Molecular Evolution of Aminoacyl tRNA Synthetase Proteins in the Early History of Life

Author's Version: Original published in
Origins of Life and Evolution of Biospheres, DOI: 10.1007/s11084-011-9261-2
The final publication is available at springerlink.com

Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life

Gregory P. Fournier[1$], Cheryl P. Andam[2,3], Eric J. Alm[1], J. Peter Gogarten[3$]

[1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge MA 02139
[2] Department of Crop and Soil Sciences, Cornell University, Ithaca NY 14853
[3] Department of Molecular and Cell Biology, University of Connecticut, Storrs CT 06269

$: Corresponding authors:
J. Peter Gogarten,
Email: jpgogarten@gmail.com
Phone: 1 860 486 4061
FAX: 1 860 486 4331

Gregory P. Fournier,
Email: g4nier@mit.edu
Phone: 1 617 253 2726
FAX: 1 617 258 8850

ABSTRACT

Aminoacyl-tRNA synthetases (aaRS) consist of several families of functionally conserved proteins essential for translation and protein synthesis. Like nearly all components of the translation machinery, most aaRS families are universally distributed across cellular life, being inherited from the time of the Last Universal Common Ancestor (LUCA). However, unlike the rest of the translation machinery, aaRS have undergone numerous ancient horizontal gene transfers, with several independent events detected between domains, and some possibly involving lineages diverging before the time of LUCA. These transfers reveal the complexity of molecular evolution at this early time, and the chimeric nature of genomes within cells that gave rise to the major domains. Additionally, given the role of these protein families in defining the amino acids used for protein synthesis, sequence reconstruction of their pre-LUCA ancestors can reveal the evolutionary processes at work in the origin of the genetic code. In particular, sequence reconstructions of the paralog ancestors of isoleucyl- and valyl- RS provide strong empirical evidence that at least for this divergence, the genetic code did not co-evolve with the aaRSs; rather, both amino acids were already part of the genetic code before their cognate aaRSs diverged from their common ancestor. The implications of this observation for the early evolution of RNA-directed protein biosynthesis are discussed.

DOI: 10.1007/s11084-011-9261-2

Key words: aminoacyl-tRNA synthetase, paralog, genetic code, LUCA, horizontal gene transfer

Introduction

Aminoacyl-tRNA synthetases (aaRS), together with tRNA, define the alphabet and syntax of the genetic code, upon which all protein synthesis depends. Aminoacylation of all 20 universal amino acids is partitioned between two major aaRS folds (class I and class II). Within each class, the aaRSs with different amino acid specificity show distant shared ancestry as revealed by structural, sequence, and enzymatic similarity. However, while nearly all families of aaRS are universally distributed within the three domains (with the exception of some aaRS requiring tRNA-dependent amino acid synthesis, and domain-specific class I and class II LysRS families (Ibba et al. 1997), many, if not most of these groups have undergone horizontal gene transfer across all taxonomic levels. This complicates attempts to reconcile their phylogeny with organismal histories. Widespread and recurring gene transfers characterize the evolution of these enzymes, in contrast to the phylogenies of the rest of the translation machinery that reflect a history dominated by vertical inheritance. In fact, not only do the molecular phylogenies of the synthetases differ significantly from the organismal phylogeny, but the phylogenies of aaRS specific for each amino acid do not show consistency with one another (Woese et al. 2000). However, transfer of aaRS between divergent organisms is not rampant, transfer events are readily recognized, and the remainder of the phylogeny has retained a phylogenetic signal similar to that in ribosomal proteins and RNAs (Wolf et al. 1999; Andam et al, 2010).

The identity of the lineages involved in an HGT event can be easily determined if groups closely related to the donor are still living. However, in ancient transfers that have occurred before the divergence of major domains, or even prior to the time of LUCA, close relatives of donor lineages are usually extinct, making it difficult or even impossible to determine transfer scenarios. Genes that diverged this early also usually have an extremely long time of coalescence to these deep branches, reflected by very low sequence similarity between homologs, further complicating phylogenetic inference (Zhaxybayeva and Gogarten 2004).

LUCA and its sister lineages

Even though many gene families coalesce back to single ancestors within a phylogeny, one cannot assume that these coexisted within the same cellular ancestor. Coalescence simulations in which HGT events have been introduced demonstrate that the molecular ancestor of a particular gene does not necessarily reside in the organismal ancestor (Zhaxybayeva and Gogarten 2004). This logic can be extended to the time of LUCA and beyond, as the most recent cellular ancestor was almost certainly not the only existing cell at that time. It is much more likely that an entire community of primordial lineages interacted with each other and inhabited different niches, in the same way that organisms do now (Gogarten et al. 2008). The evolutionary processes and mechanisms that generated the diversity of the three domains after the time of the LUCA were likely also present, a conclusion following from the many shared ancestral characters found in all extant cells, such as a universal genetic code. Hence, it is reasonable to assume that these organisms also transferred genetic material with each other, allowing metabolic inventions to spread rapidly in the community.

