of more rigorous study and larger data sets. Nagel et al. observed that the AARSs cluster monophyletically with respect to amino acid specificity, and the monophyly rule, with only two exceptions (see Findings and Analysis), remains intact. The phylogeny shows a number of superclusters that appear to be consistent or partially consistent with our data (see below) and other more recent investigations (9, 16, 46). These superclusters are indicated by parentheses: for class I (GluRS, GlnRS), (TrpRS, TyrRS), and (ValRS, IleRs, LeuRS, MetRS); for class II (ThrRS, ProRS, SerRS) and (LysRS, AspRS, AsnRS). Because the sequence identity between many of the AARSs of different specificities is below the twilight zone threshold, the connections within and among these superclusters are unreliable and in some cases misleading.

Motivation for Structural Analysis

Over a decade has passed since the work of Nagel and Doolittle, and now there are a sufficient number of experimentally determined AARS structures to construct a definitive structural phylogeny of the AARSs at the class level. Only the structure of AlaRS remains unknown. In addition, there are structures representing the archaeal and bacterial genres for many but not all of the synthetases. Chothia and coworkers showed that protein structure is more highly conserved than sequence (14). This guiding principle allows us to use structural alignments to construct an accurate phylogeny depicting the entire evolutionary course of the AARSs. In addition to a sufficiently large data set and accurate multiple structural alignments, a measure of structural homology is key to correctly reconstructing the evolutionary history of the AARSs at the class level. We introduce a robust structural measure for this purpose which is motivated by the "fraction of native contacts" or Q measure from the field of protein folding.

In the next section we give a detailed description of the data set of proteins used in this study, the structural alignment method, a new method for generating a nonredundant data set, our measure of structural homology, and the integration of data from the sequence-based analysis of the AARSs. Following this section, we present our results from the structural alignment of the AARSs, the structural phylogeny, and a discussion of structural conservation in the AARSs. We conclude with a discussion of the evolutionary implications of this structural phylogeny.

COMPUTATIONAL METHODS

AARS Domains and Coordinates

In order to study the evolutionary course of the AARSs at the class level, we confine our attention to the class I and class

II catalytic domains that are common to all members of each class separately. Domain definitions were taken from the latest version of the Structural Classification of Proteins, SCOP 1.61 (40). The Astral database mirrors the Protein Data Bank (PDB), but divides each PDB file into separate files for each SCOP domain (6). Except in a few cases where SCOP domains have yet to be defined, the Astral PDB-style files were used as a source of coordinates for the catalytic domain of the AARSs. If the SCOP domain definition was not available for a particular AARS, we used a structural alignment to the closet available homolog to define a SCOP-like domain from the PDB chain. SCOP-like domains were defined for LeuRS (1h3n), ProRS from *M. jannaschii* (1nj8) and *M. thermoautotrophicus* (1nj1), and AsnRS from *T. thermophilus* (7). At the time of this writing there is no PDB entry for this protein (or for any AsnRS), but the coordinates were provided by Stephen Cusack (personal communication). We identify this protein structure with a fictitious PDB code, 11sc.

Structural Alignment

Using the multiple structural alignment program STAMP (49), all of the available catalytic domains were aligned for the class I and class II AARSs separately. Including nearly identical crystal forms, e.g., different chains of a homodimer or the same protein in different crystallographic environments, there are a total of 56 class I AARS catalytic domains and 99 class II AARS catalytic domains in the PDB (not including AsnRS, see above). STAMP uses a dynamic programming procedure in combination with linear least-squares fitting to find the rigid-body rotation that simultaneously minimizes the C_α-C_α distance and local main-chain conformation for each pair of aligned proteins. The STAMP algorithm does not include sequence-dependent information. This program first computes all possible pairwise alignments and then uses a hierarchical clustering analysis based on structural similarity to build the multiple alignment. The program aligns the most similar structures first and moves along a structural dendrogram to add groups of aligned structures to the multiple alignment. This kind of multiple alignment was used in the following analysis.

Homology Measure

We employ a structural homology measure which is based on the structural similarity measure, Q, developed by Wolynes, Luthey-Schulten, and coworkers (18) in the field of protein folding. Our adaptation of Q is referred to as Q_H, and the measure is designed to include the effects of the gaps on the aligned portion: $Q_H = \aleph\,(q_{\mathrm{aln}} + q_{\mathrm{gap}})$, where $\aleph$ is the normalization, specifically given below. Q_H is composed of two components. q_{aln} is identical in form to the unnormalized Q measure of Eastwood et al. and accounts for the structurally aligned regions. The q_{gap} term accounts for the structural deviations induced by insertions in each protein in an aligned pair:

$$q_{\mathrm{aln}} = \sum_{i<j-2} \exp\left[-\frac{(r_{ij} - r_{i'j'})^2}{2\sigma_{ij}^2}\right]$$

$$q_{\mathrm{gap}} = \left[\sum_{g_a}\sum_{j}^{N_{\mathrm{aln}}} \max\left\{\exp\left[-\frac{(r_{g_aj} - r_{g'_aj'})^2}{2\sigma_{g_aj}^2}\right], \exp\left[-\frac{(r_{g_aj} - r_{g''_aj'})^2}{2\sigma_{g_aj}^2}\right]\right\} + \sum_{g_b}\sum_{j}^{N_{\mathrm{aln}}} \max\left\{\exp\left[-\frac{(r_{g_bj} - r_{g'_bj'})^2}{2\sigma_{g_bj}^2}\right], \exp\left[-\frac{(r_{g_bj} - r_{g''_bj'})^2}{2\sigma_{g_bj}^2}\right]\right\}\right]$$

The first term, q_{aln}, computes the unnormalized fraction of C_α-C_α pair distances that are the same or similar between two aligned structures. r_{ij} is the spatial C_α-C_α distance between residues i and j in protein a, and $r_{i'j'}$ is the C_α-C_α distance between residues i' and j' in protein b. This term is restricted to aligned positions, e.g., where i is aligned to i' and j is aligned to j'. The remaining terms account for the residues in gaps. g_a and g_b are the residues in insertions in both proteins, respectively. g'_a and g''_a are the aligned residues on either side of the insertion in protein a. The definition is analogous for g'_b and g''_b.

The normalization and the σ_{ij}^2 term are computed as:

$$\aleph = \frac{1}{\frac{1}{2}(N_{\mathrm{aln}} - 1)(N_{\mathrm{aln}} - 2) + N_{\mathrm{aln}}N_{gr} - n_{\mathrm{gaps}} - 2n_{\mathrm{cgaps}}}$$

$$\sigma_{ij}^2 = |i - j|^{0.15} \qquad (1)$$

where N_{aln} is the number of aligned residues. N_{gr} is the number of residues appearing in gaps, and n_{gaps} is the number of insertions in protein a, the number of insertions in protein b, and the number of simultaneous insertions (referred to as c gaps). n_{cgaps} is the number of c-gaps. Gap-to-gap contacts and intragap contacts do not enter into the computation. σ_{ij}^2 is a slowly growing function of sequence separation of residues i and j, and this serves to stretch the spatial tolerance of similar contacts at large sequence separations. Q ranges from 0 to 1, where Q = 1 refers to identical proteins. If there are no gaps in the alignment, then Q_H becomes Q_{aln}, which is identical to the Q measure described before (18) (Fig. 6).

Nonredundancy

Given that the class II synthetase catalytic domain is represented by 99 crystal forms in the PDB, it is necessary to systematically remove some of these examples in order to construct structural dendrograms and structural overlaps that can be displayed clearly on a single page and provide a picture that is representative of the AARSs of known structure. In addition, when Q per residue is displayed on a protein structure (see Fig. 7), it is necessary to determine a nonredundant set of proteins that contribute to the computation. Using too many examples of a particular synthetase will bias the results of this calculation. We have developed a method, based on the multidimensional *QR* factorization (28, 44), for generating a nonredundant multiple alignment which is introduced here and described in detail elsewhere (P. O'Donoghue and Z. Luthey-Schulten, submitted for publication).

The multiple structural alignment is encoded by the matrix A. A is of dimension $l_{\mathrm{aln}} \times k_{\mathrm{proteins}} \times d$. Each column of A corresponds to a single protein structure, and the multiple alignment is defined by the rows of A. The total length of the multiple alignment is l_{aln}, and k_{proteins} is the number of pro-

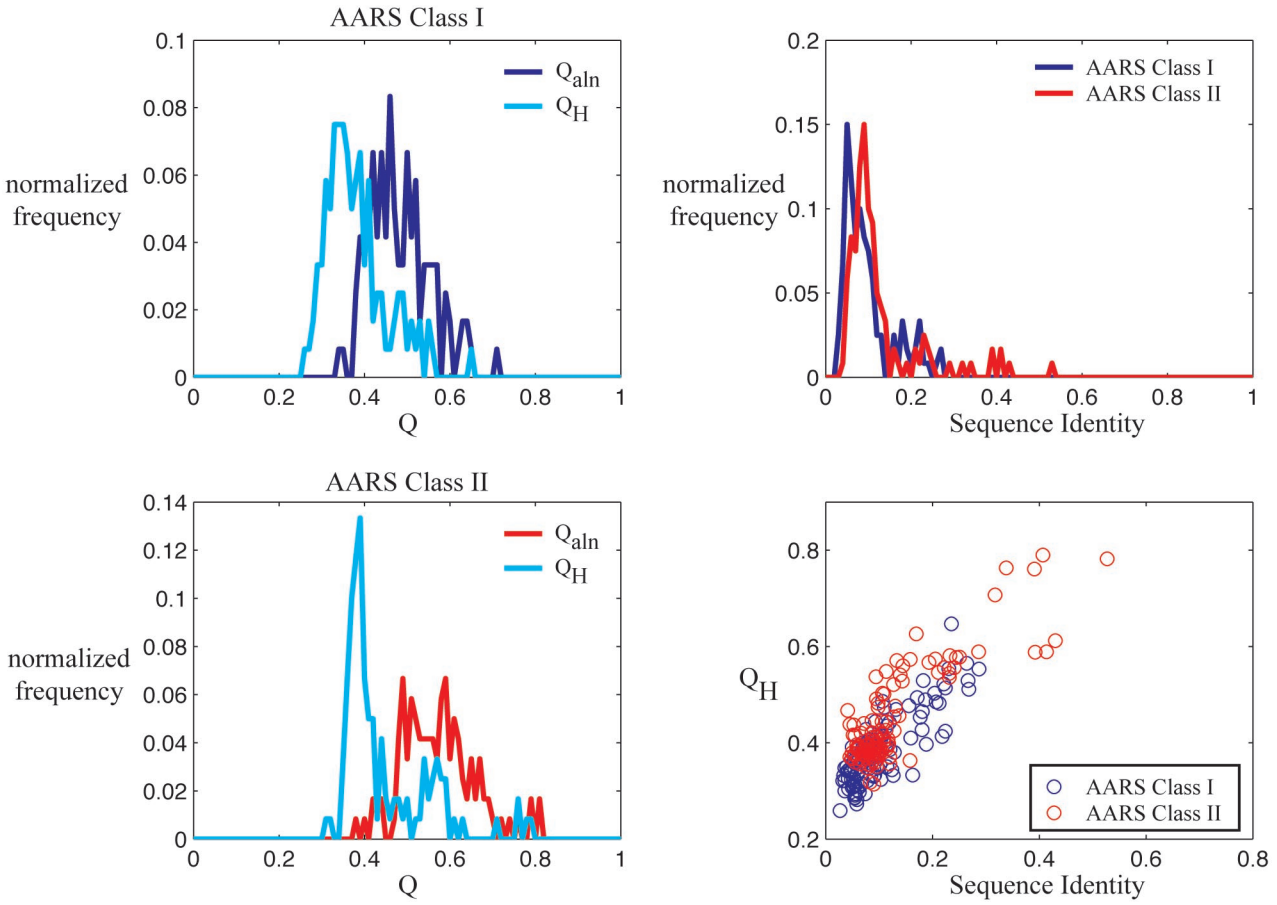

FIG. 6. Distribution of the structural and sequence identities of the nonredundant set, defined in Computational Methods, of AARSs appearing in the dendrograms in Fig. 11 and 12. The structural measure Q_H includes the effect of the gaps in the alignment, while Q_{aln} is only computed over the aligned portion (see Computational Methods). The class I AARSs are slightly more diverse in sequence and in structure than the class II AARSs.

teins in the alignment. The multiple structural superposition provides a set of rotated coordinates for each protein in the alignment. The rotated real-space coordinates of the C_α positions for the proteins are encoded in the first three components of the d dimension. Gapped positions are accounted for by the fourth component of the d dimension, so $d = 4$. The gap matrix, G, encodes gapped positions as 1 and aligned positions as 0. So that any gap position is weighted equivalently to any real-space position, we multiply G by a scalar, c, according to $c\|G\| = \|X\| + \|Y\| + \|Z\|$.

A is orthogonalized by the transformation matrix, Q^T (see Fig. 8). Orthogonalization of A gives the upper triangular matrix, R, and the columns of R are ordered by increasing linear dependence from left to right due to the action of the permutation matrix, P. The QR factorization steps through the matrix A in a columnwise fashion from left to right. At the ith step, in the factorization P is constructed to exchange the current column of A, over each d dimension simultaneously, with the kth column of maximum Frobenius norm, $\|a_k\|_F = [\sum_{d=1}^{4} \sum_{l=i+1}^{l_{aln}} |a_{lkd}|^2]^{1/2}$. A new alignment matrix is generated, $\tilde{A} = AP$, in which the proteins in $\tilde{A}$ are ordered by increasing linear dependence from left to right. Since we assume that redundancy in a multiple alignment is directly related to linear dependence between the aligned proteins, trimming proteins from right to left in $\tilde{A}$ to a desired level of redundancy gives a reduced set of proteins, which form the nonredundant set. We have also implemented this procedure for generating nonredundant multiple sequence alignments (O'Donoghue and Luthey-Schulten, submitted).

We have computed this ordering for both the class I and class II AARSs, and it can be seen in Fig. 9 and 10. We used the three-dimensional coordinates of the overlapped synthetase structures to compute the order. Note that by following the ordering, starting at the protein ranked 1, all of the major structural clusters are included before the ordering returns to one of the major clusters. We define the subset that contains all specificities and organisms, with two exceptions, as the nonredundant set. For the class I AARS this subset is composed of the first 16 ranked structures, and for the class II AARSs this subset is defined as the first 17 ranked structures. In two cases the backbone structure of the AARSs from closely related organisms is nearly indistinguishable, and these structures appear more closely related than different crystal forms of the same AARS. The TyrRSs from *B. stearothermophilus* and from

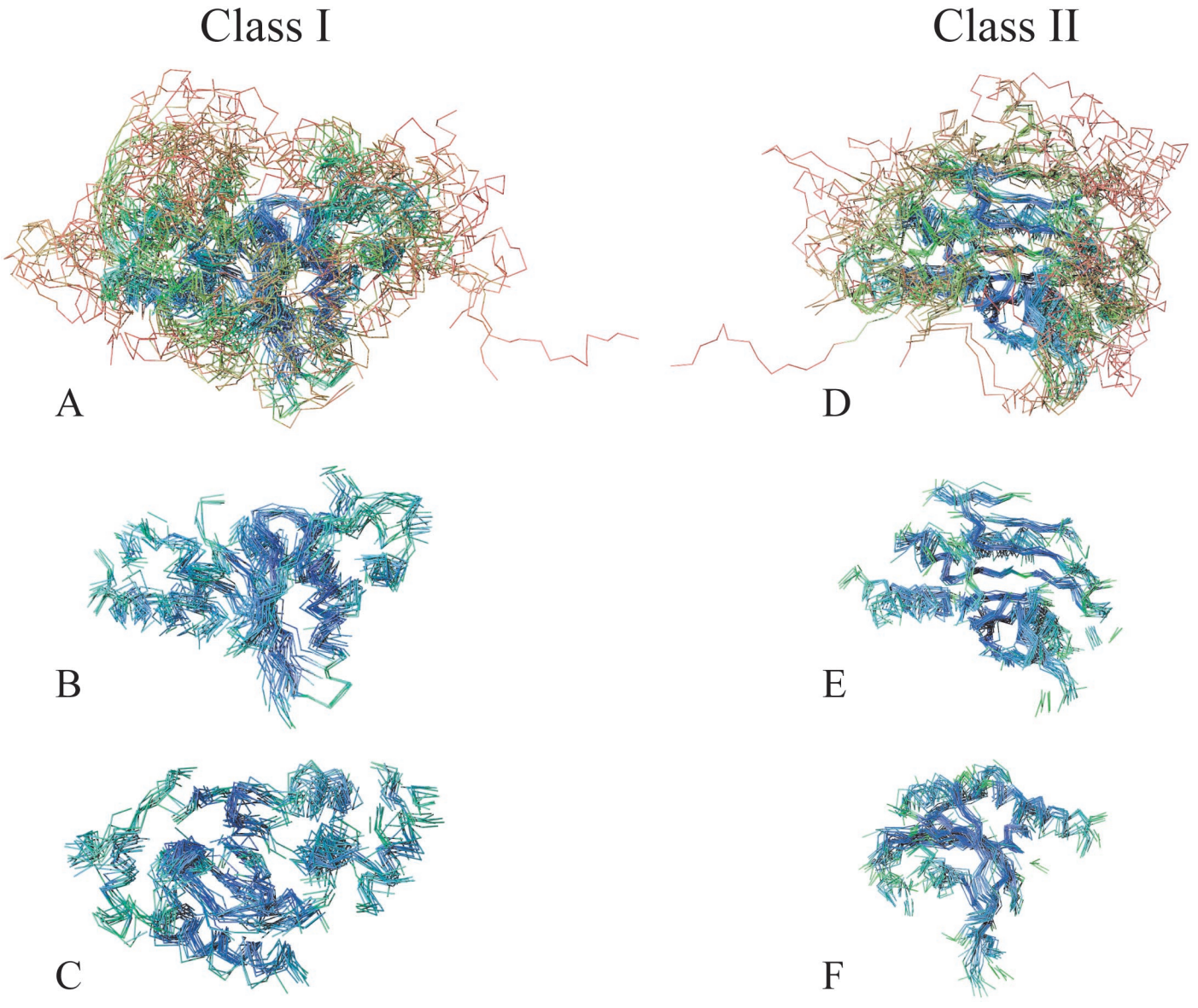

FIG. 7. Overlap of the nonredundant set of structures defined in Computational Methods and identified in the structural dendrograms for the class I (A to C) and class II (D to F) synethetases. Different views of the conserved core structures ($Q_{aln} > 0.4$) are shown in the lower panels. The proteins are colored by structural conservation (Q_{aln} per residue). The core structures may resemble the ancestral class I and class II proteins.

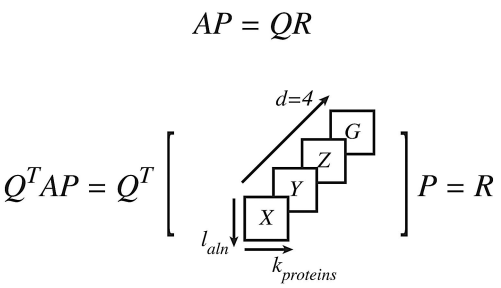

FIG. 8. Four-dimensional QR factorization.

S. aureus (ranked 18) are one example, and the other is ProRS from *M. jannaschii* and from *M. thermoautotrophicus* (ranked 29). For calculations of Q per residue, *S. aureus* TyrRS and *M. thermoautotrophicus* ProRS are excluded, but they are included in the representative structural dendrograms in Fig. 11 and 12. Since asparagine synthetase A (AsnA) is not a true AARS, it is also excluded from Q per residue computations for the class II AARSs (see Findings and Analysis).

Phylogenetic Analysis

In order to investigate the structural evolutionary relationships between the AARSs, we apply the unweighted pair group method with arithmetic averages (UPGMA) for cluster analysis (51). This method performs agglomerative clustering based on a pairwise distance measure which can be represented as a dendrogram. Since the sequence identity distribution (see Fig. 6) is tightly peaked near 10% identity, this measure is inappropriate for constructing dendrograms of such distantly related proteins. Instead, the structural homology measure Q_H is used as a pairwise similarity measure. The distance matrix required for the UPGMA method is simply a matrix of the pairwise structural dissimilarity values ($1 - Q_H$). Other popular methods for constructing phylogenies, such as maximum parsimony, which rely on generating ancestral states of modern sequences have been developed for sequence-based comparisons and are not currently applicable to constructing structural phylogenies. This presents an interesting, though, to our knowledge, undeveloped area of study. Neighbor-joining trees are available in the supplementary material (see Addendum).

Sequence-Based Annotation of Genre

The results of Woese et al. (57) were used to assign the genre, either archaeal or bacterial, to each of the protein structures. Each protein is a gene product from a particular organism. Due to the widespread occurrence of horizontal gene transfer over the course of AARS evolution, the organism and the gene product will not always belong to the same domain of life. Some AARSs for particular organisms do not appear in the sequence phylogeny of Woese et al., and in these cases we used BLAST (2) to determined the closest relative represented in the sequence phylogeny. For example, the IleRS from *T. thermophilus* is not found in the sequence phylogeny, but the closest relative appearing in the sequence phylogeny is the IleRS from *Clostridium acetobutylicum.* According to the BLAST alignment these two proteins share 49% sequence identity, so IleRS from *T. thermophilus* is clearly of the archaeal genre. The closest relative of the bacterial genre, found in the sequence phylogeny, is IleRS from *Thermotoga maritima*, which shares only 33% sequence identity with the *T. thermophilus* IleRS. This procedure was followed for genre assignments of ArgRS, ProRS, TyrRS, and LeuRS from *T. thermophilus*, TyrRS from *B. stearothermophilus* and *S. aureus*, TrpRS from *B. stearothermophilus*, and AspRS from *Pyrococcus kodakaraensis.* AARSs that do not conform to the canonical distribution cannot be assigned to the archaeal or bacterial genre.

FINDINGS AND ANALYSIS

Although the aminoacyl-tRNA synthetases have been the subject of intense scientific inquiry for nearly half a century, their complete evolutionary history has yet to emerge. Careful and accurate work (57, 62) has defined the evolutionary path of the aminoacyl-tRNA synthetases separately by specificity. Additionally, the pattern of gene duplications that gave rise to GlnRS and AsnRS from GluRS and AspRS, respectively, has been established (11, 57, 62). Except for other specific examples mentioned above, earlier evolutionary events remain undeveloped or unreliably determined. To provide a more complete and accurate picture of this evolutionary history, we present a phylogeny of the AARSs, at the class level, that is both quantitative and strictly based on protein structure.

Structural Alignment of the AARSs

The synthetases within each class are readily structurally aligned by the STAMP algorithm (see Computational Methods) and display clear structural homology. Synthetases of different classes, however, are not structurally related. In Fig. 7A to C all of the class I synthetases are overlapped, and in Fig. 7D to F all of the class II synthetases are overlapped. In each case the proteins are color coded by structural conservation. A striking feature is that the structural core is well conserved, with increasing variation toward the protein surface. The remaining panels of this figure show only the conserved structural core without the insertions. This conserved core structure likely resembles the ancestral state for each of the synthetase classes separately and allows us to glimpse a protein structure that may date back to the period of transition from the RNA world to the world of protein enzymes. There is greater structural divergence among class I synthetases than the more structurally homogeneous class II synthetases, and this hints at a somewhat more ancient origin for the class I AARSs.

In order to characterize the level of homology in the class I and II AARSs, we make reference to the similarity measure distributions shown in Fig. 6. In addition to the Q_H distribution, we also consider Q_{aln}, which accounts for only the structurally aligned regions. Note that the distribution of Q_{aln} falls off rapidly near $Q_{\mathrm{aln}} = 0.4$ and extends to $Q_{\mathrm{aln}} \sim 0.8$. These Q values can be compared to those observed in protein-folding trajectories where there is no q_{gap} component, since Q is computed based on a single protein structure in different conformations, i.e., while folding occurs. $Q \geq 0.4$ corresponds to structures with 5 Å root mean square deviation or less, and this