Through HGT, these primordial lineages may have contributed directly to the genetic composition of LUCA's descendants. Such parallel lineages may even have persisted long enough to transfer genes to the ancestors of the extant domains, after the diversification following LUCA. In cases where the transfer occurred between an extremely ancient donor lineage and a more recently derived recipient, an intermediate carrier such as another cellular lineage or virus may have facilitated the horizontal transmission, "carrying it forward" in time. However, these secondary occurrences do not contradict the notion that deep branching lineages that are now extinct have significantly contributed to the genetic composition and phenotypic diversity within present-day genomes.

Transfers predating LUCA

Ancient transfers that originated in lineages likely to have diverged prior to the time of the LUCA have been identified through the analysis of atypical and rare forms of aaRS. An unusual form of seryl-tRNA synthetase (SerRS) exists that is distributed in a few extant archaeal methanogenic species (Andam and Gogarten 2011a). Phylogenetic reconstruction reveals that each type of SerRS forms a well-supported distinct clade (Figure 1). The rare form does not exhibit high sequence or structural similarity to other SerRS found in the majority of organisms within the three domains of life (Kim et al. 1998; Andam and Gogarten 2011a). Another significant difference between the two forms is their mechanism of substrate recognition (Korencic et al. 2004). Although both reveal some similarities in their mode of $tRNA^{Ser}$ recognition, there are remarkable differences in their identity requirements, such as the G1:C72 base pair and the number of unpaired nucleotides at the base of the variable stem that are both required by the rare SerRS (Korencic et al. 2004). Serine recognition of the rare form is also dependent on a zinc ion present in the active site, which is not found in the common SerRS (Bilokapic et al. 2006). Despite their differences, the rare and common forms perform identical functions, aminoacylating serine and ligating it to its cognate tRNA molecule. In rare instances, both forms are found in the same genome, as in the case of *Methanosarcina barkeri* (Korencic et al. 2004; Andam and Gogarten 2011a), which would suggest the acquisition of a second but divergent form of the enzyme by a genome already possessing an initial copy.

We observe a similar pattern in threonyl-tRNA synthetases (ThrRS). A common form and a divergent form both exist, with the latter only found in Archaea (Andam and Gogarten, 2011b). In both SerRS and ThrRS, the rare forms of these two enzymes diverged early from the common forms, with subsequent horizontal transfer from an unknown ancient lineage that probably has gone extinct or is undiscovered. This divergence event appears to have predated the LUCA node of the common forms of the two enzymes (Figure 1), although it is possible that higher substitution rates in the rare forms could produce long branch attraction artifacts to the same effect.

Another more striking example of an enzyme that arose prior to the organismal LUCA is pyrrolysyl-tRNA synthetase (PylRS), which charges $tRNA^{Pyl}$ with the non-canonical amino acid pyrrolysine (Pyl) (Srinivasan et al. 2002). This rare enzyme has a very restricted distribution, found only in members of the archaeal order Methanosarcinales, the firmicute *Desulfitobacterium hafniense* and a Deltaproteobacterium endosymbiont (Fournier et al. 2009).

In relation to the other aaRS, PylRS is placed as a deep-branching lineage within the subclass IIb, emerging prior to the LUCA of the bacterial and archaeal domains. The phylogenetic distribution of this enzyme suggests that these extant taxa acquired PylRS through several HGT episodes from an ancient, most likely extinct, lineage. Furthermore, Pyl has an extremely narrow functional role in a few functionally similar, unrelated, non-essential proteins (corrinoid dependent mono-, di-, and trimethylamine methyltransferases), which are just as narrowly distributed (Krzycki 2004). It seems unlikely that this limited functionality and distribution would be sufficient for the selection and evolution of a unique system devoted entirely to the synthesis and aminoacylation of a novel amino acid. The more likely scenario is the existence of an ancient donor lineage that used Pyl far more extensively in the past (Fournier et al. 2009).