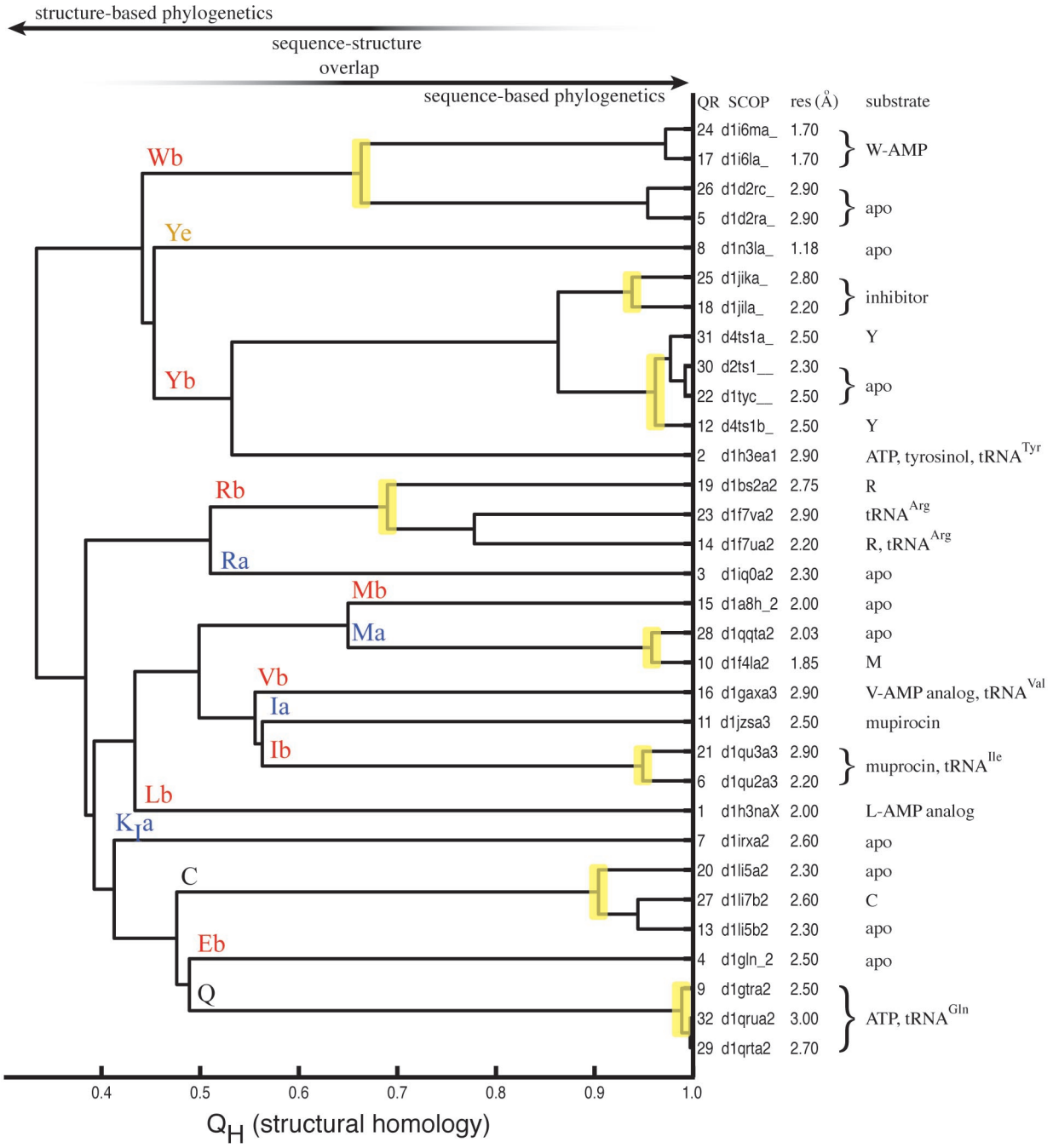

FIG. 9. Full structural dendrogram of the class I aminoacyl-tRNA synthetase catalytic domains. Crystal structures with high sequence identity, i.e., ~100%, have been omitted (see Computational Methods). The ordering according to the QR transformation and the SCOP domain codes is listed on the right. Also listed are the crystallographic resolution and the identity of cocrystallized substrates. Yellow bars indicate the noise in the crystallographic data arising either from variations in resolution or from conformation between the structures crystallized with and without substrates. In terms of Q_H, regions of reliability for sequence- and structure-based phylogenetic analysis are indicated on the graph, and there is a considerable region of overlap between the two methods. Faded lines indicate decreasing reliability of phylogenetic interpretations. (Trees were drawn with MATLAB version 6.5 [Mathworks, Inc.].)

value indicates visible structural similarity (18). Although this is qualitatively evident in the structural overlap of the synthetases (see Fig. 7), it is reassuring that the Q_{aln} distributions fit into our established notions about the relationship between the Q measure and structural similarity. The composite homology measure Q_H is shifted to lower Q values by approximately $\Delta Q = 0.12$, and this is consistent with the notion that accounting for the structural deviations introduced by gaps is equivalent to introducing a gap penalty. A structural alignment of unrelated proteins gives extremely low Q values, and the superposition of

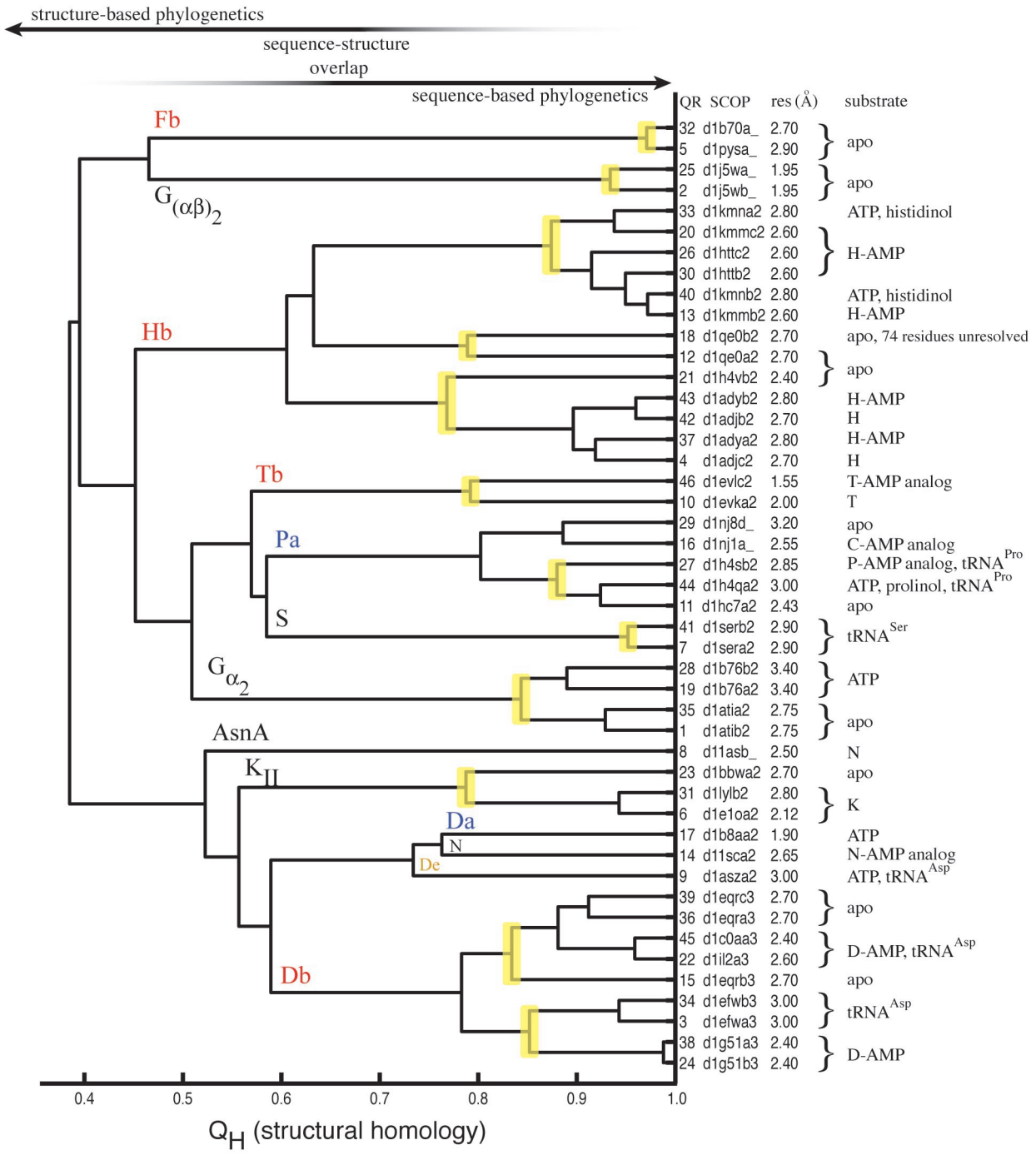

FIG. 10. Full structural dendrogram of the class II synthetases. See Fig. 9 for the details. The yellow bars indicate the noise in the structural data.

the unrelated class I and class II LysRSs is characterized by $Q_H = 0.09$ with $Q_{aln} = 0.1$.

The sequence identity distributions for the AARSs are also shown in Fig. 6. It is impressive to note the amount of structural conservation observed at the class level of the synthetases, and simultaneously to observe the striking lack of sequence conservation. The sequence identity distributions have an average of nearly 10% sequence identity, which approaches the limit seen in random alignments. The plot of Q_H versus sequence identity clearly indicates that sequence identity is an inappropriate similarity measure for such distantly related proteins. Note the broad range of Q values, $Q_H = \{0.25, 0.45\}$, which correspond to roughly 10% sequence identity. The Q measure continues to give meaningful information about the similarity of two distantly related proteins, while the sequence identity values are effectively indistinguishable from those associated with random alignments or unrelated proteins. These distributions reinforce the notion that protein structure is

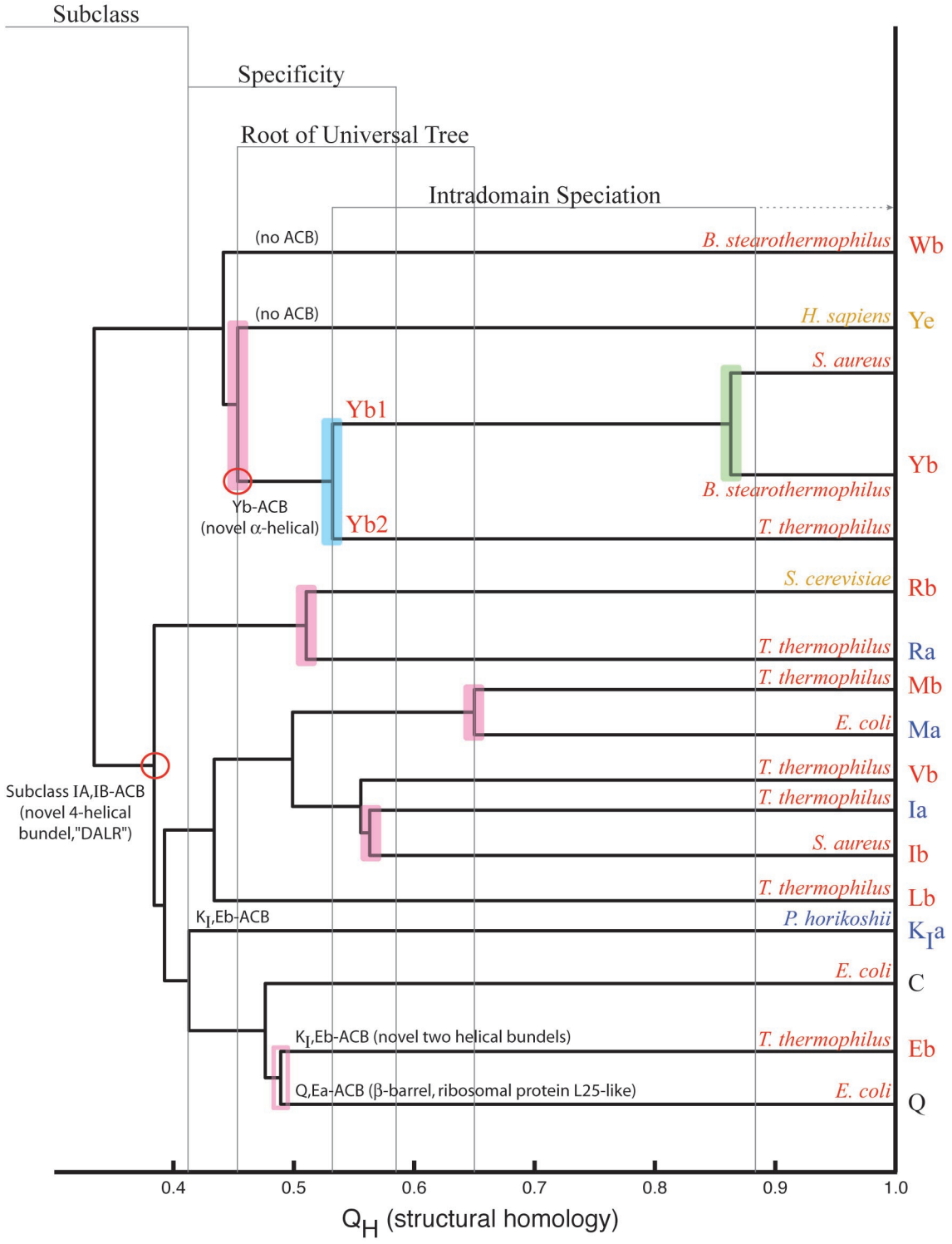

FIG. 11. Nonredundant structural dendrogram of the class I AARSs (see Computational Methods). The first 16 synthetases ordered in the full dendogram in Fig. 9 include all specificities and organisms except for TyrRS from *S. aureus*, which has also been included. In general, the features of the structural dendrogram agree with the sequence-based phylogenetic analysis of Woese et al. (57). The organism names are color coded according to the domain of life: red (Bacteria), blue (Archaea), and gold (Eucarya). The AARSs are labeled by their one-letter code for specificity and genre. Whenever an organism and enzyme differ in color, a horizontal transfer has occurred except in the cases of noncanonically distributed AARSs, which is in black. Magenta bars mark organismal domain divergence, blue bars mark subtype divergence, and green bars mark more local intradomain speciation. The open magenta bar indicates likely but unconfirmed organismal domain divergence (see text). A red circle denotes the proposed point of acquisition of an anticodon binding domain (ACB).