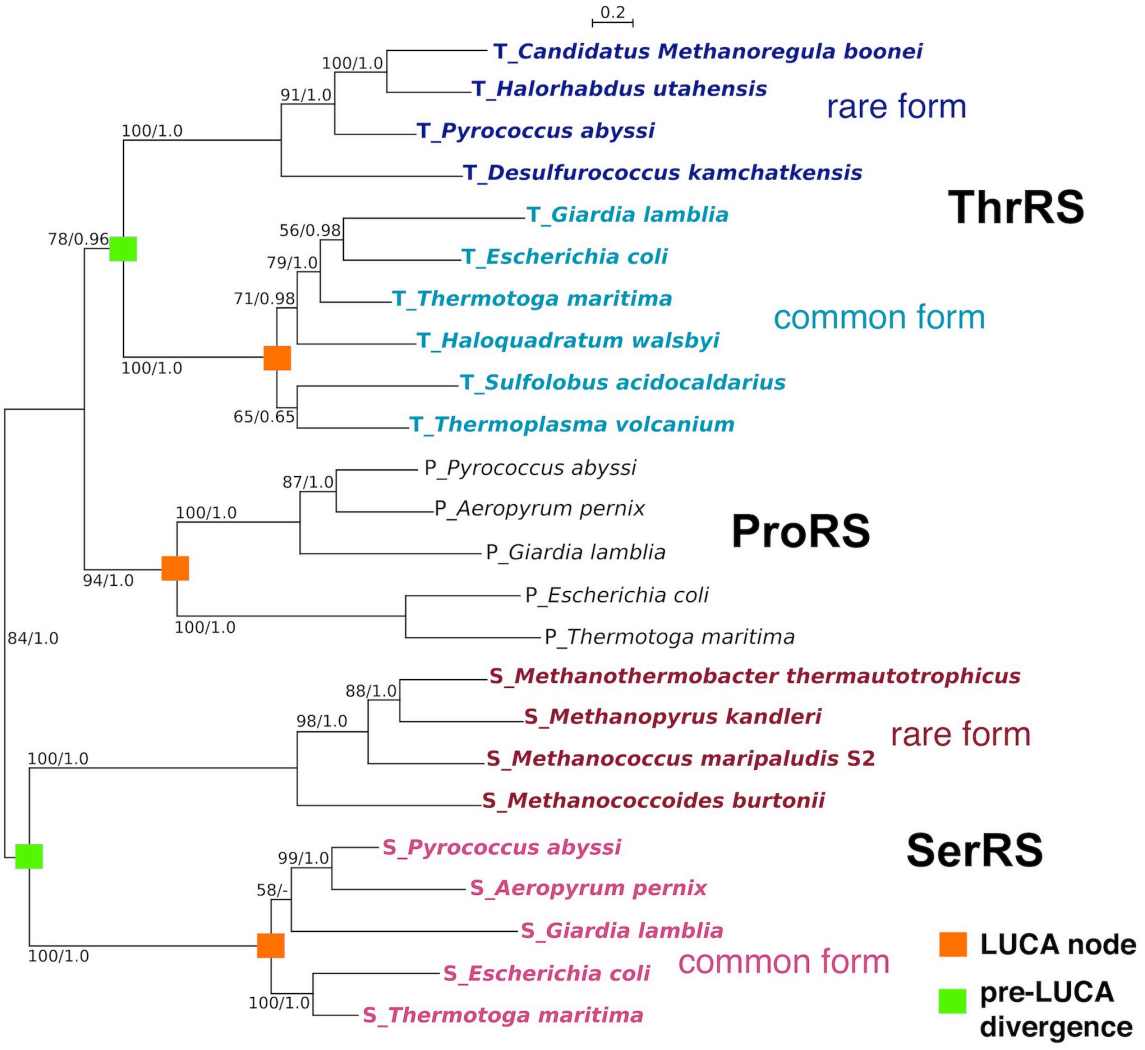

Figure 1. Phylogenetic tree of representatives of seryl-, threonyl-, and prolyl-tRNA synthases. The tree shows the divergence events that gave rise to the common and rare forms of seryl- and prolyl-tRNA synthases. Green squares represent ancient divergence events and orange squares represent nodes that have descendants in all three domains of life, and that therefore can be identifies as LUCA nodes. Adapted and modified from (Andam and Gogarten 2011b).

These transfers suggest that novel mechanisms of adaptation and selection are at work in seemingly neutral events. Conveniently, as the ancestral reconstructions of gene families are independent of organismal history, these reticulations have little effect on inferring deep ancestral states of synthetase families. In fact, reconstruction of homeoallele ancestors may even elucidate the selective forces at work behind their divergence (Andam et al. 2011). Homeoalleles are divergent enzymes of identical functions, which are transferred within higher-level exchange groups, such as class or phylum (Andam et al. 2010; Andam and Gogarten 2011a). The groups are formed through recurrent transfers among closely related organisms, generating increasing genetic similarity among the members of the group. One complication arising from such extensive transfer is an inability to evaluate the phylogenetic topology used for reconstruction: is a branch at conflict with accepted organismal groups a reconstruction artifact, or simply another transfer event? Is the agreement between gene and organismal phylogeny indeed a reflection of inheritance from a common ancestor, or does it reflect a gene transfer bias in favor of close relatives? While it has been proposed that phylogenetic uncertainty has little effect upon ancestral reconstructions (Hanson-Smith et al. 2010), it remains to be shown if this robustness is also evident in the probability distributions within sites, and not just for the ML sequence. Presumably, extensive surveys of transfers within synthetase families can identify consistent patterns of transfer (Andam and Gogarten 2011a), and may help to resolve such uncertainties.

Aminoacylation and the Genetic Code

Predating the time of LUCA, the relationship between synthetase evolution, the RNA world, and the origin of the genetic code is complex, and its understanding requires the synthesis of numerous interrelated chemical, structural, semantic, and logical narratives. Given the near-universal 1:1 relationship between synthetases and cognate amino acids, one elegant hypothesis among these narratives is that the divergence of synthetase families drove the expansion of the genetic code to include all 20 canonical amino acids, and that different nodes along the synthetase phylogeny correspond to distinct times in protein evolution where fewer than 20 amino acids were distinctly coded.

This coevolutionary model for aminoacyl-tRNA synthetases (aaRS) is intuitively satisfying, as, in many cases, similar amino acids are recognized by closely related aaRS. This fits the typical narrative of enzyme evolution: duplications of genes encoding a protein product gain different, yet similar functions, preserving and expanding upon the original functionality, increasing the fitness of the system. While it goes without saying that the genetic code also requires the synthesis of amino acids, in many cases these pathways are not solely dedicated to making the building blocks of proteins. Amino acid biosynthesis pathways are all closely tied to central metabolism, and frequently their products are an intrinsic part of this metabolism. For example, glutamate and glutamine are essential components in the GS-GOGAT system for nitrogen assimilation (Suzuki and Knaff 2005), serine and glycine are essential in single-carbon metabolism (Newman & Magasanik, 1963), cysteine and methionine are central to sulfur uptake and metabolism (Wirtz and Droux 2005); many amino acids are also involved in small molecule/cofactor synthesis (Fischer et al. 2010; White, 2001) and even RNA modification (Grosjean and Benne 1998; Ikeuchi et al. 2010; Suzuki and Miyauchi 2010). Therefore, we can

reasonably separate the question of the functional chemical origin of many (if not most) amino acids from their incorporation in proteins via the semantics of a genetic code. The same cannot be said for aaRS; encoded polypeptide biosynthesis and tRNA almost certainly pre-existed, with the earliest protein synthesis emerging in a world where functions of modern proteins were already being performed by another system (Xiao and Yu 2007; Knight et al. 1999). This could reflect a stage of the RNA world where simple polypeptides or aminoacylated RNAs were augmenting the capacities of ribozymes (Schimmel 2008; Erives 2011), or a more advanced state wherein proteins were already functionally dominant, with some functions still performed by relic RNA systems, aminoacylation among them.

Without such a primordial non-protein synthetase regime, taking the coevolutionary model to its logical extreme would be absurd: there would be two singular "Ur-aaRS", one from each class, made only from their shared cognate amino acids. From a purely protein biochemical perspective, this is almost certainly impossible, and, even if possible, reduces to a "chicken and egg" problem between the two classes. This would also contradict the existence of the conserved sites within each aaRS class necessary for recognition and aminoacylation: the HIGH and KMSKS motifs within class I, and the GRASP motif within class II. While these motifs show some flexibility, several different amino acids must still exist, in either case. Finally, in both cases the synthetase classes are part of greater protein fold domains which must have originated at an even earlier time (Rossmann fold-domain proteins for class I aaRS, ATP-grasp domain proteins for class II aaRS). Therefore, the question becomes, *at what point did protein synthetases take over the enforcement of the genetic code, and, after this point, how did their divergence expand amino acid usage to its current state?*

Phylogenetic analysis supports that, with few exceptions, each synthetase family was present at the time of the LUCA, and inherited by each of the 3 major domains of life (Nagel and Doolittle 1995). Consequently, within each synthetase class numerous divergence events happened before the LUCA, each one a point at which the genetic code may have increased in complexity, either adding a new amino acid (neofunctionalization), or resolving a previously ambiguous specificity into two distinct amino acids (subfunctionalization). Reconstructing the sequences of synthetase protein families at each of these deep ancestral nodes can allow the inference of ancestral functional states, and therefore potentially reveal pre-LUCA states of the genetic code, as well as the mechanism of its evolution.

Sequence Resurrection

Using multiple sequence alignments and their inferred phylogenies, together with models of protein evolution, it is possible to infer the state of each alignment site at any internal node within the tree, as a set of probabilities across all 20 amino acids. From this, a maximally likely ancestral sequence can be identified and synthesized, in order to experimentally test the properties of the ancestral protein (reviewed by Benner et al. 2007). While this is the most straightforward approach for inferring ancestral function, the uncertainty in the reconstruction increases dramatically as one moves backwards in time, especially to paralog ancestors before LUCA, where functional divergence has also played a role. At these early states, it is very unlikely that the ML sequence is the "true" sequence. Rather, there is a very large population of