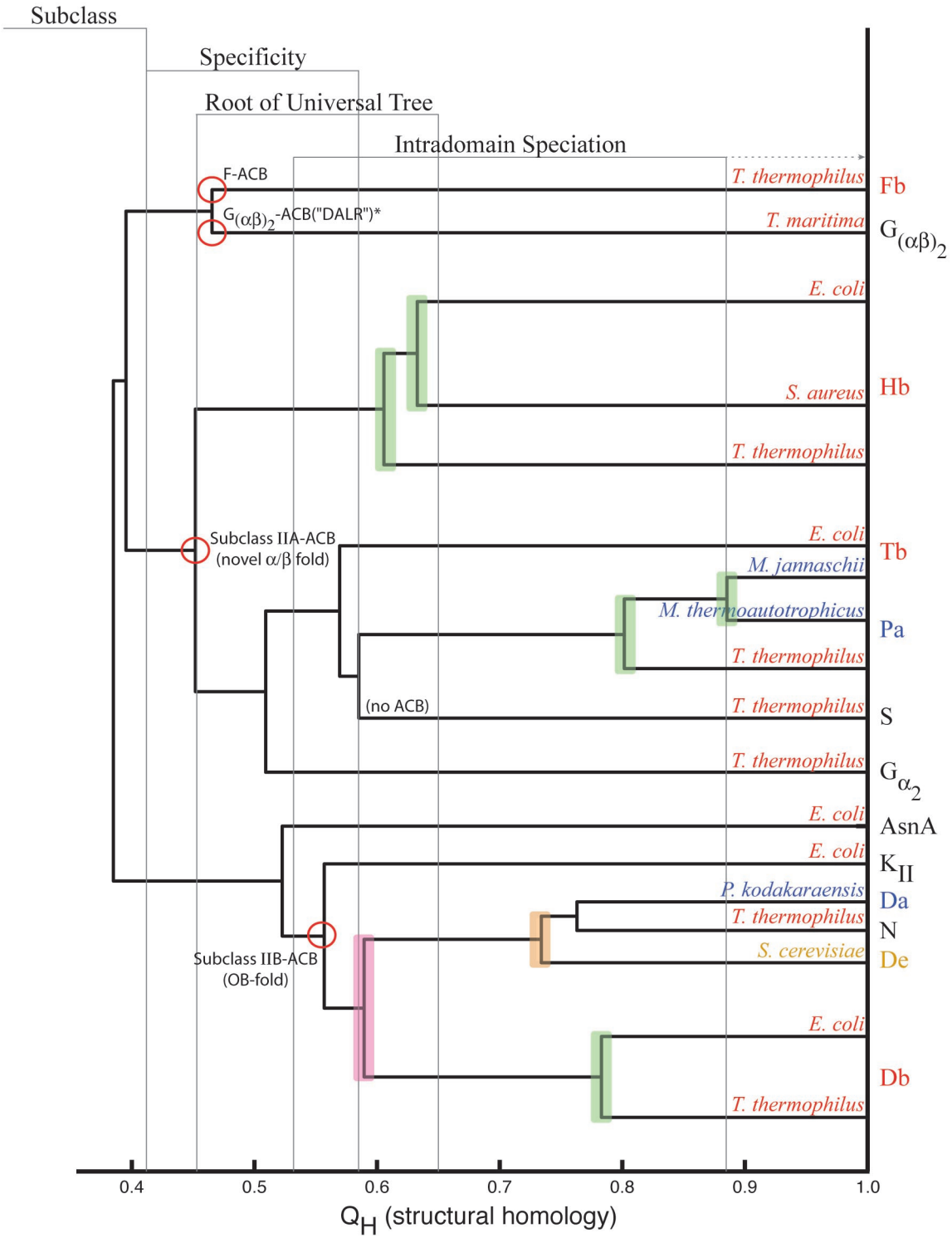

FIG. 12. Nonredundant structural dendrogram of the class II AARSs. The nonredundant set includes the first 17 ordered structures appearing in the full dendrogram in Fig. 10 except for *M. thermoautotrophicus* ProRS, which is also included. The annotations are the same as in Fig. 11 except that the orange bar marks divergence between Eucarya and Archaea. The asterisk indicates that no anticodon binding domain structure is available.

more highly conserved than sequence (14), and they are the basis for our motivation to use a structural measure to discern the early events of AARS evolution.

There is an interesting interplay between sequence and structure conservation. Figure 13 shows plots of Q per residue and sequence identity per residue averaged over the multiple alignment for the class I GlnRS and the class II AsnRS. In general, regions of sequence similarity are "encapsulated" by

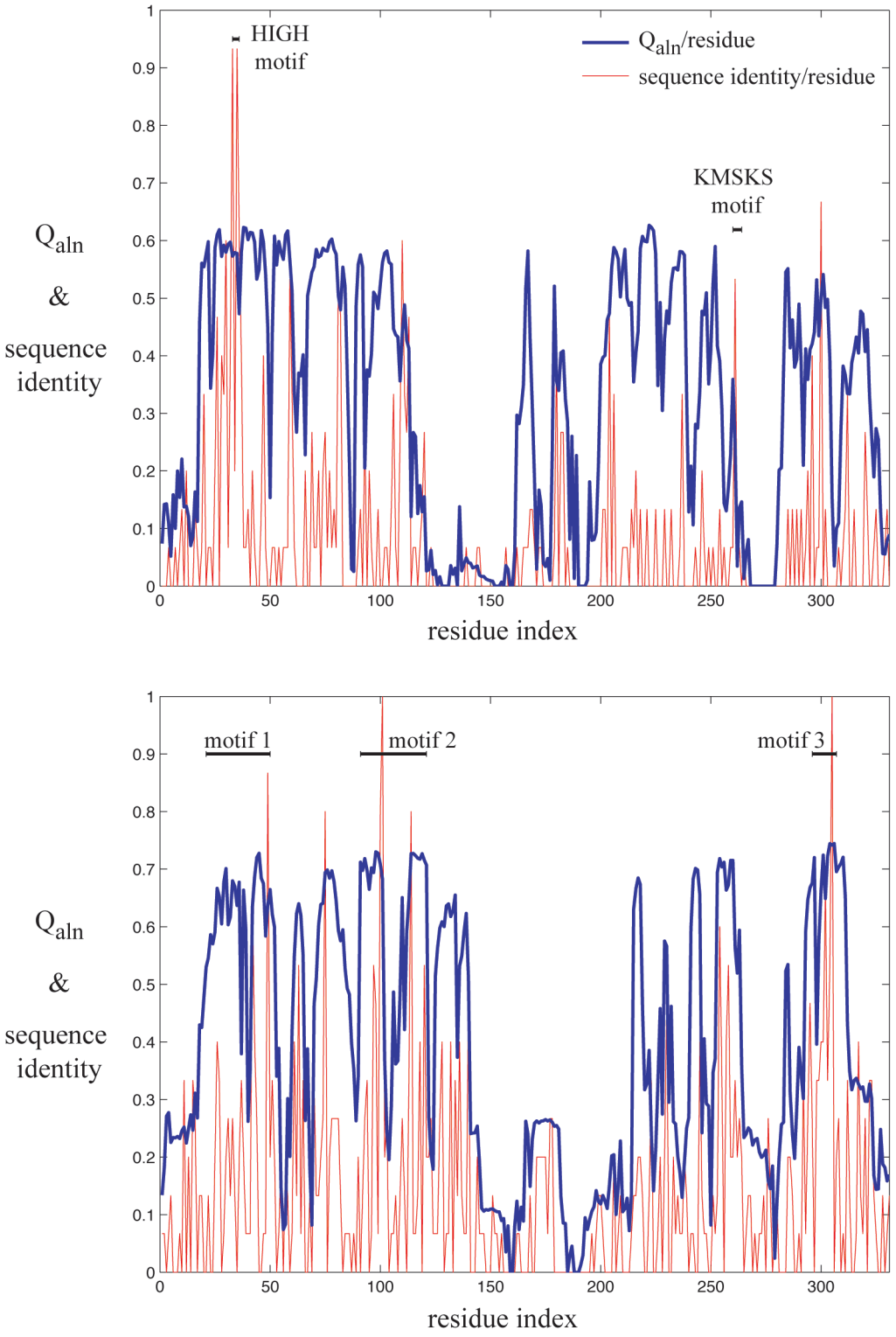

FIG. 13. Conservation of structure and sequence averaged over the multiply aligned nonredundant set (see Computational Methods) as a function of position in the protein for (top) class I GlnRS (1gtra2) and (bottom) class II AsnRS (11sca2). In general, structure is significantly conserved more than sequence information. Highly conserved sequence motifs are marked and labeled. Residue indices are related to PDB numbering in the supplementary material.

larger regions of structural similarity. The structural similarity peaks are broader and more frequent than the regions of high sequence similarity, and this is in accord with the guiding principle that conservation in sequence corresponds to conservation in structure. The reciprocal is not always true. The sequence conservation in the conserved core of the AARSs is not significantly higher than the sequence identity over the entire structure. For example, defining the conserved structural core as those residues with a Q_{aln} of >0.4 (as shown in Fig. 7) gives 13% and 22% sequence identity for GlnRS and AsnRS, respectively.

With regard to the recognized regions of sequence conservation, the class II synthetases show corresponding sequence and structural similarity in the regions referred to a motifs 1, 2, and 3. The class I AARSs show the same correspondence in the conserved HIGH region, but this is not the case for the KMSKS region. The sequence and structure conservation in the KMSKS region in moderate at best, and the KMSKS motif breaks the sequence-structure correspondence principle. This consensus motif is disrupted by gaps in the alignment due to large fluctuations in the corresponding structure. This motif is thought to be involved in stabilization of the transition state prior to formation of the aminoacyl-adenylate, and has been characterized as a "mobile loop" (32). The KMSKS region underscores the fundamental ambiguity in defining a one-dimensional representation (a multiple alignment) from a superposition of molecules in three-dimensional space (structural overlap). Typically, this ambiguity is of only minor significance.

For the class I AARSs, there are two other regions highly conserved in sequence and in structure. At residue index 110 and 300 in Fig. 13 (top) there are two peaks, corresponding to a conserved glycine and a conserved arginine, respectively, that are absent only in TrpRS and TyrRS. The arginine has been replaced by a methionine in the bacterial type ArgRS. Aside from the well-known motifs 1, 2, and 3 in the class II synthetases, we note an additional point of high sequence-structure conservation. There is a nearly completely conserved leucine (at residue index 75) which is occasionally replaced by isoleucine or valine and has been replaced by glutamate in the tetrameric GlyRS. The function of the conserved residues mentioned in this paragraph is not discussed in the literature, and their role is not obvious from examination of the crystal structures. Perhaps detailed molecular dynamics simulations will help in elucidating the function of these conserved residues. The plots also show that there is more sequence and structural conservation in the class II synthetases than in the class I AARSs.

Experimentally determined protein structures contain some inherent noise due to the limit of crystallographic resolution and experimental conditions which may result in conformational changes, e.g., apo versus ligand-bound crystal forms. In Fig. 9 and 10, we show the full structural dendrograms for the class I and class II synthetases, respectively, and the effect of the crystallographic noise in terms of Q_H of different structures for identical sequences. The majority of these effects are at $Q_H > 0.8$, with a few examples extending to $Q_H \sim 0.7$. Above this threshold, structure-based phylogenetic interpretations are unreliable, and only sequence-based analysis will yield accurate phylogenetic patterns.

Structural Evolutionary Profile of the AARSs

The modern set of AARSs are the result of a pattern of ancient gene duplications. The exact nature of this pattern, according to the structural phylogeny, is shown graphically in the nonredundant dendrograms in Fig. 11 and 12 for the separate synthetase classes. Each of the leaves on the structural trees is annotated by the specificity of the synthetase, e.g., D for AspRS, and by the appropriate genre, e.g., Da for AspRS of the archaeal genre and Db for AspRS of the bacterial genre. Synthetases that display noncanonical phylogenetic patterns are not and cannot be labeled by genre (see Fig. 5). These assignments are based on previous work (57) as described in Computational Methods. Note that the class II structural phylogeny also includes asparagine synthetase A, AsnA, which is not an AARS. The series of gene duplications that resulted in the modern forms of the AARSs also produced other proteins with related functions (50, 57). Although there are other proteins homologous to the AARSs that are not themselves AARSs, AsnA was included because its structure is known and it has a high degree of structural similarity to the AARSs. In fact, this homology was first detected at the sequence level (26) and confirmed by structural comparison (42), and the SCOP database includes AsnA as a member of the class II AARS protein family (40). AsnA uses a catalytic mechanism similar to that of the AARSs to synthesize asparagine from aspartate and ammonia in an ATP-dependent reaction.

Phylogenetic order of the AARSs. The AARSs of both classes exhibit a general monophyly with respect to amino acid specificity. This means that the AARSs display clusters that include one amino acid specificity to the exclusion of all others. For example, both the archaeal and bacterial versions of TyrRS form a distinct cluster to the exclusion of other specificities. The monophyly property can only be strictly demonstrated when all versions of a specific AARS are available. This issue is only a matter of concern for AARSs that exhibit the full canonical or basal canonical pattern (see Fig. 5), where there exist distinct archaeal and bacteria versions of a given synthetase. Except where noted for those AARSs that exhibit a noncanonical pattern, there is only one version of that molecule, and these synthetases can be adequately represented by any one example. We elaborate on these points below and compare our structural dendrogram with the sequence dendrograms of Woese et al. (57).

For many of the class I AARS that exhibit the full or basal canonical pattern, molecules for both the archaeal and bacterial genres are available in the PDB. Based on sequence analysis, both IleRS and TyrRS conform to the full canonical pattern. For IleRS, the structural dendrogram in Fig. 11 shows distinct clusters for the archaeal IleRS (now referred to simply as Ia) and bacterial IleRS (Ib), so the structural dendrogram is in agreement with the sequence phylogeny. Similarly, the branching of Ye, the archaeal type being represented by the eukaryotic TyrRS from *Homo sapiens* (see Addendum), and Yb in the structural dendrogram confirms the expected canonical pattern of TyrRS. The branching between Wb and Y is quite short, and a neighbor-joining tree (see the supplementary material) groups the *H. sapiens* TyrRS with the Wb cluster. The reason for the ambiguity of this grouping has already been deciphered at the sequence level (9, 47).

Brown et al. showed that the inclusion of archaeal sequences breaks the symmetry responsible for the ambiguous grouping and firmly supports monophyletic clustering of TrpRS and TyrRS. Within each cluster, the full canonical pattern is evident (9). We conclude that the crystal structures of TyrRS and TrpRS from the domain Archaea are required to remove the ambiguity of this branching in the structural dendrogram. This is one of three minor conflicts between the UPGMA and neighbor-joining tree representations. The others are mentioned below.

Although the canonical pattern is evident in both IleRS and TyrRS, we cannot completely verify the full canonical pattern. Since the archaeal TyrRS branch is being represented by TyrRS from *H. sapiens* and no structure for TyrRS from an archaeon is available, we cannot demonstrate that the eukaryotic structure forms a distinct group within the archaeal branch (see Addendum). The case is similar for IleRS. We have an additional point in common with the sequence phylogeny with regard to the bacterial subtypes of TyrRS. The term subtype is in reference to distinct clades within an organismal domain, and the two bacterial subtypes can be thought of as separate bacterial versions of a molecule. In the sequence phylogeny, *S. aureus* and *B. stearothermophilus* occupy part of one subtype (Yb1) and *T. thermophilus* is part of the second subtype (Yb2). This distinction is equally clear in the structural dendrogram.

Both MetRS and ArgRS exhibit the basal canonical pattern, and this is completely supported by the structural dendrogram. In each case, there is a clear distinction between the archaeal and bacterial genres. Woese et al. noted that the ArgRS tree cannot be reliably rooted between the archaeal and bacterial branches. The structural phylogeny confirms the existence of distinct archaeal (Ra) and bacterial (Rb) genres and is reliably rooted by the other class I synthetases. Here we appeal to the technique of reciprocal rooting (10, 33). Class I synthetases specific for C and Q show noncanonical patterns and thus are each appropriately represented in the structural dendrogram by a single structure.

For class II (see Fig. 12), AspRS conforms to the full canonical pattern, and this is supported in both the sequence and structure phylogenies. In this case structures are available which represent the bacterial, archaeal, and eukaryotic versions. In this region of the structural dendrogram, the full canonical pattern is evident. The Pa cluster contains two archaea and one bacterium, *T. thermophilus*, that acquired the archaeal type ProRS via horizontal transfer (57). The sequence phylogeny shows that the horizontally acquired ProRSs from a subgroup within the archaeal cluster, and this is also seen in the structural phylogeny. The grouping of bacterial species in the Hb cluster shows a minor disagreement with the sequence phylogeny, which indicates that the *Escherichia coli* and *T. thermophilus* HisRSs should group together to the exclusion of the *S. aureus* enzyme. This branching could probably be resolved with additional HisRS structures. AsnRS, SerRS, the class II LysRS, and the dimeric and tetrameric forms of GlyRS are all noncanonically distributed and are completely represented in the structural dendrogram.

The crystal structures of Wa, Ea, and Va and the bacterial version of the class I LysRS (K_Ib) would allow structural verification of the presence of the canonical pattern in these class I synthetases. Determination of the class II AARSs structures of Ha, Pb, Ta, Aa, Ab, and Fa would allow a more complete structural phylogeny. The sequence phylogenies of Woese et al. of the AARSs individually by specificity were rooted with synthetases of different specificities, and they observed that the monophyly rule is maintained in general. Additional crystal structures would therefore enrich the present picture, but we can be confident that the order of gene duplications established in the nonredundant structural dendrograms in Fig. 11 and 12 will likely remain unchanged. When the structural dendrogram can be verified by the sequence phylogeny, the two phylogenies are generally in agreement, and we expect that this agreement will not change with additional synthetase structures.

Some of the synthetases do not group monophyletically with respect to functional specificity. In Fig. 12, the cluster for AspRS includes not only the archaeal and bacterial versions of these synthetases, but an additional synthetase specificity, AsnRS, which arises within the archaeal branch. The AspRS-AsnRS cluster is a paraphyletic group. As mentioned above, paraphyly of the AspRS-AsnRS cluster and the GluRS-GlnRS cluster has been documented (57). We note that Eb and Q form a definite group, but without the structure of Ea we cannot add clear support to this case of paraphyly. The paraphyletic AspRS-AsnRS cluster is in close agreement with the sequence phylogeny. Woese et al. reported that AsnRS arises after the division between Da and Db but before the Eucarya-Archaea division. We observe that AsnRS evolves from the Da branch after the division between Da and De. The neighbor-joining structural dendrogram (supplementary material) agrees exactly with the sequence phylogeny. This ambiguity may be resolved with additional AspRS and AsnRS structures.

LysRS and GlyRS violate the monophyly rule in a different way. These synthetases are examples of polyphyly, which refers to synthetases of the same specificity that do not directly share a common ancestor of that specificity. LysRS is found in both a class I (K_I) and a class II (K_{II}) version and thus violates the class rule as well as the monophyly rule. The two versions of GlyRS are both class II synthetases. The form that is part of subclass IIA (see Fig. 4) is a homodimer, as are most class II synthetases. The second form is an $(\alpha\beta)_2$ tetramer. The class II synthetase fold is found only in the α-subunit, and conveniently this is the subunit that has been crystallized. The tetrameric GlyRS, $G_{(\alpha\beta)_2}$, is clearly one of the most divergent with respect to all other class II synthetases. Although it appears to be distantly associated with PheRS, this relationship is so distant that the two cannot be considered specifically related. The tetrameric GlyRS is found exclusively in most Bacteria, and the dimer GlyRS is found in some Bacteria and the other two domains of life in a noncanonical distribution (57). SerRS also has two distinct forms which are thought to be distributed polyphyletically. The second form, which is found in *M. thermoautotrophicus* and two species of the *Methanococcus* genus, is of unknown structure (57).

We now turn our attention to the complex history of enzyme displacement and horizontal gene transfer that characterizes the evolutionary path of the LysRSs. From our structural dendrogram, it is apparent that K_I branches very deeply with respect to the other class I enzymes (see Fig. 11). The meaning of a deep branching is that the gene duplication event that gave

rise to K_I also gave rise to an ancestral state for the protein that would evolve, through further gene duplications, to synthetases specific for C, E, and Q. Equivalently the magnitude of structural divergence between the modern K_I synthetase and its closest synthetase relative of a different specificity is large in comparison with other gene duplications that generated new specificities. The deep branch of K_I is an indication that the existence of this enzyme far predates the root of the universal tree. The phylogenetic distribution of K_I therefore should conform to some extent to the canonical pattern. Although initial sequence analysis indicated that this was not the case, an updated analysis, including more sequences of K_I, demonstrated that the K_I enzymes fit the basal canonical pattern (3). The authors reported a limited amount of HGT from Archaea to Bacteria (*Borrelia burgdorferi*, *Streptomyces coelicolor*, *Treponema denticola*, *Treponema pallidum*) and from Bacteria to Archaea (*C. symbiosum*), but there are distinct archaeal and bacterial genres for the K_I enzyme. Since GluRS was used to root the K_I sequence phylogeny of Ambrogelly et al., there is a strong indication that K_I of the bacterial genre should group monophyletically with the archaeal type K_I in our structural dendrogram.

The evolution of K_{II} is somewhat more complex. K_{II} is found in the majority of Bacteria, and this synthetase shows a noncanonical distribution with eukaryotes and two members of the Crenarchaeota acquiring the enzyme through horizontal gene transfer (57). K_{II} emerges as the result of a gene duplication, which gave rise to the Lys-Asp-Asn supercluster that occurred only shortly before the divergence of the Da and Db branches (see Fig. 12). In contrast to K_I, therefore, the emergence of K_{II} only slightly predates or possibly arises simultaneously with the formation of the root of the universal phylogenetic tree and the emergence of the archaeal and bacterial domains.

K_{II} is not incorporated into the archaeal genomes at this point in evolution, or it is quickly displaced by the existing K_I. K_{II}, however, is maintained through the present day in most of the bacterial lineages. Since K_I displays the basal canonical pattern and yet K_I is present in a minority of bacteria, in most ancestral Bacteria K_{II} displaced K_I. All known eukaryotes possess the noncanonically distributed K_{II} enzyme, and K_I has not been found in this domain of life (3, 57). This deviation from the full canonical pattern suggests a crucial horizontal transfer event in which the eukaryotes accepted the K_{II} enzyme from the Bacteria at an early stage of eukaryotic evolution. The HGT event appears to have completely displaced K_I in the eukaryotic lineage. The above results taken together lead us to propose a scenario for the coevolution of LysRS from class I and II, as shown in Fig. 14.

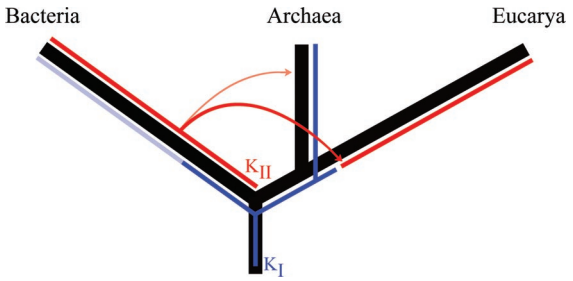

FIG. 14. Coevolutionary scenario of the class I and class II LysRS. The proposed points of appearance of the proteins are marked by the labels K_I and K_{II}. Red (blue) bars indicate the evolutionary paths of K_{II} (K_I) in the reference frame of the universal phylogenetic tree (black). The bold arrow marks a major HGT of K_{II} from Bacteria to Eucarya which displaces K_I, and the thin arrow shows a minor transfer from Bacteria to Archaea. The transparent blue line indicates displacement of K_I by K_{II} in all but a minority of the Bacteria. Other minor HGT events are not marked but are mentioned in the text.

Subclass definitions and supercluster order. The AARSs have been divided into subclasses which are meant to group the most closely related synthetases of different specificities. The standard subclass divisions are shown in Fig. 4, and these divisions were determined by a combination of sequence and structure comparison in the absence of an overall structural phylogeny, as presented here (16, 17, 21, 48). Since the sequence comparisons at the subclass level are at or beyond the limit of reliability, there is some disagreement with the structural subclasses that we suggest. In reference to our structural dendrograms, a cutoff of $Q_H \cong 0.41$ sufficiently defines the subclass level for the AARSs of each class. For the class I AARSs we propose five subclasses instead of the standard three. The structural phylogeny clearly shows that K_I cannot be included in a group with GlnRS and GluRS, or in any other subclass for that matter. A similar argument follows for ArgRS, so K_I and R each form new and separate subclasses. We also observe a close structural grouping of CysRS with GlnRS and GluRS and shift CysRS from the standard subclass IA to the structural subclass IB. Wolf et al. observed that GlnRS, GluRS, and CysRS each share a small homologous subdomain insertion in the center of the catalytic domain that is not found in any other synthetase (62). The root of the Cys-Glu-Gln supercluster, therefore, represents the point of this insertion. Due to a lack of crystallographic resolution in this region, much of the insertion is missing in the CysRS structures. The remainder of the catalytic domain, however, is enough to suggest a structural grouping of CysRS with GluRS and GlnRS.

Within subclass IA, Brown and Doolittle present a sequence-based phylogeny of IleRS, ValRS, and LeuRS, and they concluded that these synthetases group monophyletically (10). Our structural phylogeny supports this point, and additionally both phylogenies show that ValRS and IleRS group closely together to the exclusion of LeuRS, which forms an outgroup. Further, we show that MetRS groups in with the Val-Ile cluster and that LeuRS is the most distant relative of the structural subclass IA.