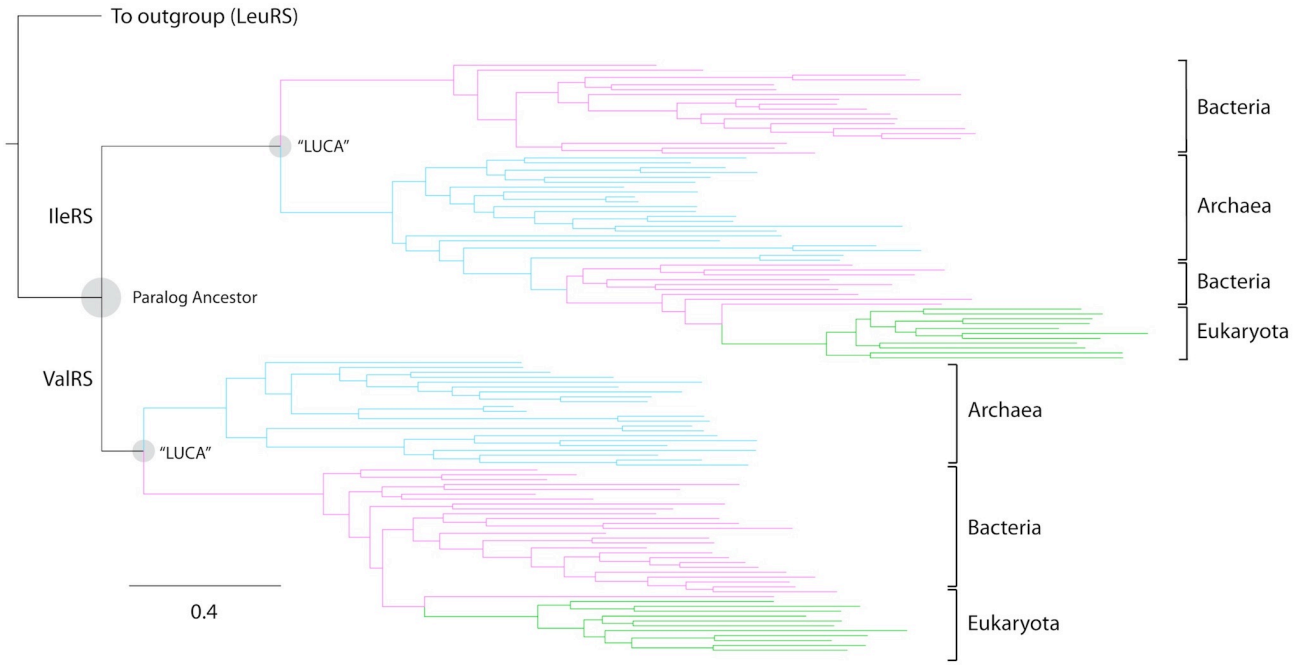

Figure 2: Rooted phylogeny of aliphatic aminoacyl-tRNA synthetases. IleRS and ValRS are sister paralogs, with LeuRS (not shown) included as outgroup. Domains within each paralog (colored) show differing topologies due to deep horizontal gene transfer events

possible ancestors that must be synthesized and assessed. The true ancestral function can then be inferred by correlating candidate sequence function with sequence likelihood. Estimates based on preliminary ancestral sequence reconstructions suggest $4x10^6$ likely variants for an aaRS amino acid-binding pocket alone, which consists of only 10-20 sites.

Compositional Analysis

Until high-throughput processes are adopted to experimentally test the large ancestor space of synthetase reconstructions, compositional methods provide a reasonable alternative, using logical inference to predict ancestral functionality.

If divergences giving rise to synthetase families coincide with genetic code expansion events, the composition of the paralog ancestor at each node gives clues to the lexicon of the genetic code before the divergence occurred. For example, if the paralog ancestor of IleRS and ValRS encoded a synthetase with Val as its cognate amino acid, and this state reflects a simpler stage in the genetic code, then there should be no Ile in its reconstructed sequence. Reversing this logic, an observed absence of Val in the reconstruction implies IleRS is the ancestral functional state, while conversely, an absence of Ile implies ValRS is the ancestral functional state. However, other possibilities are possible besides neofunctionalization. If sites reconstructing for either Ile or Val always show co-occurring ambiguity (e.g., 45% Ile, 55% Val), then this ambiguity may in itself have been the ancestral state, implying an ambivalent coding that prevented purifying selection from fixing sites for either Ile or Val, up until divergence partitioned their sequence space (subfunctionalization). The third alternative is that reconstructed ancestors contain distinct

sites for both Ile and Val, implying that both were specifically part of the code at this time, and, therefore, must have been aminoacylated by an alternative, more ancient system. In such a case, these functions would have been taken over by aaRS following their divergence (parafunctionalization), with the functions of paralog ancestors being either redundant or complementary to the primordial system.

Results

Cognate synthetases for aliphatic amino acids are all found within the same synthetase super-family, with IleRS and ValRS as sister paralogs, and LeuRS as the outgroup (Figure 2). As these amino acids are functionally and structurally very similar, and even occupy the same column of the genetic code (NUN), it has been proposed that their cognate synthetases were indiscriminate in their aminoacylation (Cusack et al. 2000). However, ancestral reconstruction of the IleRS/ValRS paralog ancestor shows several high-probability sites for both Ile and Val within the ML sequence, with far fewer sites showing ambiguity between these two amino acids (Figure 3). Out of 421 reconstructed sites, 90 show a supermajority consensus (>80%) for aliphatic amino acids (Ile, Val, or Leu). Of these sites, 22 have a high relative probability for Ile, and 28 have a high relative probability of Val, in each case, similar to the number for Leu (24), the cognate amino acid to the outgroup LeuRS protein family which, under a coevolutionary model, would already have diverged at this point and be expected to be part of the code. Additionally, only 9 sites were identified as being ambiguous between Ile and Val. These observations strongly support the parafunctionalization model for IleRS and ValRS. Furthermore, mapping these sites to published structures of ValRS (Fukai et al. 2003) shows that Ile, Val, and ambiguous sites occur across all regions of the protein, demonstrating that this signal is unlikely confounded by the relative timing of protein evolution (e.g., an older catalytic core containing only Val, with an anticodon recognition domain containing both Ile and Val, would suggest early neofunctionalization with later addition of a protein domain).