With respect to the class II synthetases, there appears to be a general confusion in the literature regarding the subclass assignments of tetrameric GlyRS and AlaRS. There are two structurally distinct forms of GlyRS: the tetrameric form, $G_{(\alpha\beta)_2}$, and the dimeric form, $G_{\alpha 2}$. The dimeric form is part of the H-T-P-S supercluster and is properly classified in the standard subclass IIA. Our structural dendrogram upholds this classification. In some cases, $G_{(\alpha\beta)_2}$ is classified as a subclass IIA enzyme (23), while elsewhere it is classified as a subclass IIC enzyme with PheRS (17). Occasionally, the subclasses appear without mention of the two distinct forms of GlyRS, and a generic GlyRS is simply placed in subclass IIA (48).

Our data suggest that $G_{(\alpha\beta)_2}$ clusters with PheRS. Although this relationship is distant, it falls within the subclass threshold in terms of Q_H, so we include $G_{(\alpha\beta)_2}$ in subclass IIC, which is in

agreement with previous work (17). There is a further similarity between PheRS and $G_{(\alpha\beta)_2}$ as both enzymes have an $(\alpha\beta)_2$ quaternary arrangement, and these are the only synthetases with this quaternary structure. The minor form of SerRS (see above) and AlaRS cannot be reliably grouped into subclasses until their structures are determined. In fact, AlaRS has been classified as a IIC enzyme (17) and alternatively with subclass IIA (48). So as not to add to this confusion, we refrain from including AlaRS in any of the subclasses.

The phylogenetic order of the subclass IIA synthetases (46) agrees qualitatively with the H-$G_{\alpha 2}$-T-P-S supercluster organization observed in our structural dendrogram. The neighbor-joining tree (supplementary material) exchanges the positions of $G_{\alpha 2}$ and Tb. This same ambiguity has been observed before (46), and additional structures, specifically of Ta and Pb, may resolve this discrepancy. The phyletic order given before (46) differs quantitatively in that the relative branch lengths are not the same. This minor difference is due to the fact that we include the entire catalytic domain in combination with a structure-based measure to generate the phylogenetic order in this supercluster, while Ribas de Pouplana et al. employed a sequence-based measure over the most conserved fragments of this domain.

Evolutionary Events and Structural Divergence

There are two major questions that we strive to answer by comparing homologous protein structures. In what specific ways has the evolutionary dynamic changed protein shape over time? What can monitoring the change in protein shape tell us about the evolutionary process?

A general hierarchy of evolutionary events is recorded in the structures of the AARSs. Protein shape changes over the evolutionary course in response to a combination of physical forces and natural selection. The response of organisms and genetic elements to these forces, which drive the evolutionary dynamic, is observed in patterns of vertical and horizontal gene transfer. Accumulated effects of these evolutionary mechanisms, lead to specific evolutionary events.

We can categorize these events in a hierarchy from the most distant to the more recent events in time. The most distantly detectable events predate the root of the universal phylogenetic tree. In Fig. 11 and 12 these correspond to the period of early AARS evolution that gave rise, though a series of gene duplications, to the AARS subclasses. Quantitatively this region of similarity occurs in the range $Q_H \cong \{0.35, 0.41\}$. This value and others quoted below are in reference to both the class I and class II phylogenies combined. The next regime, $Q_H \cong \{0.41, 0.58\}$, encompasses further gene duplications giving rise to most of the synthetase specificities. The specificities that are distributed paraphyletically arise later.

During the next major evolutionary stage, the formation of the root of the universal phylogenetic tree, the archaeal and bacterial organismal domains diverge from one another, and gene duplications that mark this divergence fall in the range $Q_H \cong \{0.45, 0.65\}$. Since there is only one example (AspRS) where the archaeal and eukaryotic versions for a given synthetase have both been structurally determined, we cannot reliably assign a regime in Q space that corresponds to the separation of the domains Archaea and Eucarya from their common ancestor. In the case of AspRS, the divergence of the eukaryote from the ancestral archaeal AspRS occurs at $Q_H \cong 0.73$. The final level of divergence is accounted for by intradomain speciation events in the range $Q_H \cong \{0.53, 0.88\}$.

The above results indicate that there is a general hierarchy to the evolutionary development of the AARSs across specificities and even across synthetase classes. This fits with the notion proposed by Nagel and Doolittle that the synthetases of both classes underwent "coordinate development" (41). The meaning of this notion is that the synthetases of the two classes were evolving analogous functionalities concurrently. Although we note that there is a general hierarchy to these evolutionary stages, there is considerable overlap of these stages as well (see Fig. 11 and 12). Examining the evolution of any one AARS specificity shows that these stages are precisely hierarchical, meaning that once the specificity is established, with the exception of GlnRS and AsnRS, the next event to follow is the split between archaeal and bacterial types. This is followed by intradomain subtype divergence, e.g., TyrRS, which is in turn followed by more local intradomain speciation events. When looking across the synthetases of different specificities and even across the synthetase classes, these events broaden into evolutionary stages. In fact, the stages become so broad they overlap.

With these ideas in mind, we focus on some specific cases of particular interest. For MetRS, the amount of structural divergence at the split between archaeal and bacterial genres is smaller than the amount of structural divergence observed for any "genre split" across all of the other AARSs of both classes (see Fig. 15). The archaeal and bacterial type MetRSs are even more structurally similar to each other than are the two bacterial subtypes of the TyrRS clade. In addition, the archaeal and bacterial types of TyrRS are structurally more divergent than are some synthetases of different specificities, e.g., the relationship between MetRS, ValRS, and IleRS.

There are two limiting interpretations of these results. Evolutionary events are not specifically localized in time but are characterized by broad stages where each genetic element is diverging at a similar rate but independently at a different point in time. Under this assumption, the broadening of evolutinary events that we observe in terms of structural divergence may indicate that these evolutionary events are also broad in time. Regarding the example of TyrRS and MetRS, it is amazing to contemplate that while for TyrRS the bacteria are already beginning to form distinct subtypes, the formation of distinct archaeal and bacterial type MetRSs has yet to occur. Further, the genre split for TyrRS predates the advent of many of the other synthetase specificities. It is clear, however, that there is a limitation to this broadening effect. GlnRS and AsnRS are late evolutionary inventions (57), and neither shows even the basal canonical distribution. Presumably that is because the major organismal domains had already diverged before the invention of this functionality.

The genre split for MetRS occurs at $Q_H \cong 0.65$ and AsnRS arises at $Q_H \cong 0.77$. Thus, there appears to be a very short window in Q space between the evolution of new specificities that may display the canonical pattern and new specificities that arise too late to possibly contain this historical trace. We must consider that our structural homology measure may not

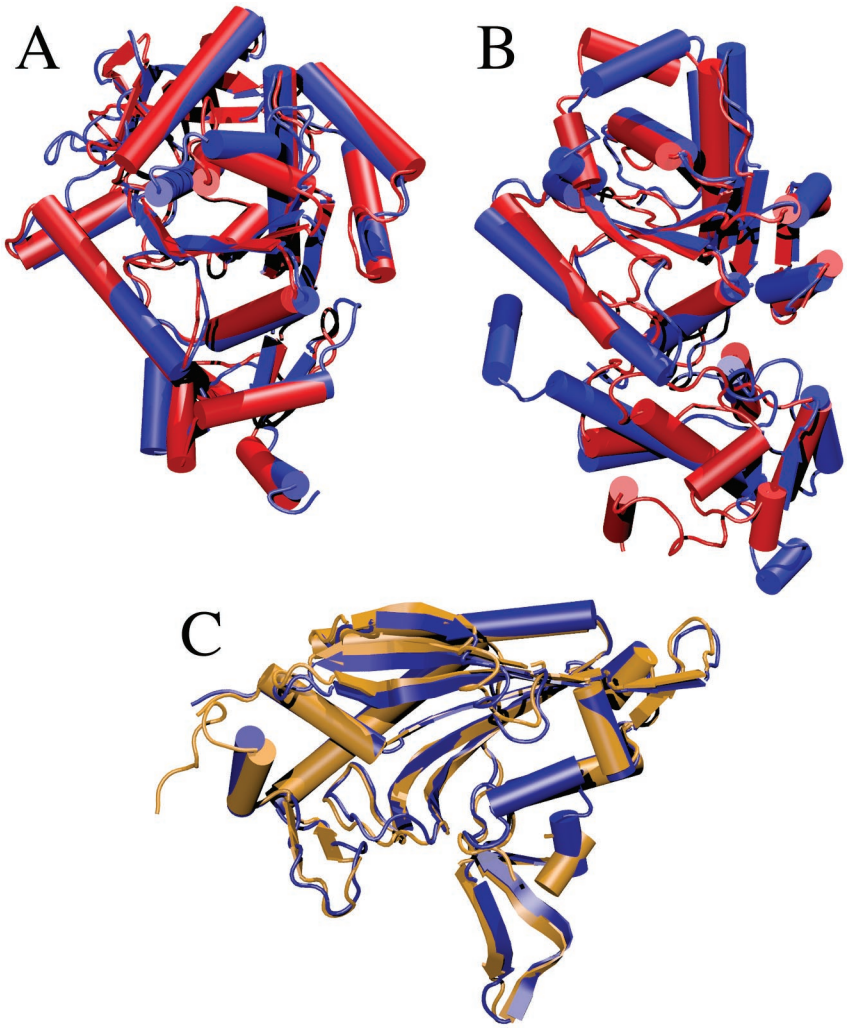

FIG. 15. In panel A the structural overlap of Ma (*E. coli* MetRS, blue) and Mb (*T. thermophilus* MetRS, red) is shown, and Ya (the archaeal type is being represented by the eukaryotic *H. sapiens* TyrRS, [see Addendum], blue) overlapped with Yb (*T. thermophilus* TyrRS, red) is shown in panel B. Both pictures show the magnitude of structural divergence associated with the split between the bacterial and archaeal organismal domains. Note that there is visibly more structural divergence between Ya and Yb as opposed to the relatively higher structural homology of Ma and Mb. Panel C depicts the effect of horizontal gene transfer on protein shape. Two ProRS structures of the archaeal genre are shown. ProRS was acquired by the bacterium *T. thermophilus* via HGT (orange) and it is overlapped with the "native" archaeal ProRS of *M. thermoautotrophicum* (blue). The structures are nearly indistinguishable.

scale linearly with time or that the evolution of structure itself is not a linear function of time.

The second interpretation is that the evolutionary events are localized in time and the broadening is an artifact of varying rates of divergence between different genetic elements. The actual situation is likely to be some synthesis of these two hypotheses. Certainly the evolutionary profile of the AARSs indicates that these events or stages are broad in nature, but their evolutionary course does not seem reasonable without invoking the notion that structural change in proteins occurs at various rates. Although we are not the first to identify these hypotheses, the structural phylogeny of the AARSs presents an opportunity to explore these ideas concretely.

After the three organismal domains were established, the AARSs continued to display frequent interdomain horizontal gene transfer. These events have been clearly documented (57, 62), but here we discuss the effect of this kind of HGT on protein shape. The effect of horizontal gene transfer varies from subtle to dramatic, and this depends on the range and time of the transfer. A local transfer event between two closely related bacteria may be nearly undetectable, while a long-range horizontal transfer event between an archaeon and a bacterium may leave the bacterium with a protein that is almost unrecognizably related to a protein of the same function in other bacteria. Proteins can be retained over the evolutionary course though vertical evolution, and they will differ from orthologs by the expected amount of divergence in accord with the general divergence of the two organisms in question. Horizontal transfer has the apparent effect of increasing the amount divergence observed within a group. A single horizontal transfer, between say an archaeon and a bacterium, can potentially have the effect (on the gene in question) of increas-

ing the amount of divergence by at most 3 billion years. Essentially this gene, now resident in a bacterium, appears to have acquired 3 billion years of vertical divergence with respect to the orthologous gene in other bacteria. The process of horizontal transfer followed by acquisition and selective retention of "nonnative" genetic elements appears to be a mechanism for rapid introduction of genetic novelty.

The effect of horizontal gene transfer on protein shape is both striking and simple. The overlap between the two ProRSs of the archaeal genre shows that the two backbone structures are nearly indistinguishable, with $Q_H \cong 0.75$, as shown in Fig. 15C. One structure is from the archaeon *M. thermoautotrophicus*, and the other is from the bacterium *T. thermophilus.* Since the structure of the bacterial type ProRS is unavailable, we note that the expected difference between bacterial and archaeal genres is in the range $Q_H \cong \{0.45, 0.65\}$. The boundaries of this range are depicted by structural overlaps in Fig. 15A and B. The discrepancy between the expected amount of structural divergence for protein from a bacterium and an archaeon and that observed for the ProRSs is accounted for by horizontal transfer. *T. thermophilus* acquired ProRS from the Archaea (57), and the effect was that the ProRS of *T. thermophilus* appears nearly structurally indistinguishable from the "native" archaeal type ProRS.

Structural Conservation of Substrate Interactions

In the top panel of Fig. 16, we examine the conservation of the protein backbone structure of the 16 residues that make contact with the amino acid substrate tyrosine of the class I TyrRS enzyme from *B. stearothermophilus.* Structural conservation is plotted in terms of Q_{aln} for these residues averaged over the multiple alignment for five separate levels of divergence: the class I level, subclass IC level, the specificity level for TyrRS, the bacterial version of TyrRS only, and the bacterial subtype level (Yb1). There is a general trend for Q_{aln} to increase as the level of divergence decreases, but we note that the majority of the structure associated with amino acid binding is conserved for all class I synthetases. In the plot residue indices 2 to 6, 10 to 14, and 16 are well conserved at the class level. The structural positions represented by the residue indices 1, 7, and 15 are established at the subclass level. Presumably these backbone positions were inserted to accommodate recognition of the large aromatic residues tyrosine and tryptophan. We note very little change in backbone conservation between the subclass level, including TrpRS and TyrRS, and the specificity level for TyrRS. The backbone structure need not change in shape to account for the generation of enzyme specificity, so the amino acid side chains must be the main source of discrimination between tyrosine and tryptophan. The protein structure associated with residues indices 8 and 9 appears to reach its modern state at the bacterial genre level. These residues are part of an inserted loop in Yb that is not found in Ya. They make contact with the carbonyl and amide termini of the substrate tyrosine and seem to confer added stability to the enzyme substrate interaction.

We now examine the structural conservation associated with the interactions between the AspRS catalytic domain of *Escherichia coli* and the aspartyl-adenylate and tRNA substrates (see Fig. 16, bottom). We divide the 75 residues in contact with these substrates into four groups: residues which make contact with the aspartyl portion of the Asp-AMP, the adenylate region of this substrate, the tRNA molecule except the CCA stem, and residues making contact with only the CCA trinucleotide. For clarity we show the structural conservation for three levels of divergence: the class II level, subclass IIB level, and the blue curve shows structural differences between the apo and ligand bound forms of the *E. coli* AspRS. As for the class I TyrRS enzyme, we note that the backbone structure associated with amino acid binding, in the *E. coli* AspRS structure, is well conserved across the entire class with the exception of two backbone positions that appear to be specific for aspartate binding (residue indices 9 and 10) and one (residue index 15) that is established at the subclass IIB level. The same is true, albeit with somewhat more structural fluctuation, for the residues in contact with the AMP portion of the aspartyl-adenylate.

The only completely conserved residues in all of the class II AARSs are two arginines. One arginine interacts with the AMP (residue index 18 in Fig. 16), and the other interacts with both the AMP and the CCA trinucleotide (residue indices 30 and 65). The backbone positions of these residues are well conserved at the class level. Taken together, these results suggest that aminoacyl-adenylate binding is a function of the AARSs that dates back to their inception, and the amino acid substrate specificity is largely a function of the amino acid side chains present in the active site and not the protein backbone structure.

Structural conservation for residues contacting the tRNA tells a different story. These contact residues do not include contacts to the tRNA anticodon, which are made by a separate anticodon binding domain in most of the synthetases. The synthetase catalytic domain, however, does make a significant amount of contacts the the helical stem of the tRNA leading down to the anticodon loop. At the class level, there is a striking difference between the protein backbone conservation associated with residues contacting the CCA trinucleotide as opposed to the rest of the tRNA. A significant portion of the residues contacting the CCA are associated with backbone positions that are structurally very well conserved at the class level, while contacts to the remainder of the tRNA are not conserved at this level. This indicates that binding to the CCA is an aboriginal feature of the class II synthetases while protein structure associated with binding to other regions of the tRNA arose at a later time. Some of this structure appears at the subclass level (yellow curve in Fig. 16), and is conserved for LysRS, AsnRS, and both bacterial and archaeal versions of AspRS. This structure corresponds to residues indices 40 to 48 and 53, and these residues mainly make contact to the acceptor stem and the junction between the acceptor stem and the anticodon stem. It is interesting that this part of the catalytic domain structure emerges coincidently with the acquisition of the OB fold anticodon binding domain which is common only to the subclass IIB synthetases.

The above results indicate that the regions of protein structure associated with aminoacyladenylate and CCA binding are common and thus ancestral to all the AARSs at the class level while structure associated with the reminder of the tRNA arose later and in stages. Much of the protein structure responsible for interacting with the tRNA seems to be in place by

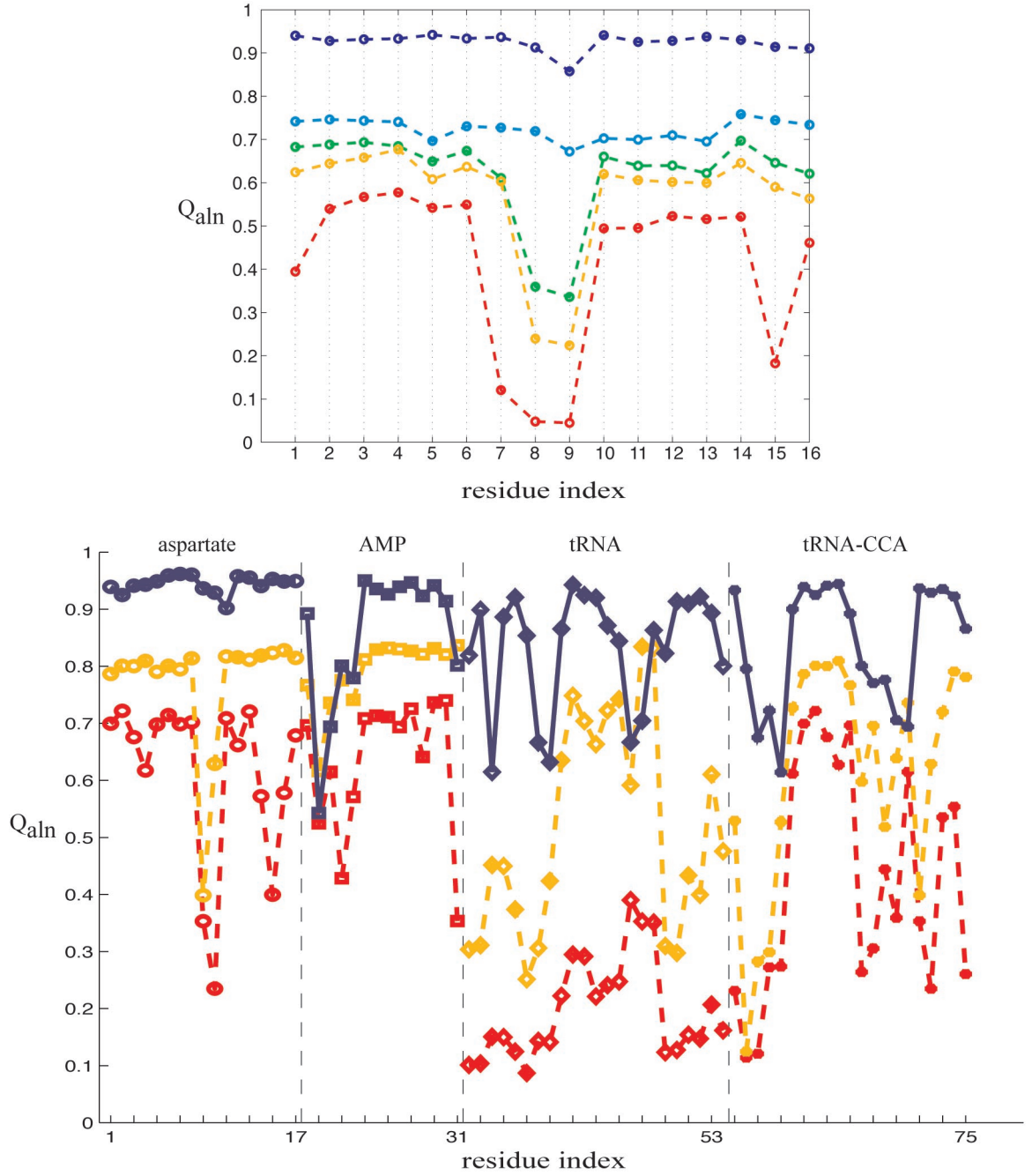

FIG. 16. Conservation of contacts between the AARS catalytic domain and the amino acid substrate, AMP, tRNA-CCA acceptor trinucleotide, and the rest of the tRNA. The top panel shows conserved backbone structure of TyrRS (*B. stearothermophilus*) that makes contact with the substrate tyrosine at five levels of divergence: class I (red), subclass IC (yellow), TyrRS specificity (green), Yb (cyan), and Yb1 (blue). The bottom panel shows backbone conservation of AspRS from *E. coli* for residues that make contact with the aspartyladenylate and tRNA substrate at three levels of divergence: class II (red), subclass IIB (yellow), and apo versus ligand-bound AspRS structures (blue). Aminoacyl-adenylate binding contacts are well conserved already at the class level, while the tRNA contacts are only partially present. The specific residues in the PDB files corresponding to the indices are given in the supplementary material.

the formation of the synthetase subclasses with the rest inserting gradually as the synthetases themselves become specific for particular amino acids. A possible interpretation of Fig. 16 may be that the structure of tRNA changed during the evolution of the synthetases (19, 55), or the synthetases may have evolved to adapt in shape to an already established tRNA. Alternatively, there may have been another molecule involved, e.g., a ribozyme no longer extant, in the aminoacylation reaction as the AARSs were evolving into their modern form. A structural study of tRNA evolution may resolve this issue.

CONCLUSION

Where comparison is possible, the sequence and structural phylogenies are generally in agreement. The sequence analysis complements the structural phylogeny. The sequence phylogeny of Woese et al. (57) has been important not only for annotating the structural phylogeny but also for determining an appropriate measure of structural homology. The consideration of insertions is crucial for constructing a structural phylogeny of distantly related proteins. Using only the aligned portions results in structure-based phylogenies that are in fundamental disagreement with established sequence phylogenies. For example, a structural dendrogram based only on Q_{aln} places HisRS in the Lys-Asp-Asn supercluster and not with the other class IIA synthetases.

The rapid rate at which protein sequences and structures are being added to the public databases will soon allow construction of the complete evolutionary history for the AARSs, and it will become possible to conduct a similar analysis for many more enzyme families. Additional structures are predicted not to alter the phylogeny presented here, but crystal structures of the class I synthetases Wa, Ea, Va, and K_Ib and the class II synthetases Ha, Pb, Ta, Aa, Ab, and Fa would complete the evolutionary history of the AARSs.

Conservation of sequence implies conservation of structure, but the reverse is not always true. As Fig. 13 clearly shows, the core structure is conserved, even with less than approximately 15% sequence identity. The structural conservation of the aminoacyl-adenylate binding pocket has important implications for protein design. As the backbone structure is well conserved and specificity is largely a function of the amino acid side chains lining the active site pocket, this enzyme has already been shown to allow incorporation of unnatural amino acids (30).

The variation in protein shape observed in our study provides clues about the nature of major evolutionary events. With the exception of AsnRS and GlnRS, the separate synthetase specificities, as part of the translational process, were in their modern form before the root of the universal phylogenetic tree (UPT). But this is not the full story. The root of the UPT is defined by the separation of the domains Bacteria and Archaea (plus Eucarya) from their common ancestral states, the first recognizable speciation event. What is the nature of this root? According to Woese (59–61), the formation of the root of the UPT appears to have been a gradual transition from a phase of evolution where horizontal gene flow dominated the evolutionary dynamic to a phase where vertical inheritance was the dominant evolutionary force, leading to speciation. The boundary between these qualitatively different evolutionary phases is referred to as the Darwinian threshold.

The theory asserted by Woese is that as different subsystems of the cell (say in the protobacterial cell type) became sufficiently complex and specialized to that cellular environment, they became more and more refractory to displacement by horizontally acquired genes from contemporaneously evolving cell types. At the gene level, these ideas may be interpreted to indicate that we can identify the Darwinian threshold for each gene contributing a gene product to a cellular subsystem (such as translation). The Darwinian threshold for each gene is marked by the first emergence of species-specific (archaeal and bacterial) versions of the gene. Woese further identifies the translation apparatus as one of the first systems to cross the Darwinian threshold, as a stable and reliable translation machinery is necessary for vertical inheritance to dominate the evolutionary course. Since the AARSs are an integral part of the translation apparatus, these proteins may have been among the first molecular pioneers to cross the Darwinian threshold.

We observe a broad range of structural divergence associated with comparing archaeal and bacterial versions of the different AARS specificities, MetRS (Ma versus Mb) having the smallest amount of structural divergence and TyrRS displaying the largest amount of structural divergence between archaeal and bacterial types. What does this tell us about the evolution of translation? TyrRS may simply have evolved at a faster rate than the other AARSs since the time of the root of the UPT. We speculate that it is also possible that TyrRS crossed the Darwinian threshold before any of the other AARSs and that MetRS was the last to cross the Darwinian threshold. In this scenario, the formation of the root of the UPT is broad in time, and perhaps this is evidence for the gradual "crystallization" of different cellular subsystems over time (60).

Although the root of the UPT may be broad, it has a clear boundary, and this is demonstrated by the emergence of AsnRS and GlnRS. These synthetases evolved after the formation of the root of the UPT was essentially complete, and the canonical pattern cannot be and is not detected for these enzymes (57). The dissemination of AsnRS and GlnRS across the three domains of life, though GlnRS has not been found in Archaea, occurred through HGT. The frequent presence of HGT in the evolution of the synthetases reminds us that the barrier separating species and even organismal domains is not insurmountable.