Reconstructions of simulated alignments further support parafunctionalization. Simulations show similar overall counts for aliphatic sites (supermajority average site count = 78.1, SD = 5.97), as well as similar rates for highly specific and ambiguous sites across Ile, Val, and Leu (Figure 3). The only statistically significant differences observed were for Ile and ambiguous Ile/Val/Leu site categories within the simulated majority analysis, in comparison to simulated supermajority and data majority analyses. These differences are likely due to the lack of site rate context within simulated data, resulting in a more rapid decrease in absolute probabilites of amino acids at sites with lower *p*(aliphatic). In comparison, within actual data aliphatic sites are likely to correspond to buried positions within the protein, that evolve at a slower rate and are more physiochemically constrained. As such, *p*(aliphatic) should be less correlated with sequence ambiguity across sites. This supports that observed composition biases at the reconstructed ancestor fit under the expected model of sequence evolution (indicating parafunctionalization), and are not due to historical bias at the root. Additionally, this shows that the excess of positions ambiguous for Ile/Val compared to those for Ile/Leu and Leu/Val is observed in both simulated and real reconstructions. As such, it is likely due to substitution model and/or frequency effects, and is not indicative that actual ambiguity of within the genetic code being more likely for this amino acid pair at the time of the paralog ancestor.

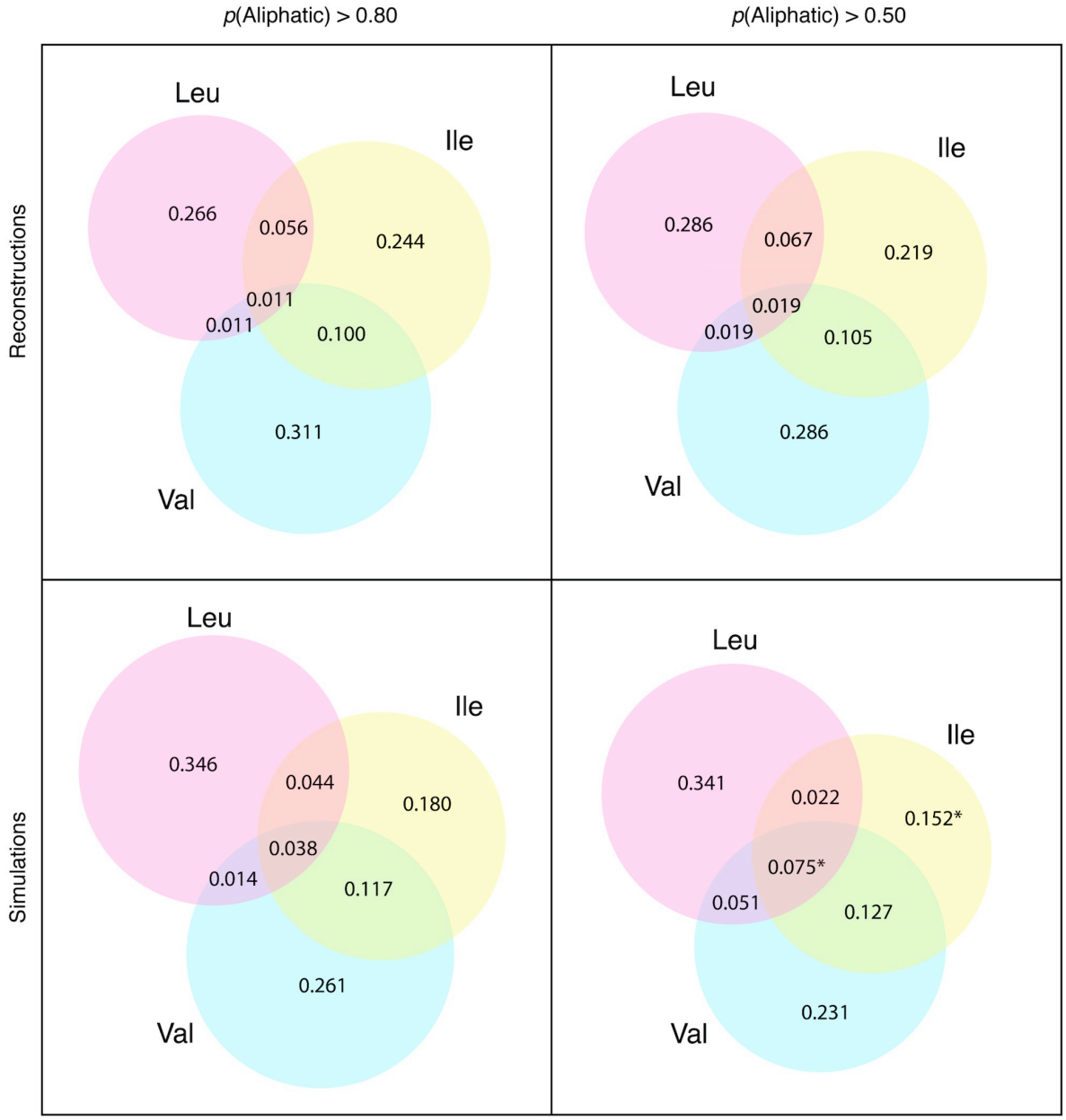

Figure 3: Aliphatic amino acid sites within IleRS-ValRS paralog ancestors. Venn diagrams represent the proportion of different site identities at reconstructed positions within data and simulations. Partitioned areas represent positions that have <20% relative probability for the excluded sets, thus defining highly resolved and ambiguous sites in all combinations. *p*(Aliphatic) refers to the cutoff for inclusion for sites with a majority (>0.50) or supermajority (>0.80) probability of containing an aliphatic amino acid (*p*(Ile)+*p*(Val)+*p*(Leu)). "*" denotes categories showing a statistically significant difference from other analyses. See text for further discussion.

These results show that at least a subset of amino acids pre-date their cognate aaRS, requiring that alternative mechanisms of aminoacylation existed before and during protein synthetase evolution. This is suggestive of a protracted "RNA-peptide world" which only

gradually became dominated by proteins. This model is also supported by structural insights, as the hydrolase domain uniquely shared by IleRS, ValRS, and LeuRS (Cusack et al. 2000) also reconstructs back to these paralog ancestors via vertical inheritance, and contains sites specific for Ile, Leu, and Val. Therefore, editing to prevent mis-aminoacylation was necessary within these enzymes before the divergence of each paralog, implying once again that these cognate amino acids preceded their cognate synthetases.

Methods

Tree reconstruction

Multiple sequence alignments of IleRS, ValRS, and LeuRS homologs from 60 representative genomes sampled across Eukaryota, Bacteria and Archaea were each generated independently, and then profile-aligned using MUSCLE (Edgar 2004). Phylogenetic trees were generated using PhyML (Guindon and Gascuel 2003) (WAG model, estimated proportion of invariant sites, 4 rate categories, estimated a, SPR tree topology search).

Given the observed re-ordering of the ZN-1 and hydrolase domains in bacterial LeuRS (Cusack et al. 2000), it was hypothesized that either of these domains may have been displaced or later acquired via horizontal gene transfer, and that these regions may be associated with a different phylogenetic topology. To test this, phylogenetic trees of the sub-alignments consisting of the ZN-1 domain, hydrolase domain, and remainder of the protein were constructed across all paralogs using PhyML (WAG model, estimated proportion of invariant sites, 4 rate categories, estimated alpha, SPR tree topology search). It was determined that the hydrolase domain shows the same phylogeny as the rest of the protein, while the ZN-1 domain showed bacterial LeuRS grouping with bacterial ValRS, suggesting domain shuffling and non-orthologous displacement occurring early in the history of the bacterial domain. Therefore, because of its uncertain phylogenetic history, the ZN-1 domain was excluded from ancestral reconstruction.

Simulations

Simulated sequences (ten replicates) with a length of 421 sites (equivalent to the number of homologous positions used in the ancestral reconstruction), were generated using the EVOLVER program in PAML (Yang 2007), using the above-described tree topology (i.e., constructed with sites excluding the ZN-1 domain) under a WAG model with 4 rate categories, and same rate parameter a=1.23 estimated under PhyML.

Ancestral Reconstructions

Using the actual and simulated multiple sequence alignments with their shared phylogenetic tree, the Bio++ v..2.0.2 software package (Dutheil et al. 2006) was used to generate ancestral reconstructions of all internal nodes under a homogeneous model (WAG model, 4 rate categories, a=1.23, observed amino acid frequencies, all other settings default). Sites within the reconstructed IleRS/ValRS paralog ancestor were counted as “aliphatic” based on the sum probability of aliphatic amino acids (p(Ile)+p(Val)+p(Leu)). Analyses were performed using

two separate cutoffs for including sites, with a majority (>0.50) or supermajority (>0.80) probability, in order to determine if a significant excess of ambiguous positions was only introduced at sites with lower absolute probability. Relative usage of Ile, Val, and Leu was then determined within these sites, with categories for ambiguity defined by the exclusion of amino acids with <0.200 relative probability. For example, reconstruction of the paralog ancestor site V657 has a p(Aliphatic)=0.830. The relative probabilities within this site (p(Ile)=0.362, p(Leu)=0.131, p(Val)=0.507) therefore categorize it as ambiguous for Ile/Val, as p(Leu) is below the probability threshold for inclusion. Normalized counts within categories were compared between datasets and simulation averages using a two-sided, one-sample t-test.

Conclusions

Aminoacyl-tRNA synthetases have unique phylogenetic histories and functional roles that provide key insights into the early evolution of life on Earth. Ancient horizontal transfers of these genes reveal important details about the complex nature of the LUCA, and the evolutionary processes leading to the diversity observed within the extant domains of life. Additionally, sequence reconstructions of synthetase paralog ancestors before the time of LUCA provide strong empirical evidence that parts of the genetic code predate the divergence of these proteins. In the future, expanding this analysis to other synthetase families will produce a more comprehensive picture of the evolutionary mechanisms behind the evolution of the genetic code, the demise of the RNA world, and the nature of life at the time of the last universal cellular ancestor.

Acknowledgments

This work was supported by National Science Foundation Grants DEB 0830024 to JPG and DEB 0936234 to EJA. GPF is recipient of a NASA Postdoctoral Fellowship at the Massachusetts Institute of Technology.

References

Andam CP, Gogarten JP (2011a) Biased gene transfer in microbial evolution. Nat Rev Microbiol 9(7): 543-555

Andam CP, Gogarten JP (2011b) Biased gene transfer and its implications for the concept of lineage. Biol Direct 23:47

Andam CP, Williams D, Gogarten JP (2010) Biased gene transfer mimics patterns created through shared ancestry. Proc Natl Acad Sci USA 107:10679-10684

Andam, CP, Fournier, GP and Gogarten, JP (2011). Multilevel populations and the evolution of antibiotic resistance through horizontal gene transfer. FEMS microbiology reviews **35**(5): 756-767.