ACKNOWLEDGMENTS

We are grateful to Carl Woese and Gary Olsen for open discussions and continuous encouragement. Their comments regarding the manuscript were invaluable. We thank Stephen Cusack for providing the structure of asparaginyl-tRNA synthetase from *T. thermophilus* and Rob Russell and Geoffrey Barton for providing the STAMP algorithm. Many thanks to Michael Heath for discussions concerning the QR factorization.

P. O. was supported by an NIH Institutional NRSA in Molecular Biophysics (5T32GM08276).

ADDENDUM

Recently, a crystal structure of TyrRS from the archaeon *M. jannaschii* (36) has become available, which now allows us to verify the full canonical pattern structurally for this synthetase. A structure of the non-discriminating archaeal type AspRS from *T. thermophilus* (13) has also become available. We present updated structure-based phylogenies in the supplementary material (supplementary figures 3 and 4) available at http://www.scs.uiuc.edu/~schulten/aars_supmat.pdf. Inclusion of these new structures strengthens the agreement between the structure- and sequence-based (57) phylogenies of the AARSs.

REFERENCES

1. **Ahel, I., C. Stathopoulos, A. Ambrogelly, A. Sauerwald, H. Toogood,T. Hartsch, and D. Söll.** 2002. Cysteine activation is an inherent *in vivo* property of prolyl-tRNA synthetases. J. Biol. Chem. **277:**34743–34748.

2. **Altschul, S. F., T. L. Madden, A. A. Schaeffer, 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.
3. **Ambrogelly, A., D. Korencic, and M. Ibba.** 2002. Functional annotation of class I lysyl-tRNA synthetase phylogeny indicates a limited role for gene transfer. J. Bacteriol. **184:**4594–4600.
4. **Ambrogelly, A., I. Ahel, C. Polycarpo, S. Bunjun-Srihari, B. Krett, C. Jacquin-Becker, B. Ruan, C. Kohrer, C. Stathopoulos, U. RajBhandary, and D. Söll.** 2002. *Methanocaldococcus jannaschii* prolyl-tRNA synthetase charges $tRNA^{Pro}$ with cysteine. J. Biol. Chem. **277:**34749–34754.
5. **Becker, H. D., and D. Kern.** 1998. *Thermus thermophilus*: A link in evolution of the tRNA-dependent amino acid amidation pathways. Proc. Natl. Acad. Sci. USA **95:**12832–12837.
6. **Benner, S. E., P. Koehl, and M. Levitt.** 2000. The ASTRAL compendium for structure and sequence analysis. Nucleic Acids Res. **28:**254–256.
7. **Berthet-Colominas, C., L. Seignovert, M. Hartlein, M. Grotli, S. Cusack, and R. Leberman.** 1998. The crystal structure of asparaginyl-tRNA synthetase from *Thermus thermophilus* and its complexes with ATP and asparginy-adenylate: the mechanism of discrimination between asparagine and aspartic acid. EMBO J. **17:**2947–2960.
8. **Blake, J. D., and F. E. Cohen.** 2001. Pairwise sequence alignment below the twilight zone. J. Mol. Biol. **307:**721–735.
9. **Brown, J. R., F. 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.
10. **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. USA **92:**2441–2445.
11. **Brown, J. R., and W. F. Doolittle.** 1999. Gene descent, duplication and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. J. Mol. Evol. **49:**485–495.
12. **Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter.** 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273:**1058–1073.
13. **Charron, C., H. Roy, M. Blaise, R. Giege, and D. Kern.** 2003. Non-discriminating and discriminating aspartyl-tRNA synthetases differ in the anticodon-binding domain. EMBO J. **22:**1632–1643.
14. **Chothia, C., and A. M. Lesk.** 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. **5:**823–826.
15. **Curnow, A., D. Tumbula, J. Pelaschier, B. Min, and D. Söll.** 1998. Glutamyl-tRNA amidotransferase in *Deinococcus radiodurans* may be confined to asparagine biosynthesis. Proc. Natl. Acad. Sci. USA **95:**12838–12843.
16. **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.
17. **Cusack, S.** 1995. Eleven down and nine to go. Nat. Struct. Biol. **2:**824–831.
18. **Eastwood, M. P., C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes.** 2001. Evaluating protein structure-prediction schemes using energy landscape theory. IBM J. Res. Dev **45:**475–497.
19. **Eigen, M., and R. Winkler-Oswatitsch.** 1981. Transfer-RNA, and early gene. Naturwissenschaften **68:**282–292.
20. **Eriani, G., G. Dirheimer, and J. Gangloff.** 1991. Cysteinyl-tRNA synthetase: determination of the last *E. coli* aminoacyl-tRNA synthetase primary structure. Nucleic Acids Res. **19:**265–269.
21. **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.
22. **Fox, G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen, and C. R. Woese.** 1980. The phylogeny of prokaryotes. Science **209:**475–497.
23. **Franckyln, C., J. J. Perona, J. Puetz, and Y.-M. Hou.** 2002. Aminoacyl-tRNA synthetases: versatile players in the changing theater of translation. RNA **8:**1363–1372.
24. **Galagan, J. E., C. Nusbaum, A. Roy, M. G. Endrizzi, P. Macdonald, W. FitzHugh, S. Calvo, R. Engels, S. Smirnov, D. Atnoor, A. Brown, N. Allen, J. Naylor, N. Stange-Thomann, K. DeArellano, R. Johnson, L. Linton, P. McEwan, K. McKernan, J. Talamas, A. Tirrell, W. Ye, A. Zimmer, R. D. Barber, I. Cann, D. E. Graham, D. A. Grahame, A. M. Guss, R. Hedderich, C. Ingram-Smith, H. C. Kuettner, J. A. Krzycki, J. A. Leigh, W. Li, J. Liu, B. Mukhopadhyay, J. N. Reeve, K. Smith, T. A. Springer, L. A. Umayam, O. White, R. H. White, E. C. de Macario, J. G. Ferry, K. F. Jarrell, H. Jing, A. J. Macario, I. Paulsen, M. Pritchett, K. R. Sowers, R. V. Swanson, S. H. Zinder, E. Lander, W. W. Metcalf, and B. Birren.** 2002. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. Genome Res. **12:**532–542.
25. **Gesteland, R. F., T. R. Cech, and J. F. Atkins.** 1999. The RNA world: the nature of modern RNA suggests a prebiotic RNA. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
26. **Hinchman, S. K., S. Henikoff, and S. M. Schuster.** 1992. A relationship between asparagine synthetase A and aspartyl tRNA synthetase. J. Biol. Chem. **5:**144–149.
27. **Hountondji, C., F. Lederer, P. Dessen, and S. Blanquet.** 1986. *Escherichia coli* tyrosyl- and methionyl-tRNA synthetases display sequence similarity at the binding site for the 3′-end of tRNA. Biochemistry **25:**16–21.
28. **Householder, A. S.** 1958. Unitary triangularization of a nonsymmetric matrix. J. Assoc. Comput. Mach. **5:**339–342.

28a. **Humphrey, W., A. Dalke, and K. Schulten.** 1996. VMD—visual molecular dynamics. J. Mol. Graphics **14.1:**33–38.

29. **Ibba, M., A. Curnow, and D. Söll.** 1997. Aminoacyl-tRNA synthesis: divergent routes to a common goal. Trends Biochem. Sci. **22:**39–42.
30. **Ibba, M., P. Kast, and H. Hennecke.** 1994. Substrate specificity is determined by amino acid binding pocket size in *Escherichia coli* phenylalanyl-tRNA synthetase. Biochemistry **33:**7107–7112.
31. **Ibba, M., S. Morgan, A. W. Curnow, D. R. Pridmore, U. C. Vothknecht, W. Gardner, W. Lin, C. R. Woese, and D. Söll.** 1997. A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. Science **278:**1119–1122.
32. **Ibba, M., and D. Söll.** 2000. Aminoacyl-tRNA synthesis. Annu. Rev. Biochem. **69:**617–50.
33. **Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata.** 1989. Evolutionary relationship of Archaebacteria, Eubacteria, and erukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86:**9355–9359.
34. **Jacquin-Becker, C., I. Ahel, A. Ambrogelly, B. Ruan, D. Söll, and C. Stathopoulos.** 2002. Cysteinyl-tRNA formation and prolyl-tRNA synthetase. FEBS Lett. **514:**34–36.
35. **Kamtekar, S., W. Kennedy, J. Wang, C. Stathopoulos, D. Söll, and T. Steitz.** 2003. The structural basis of cysteine aminoacylation of $tRNA^{Pro}$ by prolyl-tRNA synthetases. Proc. Natl. Acad. Sci. USA **100:**1673–1678.
36. **Kobayashi, T., O. Nureki, R. Ishitani, A. Yaremchuk, M. Tukalo, S. Cusack, K. Sakamoto, and S. Yokoyama.** 2003. Structural basis for orthogonal tRNA specificities of tyrosyl-tRNA synthetases for genetic code expansion. Nat. Struct. Biol. **10:**425–432.
37. **Lee, N., Y. Bessho, K. Wei, J. Szostak, and H. Suga.** 2000. Ribozyme-catalyzed tRNA aminoacylation. Nat. Struct. Biol. **7:**28–33.
38. **Low, S. C., and M. J. Berry.** 1996. Knowing when not to stop: selenocysteine incorporation in eukaryotes. Trends Biochem. Sci. **21:**203–208.
39. **Min, B., J. T. Pelaschier, D. E. Graham, D. Tumbula-Hansen, and D. Söll.** 2002. Transfer RNA-dependent amino acid biosynthesis: An essential route to asparagine formation. Proc. Natl. Acad. Sci. USA **99:**2678–2683.
40. **Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia.** 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247:**536–540.
41. **Nagel, G. M., and R. F. Doolittle.** 1991. Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proc. Natl. Acad. Sci. USA **88:**8121–8125.
42. **Nakatsu, T., H. Kato, and J. Oda.** 1998. Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. Nat. Struct. Biol. **5:**15–19.
43. Reference deleted.
44. **Olkin, J. A., L. P. Heck, and K. Naghshineh.** 1996. Automated placement of transducers for active noise control: performance measures. International Conference on Acoustics, Speech, and Signal Processing, Atlanta, Ga.
45. **RajBhandary, U. L.** 1994. Initiator transfer RNAs. J. Bacteriol. **176:**547–552.
46. **Ribas de Pouplana, L., J. R. Brown, and P. Schimmel.** 2001. Structure-based phylogeny of class IIa tRNA synthetases in relation to an unusual biochemistry. J. Mol. Evol. **53:**261–268.
47. **Ribas de Pouplana, L., M. Frugier, C. L. Quinn, and P. Schimmel.** 1996. Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria. Proc. Natl. Acad. Sci. USA **93:**166–170.
48. **Ribas de Pouplana, L., and P. Schimmel.** 2001. Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. Cell **104:**191–193.
49. **Russell, R. B., and G. J. Barton.** 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins Structure Function Genet. **14:**309–323.
50. **Schimmel, P., and L. Ribas de Pouplana.** 2000. Footprints of aminoacyl-tRNA synthetases are everywhere. Trends Biochem. Sci. **25:**207–209.
51. **Sokal, R. R., and C. D. Michener.** 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **28:**1409–1438.
52. **Srinivasan, G., C. James, and J. Krzycki.** 2002. Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science **296:**1459–1462.
53. **Stathopoulos, C., T. Li, R. Longman, U. Vothknecht, H. Becker, M. Ibba,**

and D. Söll. 2000. One polypeptide with two aminoacyl-tRNA synthetase activities. Science **287:**479–482.
54. **Tumbula, D., U. Vothknecht, H. Kim, M. Ibba, B. Min, T. Li, J. Pelaschier, C. Stathopoulos, H. Becker, and D. Söll.** 1999. Archaeal aminoacyl-tRNA synthesis: diversity replaces dogma. Genetics **152:**1269–1276.
55. **Weiner, A. M., and N. Maizels.** 1999. The genomic tag hypothesis: modern viruses as molecular fossils and ancient strategies for genomic replication, and clues regarding the origin of protein synthetsis. Biol. Bull. **196:**327–330.
56. **Wilcox, M.** 1969. Gamma-glutamyl phosphate attached to glutamine-specific tRNA. A precursor of glutaminyl-tRNA in *Bacillus subtilis*. Eur. J. Biochem. **11:**405–412.
57. **Woese, C. R., G. Olsen, M. Ibba, and D. Söll.** 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. **64:**202–236.
58. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.
59. **Woese, C. R.** 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA **95:**6854–6859.
60. **Woese, C. R.** 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. USA **97:**8392–8396.
61. **Woese, C. R.** 2002. On the evolution of cells. Proc. Natl. Acad. Sci. USA **99:**8742–8747.
62. **Wolf, Y., L. Aravind, N. Grishin, and E. Koonin.** 1999. Evolution of aminoacyl-tRNA synthetases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. **9:**689–710.