Benner SA, Sassi SO, Gaucher EA (2007) Molecular paleoscience: systems biology from the past. Adv Enzymol Relat Areas Mol Biol. 75:1-132.

Bilokapic S, Maier T, Ahel D, Gruic-Sovulj I, Söll D, Weygand-Durasevic I, Ban N. (2006) Structure of the unusual seryl-tRNA synthetase reveals a distinct zinc-dependent mode of substrate recognition. EMBO J 25: 2498-2509

Cusack S, Yaremchuk A, Tukalo M (2000) The 2 Å crystal structure of leucyl-tRNA synthetase and its complex with a leucyl-adenylate analogue. EMBO J. 19(10):2351-61.

Dutheil J, Gaillard S, Bazin E, Glemin S, Ranwez V, Galtier N, Belkhir K (2006) Bio++: a set of C++ libraries for sequence analysis phylogenies, molecular evolution and population genetics. BMC Bioinformatics 7:188.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

Erives A (2011) A Model of Proto-Anti-Codon RNA Enzymes Requiring L:-Amino Acid Homochirality. J Mol Evol. Jul 22. [Epub ahead of print]

Fischer JD, Holliday GL, Rahman SA, Thornton JM (2010) The structures and physiochemical properties of organic cofactors in biocatalysis. J Mol Biol. 403(5):803-24.

Fournier GP, Huang J, Gogarten JP (2009) Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. Philos Trans R Soc Lond B Biol Sci 364:2229-2239

Fukai S, Nureki O, Sekine SI, Shimada A, Vassylyev DG, Yokoyama S (2003) Mechanism of molecular interactions for tRNA(Val) recognition by valyl-tRNA synthetase. RNA. 9:100-111.

Gogarten, JP, Fournier, GP and Zhaxybayeva, O (2008). Gene Transfer and the Reconstruction of Life's Early History from Genomic Data. Space Science Reviews **135**: 115-131.

Grosjean H, Benne R (1998) Modification and Editing of RNA. ASM press, Washington DC, USA.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Nucleic Acids Res. 33:W557-9.

Hanson-Smith V, Kolaczkowski B, Thornton JW (2010) Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol Biol Evol. 27(9):1988-99.

Huang J, Xu Y, Gogarten JP (2005) The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. Mol Biol Evol. 22(11):2142-6.

Ibba M, Celic I, Curnow A, Kim H, Pelaschier J, Tumbula D, Vothknecht U, Woese C, Söll D (1997) Aminoacyl-tRNA synthesis in Archaea. Nucleic Acids Symp Ser. (37):305-6.

Ikeuchi Y, Kimura S, Numata T, Nakamura D, Yokogawa T, Ogata T, Wada T, Suzuki T, Suzuki T (2010) Agmatine-conjungated cytidine in a tRNA anticodon is essential for AUA decoding in archaea. Nat Chem Biol. 6(4):277-82.

Kim HS, Vothknecht UC, Hedderich R, Celic I, Söll D (1998) Sequence divergence of seryl-tRNA synthetases in archaea. J Bacteriol 180: 6446-6449

Knight RD, Freeland SJ, Landweber LF (1999) Selection, history and chemistry: the three faces of the genetic code. Trends Biochem Sci. 24(6):241-7.

Korencic D, Polycarpo C, Weygand-Durasevic I, Söll D (2004) Differential modes of transfer RNASer recognition in *Methanosarcina barkeri*. J Biol Chem 279: 48780-48786

Krzycki JA (2004) Function of genetically encoded pyrrolysine in corrinoid-dependent methylamine methyltransferases. Curr Opin Chem Biol. 8(5):484-91.

Nagel GM, Doolittle RF (1995) Phylogenetic analysis of the aminoacyl-tRNA synthetases. J Mol Evol. 40(5):487-98.

Newman EB, Magasanik B (1963) The relation of serine-glycine metabolism to the formation of single-carbon units. Biochim Biophys Acta. 78:437-48.

Schimmel P (2008) Development of tRNA synthetases and connection to genetic code and disease. Protein Sci 17:1643-1652

Srinivasan G, James CM, Krzycki JA (2002) Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science 296:1459-1462

Suzuki A, Knaff DB (2005) Glutamate synthetase: structural, mechanistic and regulatory properties, and role in the amino acid metabolism. Photosynth Res. 83(2):191-217.

Suzuki T, Miyauchi K (2010) Discover and characterization of tRNAIle lysidine synthetase (TilS). FEBS Lett. 584(2):272-7.

White RH (2001) Biosynthesis of the methanogenic cofactors. Vitam Horm. 61:299-337.

Wirtz M, Droux M (2005) Synthesis of the sulfur amino acids; cysteine and methionine. Photosynth Res. 86(3):345-62.

Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Molec Biol Rev 64:202-236

Wolf, YI, Aravind, L, Grishin, NV and Koonin, EV (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.

Xiao JF, Yu J (2007) A scenario on the stepwise evolution of the genetic code. Genomics Proteomics Bioinformatics. 5(3-4):143-51.

Yang Z (2007) PAML 4:phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586-91.

Zhaxybayeva O, Gogarten JP (2004) Cladogenesis, coalescence and the evolution of the three domains of life. Trends Gen 20:182-187