Common evolutionary origins of the bacterial glycyl tRNA synthetase and alanyl tRNA synthetase

Claudia Alvarez-Carreño[1,2], Marcelino Arciniega[3], Lluís Ribas de Pouplana[4], Anton S. Petrov[1,2], Adriana Hernández-González[3], Marco Igor Valencia-Sánchez[3]*, Loren Dean Williams[1,2,]*, and Alfredo Torres-Larios[3]

[1]NASA Center for the Origin of Life, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA

[2]School of Chemistry and Biochemistry, Georgia Institute of Technology, 901 Atlantic Drive, Atlanta, GA 30332, USA

[3]Departamento de Bioquímica y Biología Estructural, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Circuito Exterior s/n, Ciudad Universitaria, Apartado postal 70-243, Mexico City, 04510, México.

[4]Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia 08010, Spain and 6 University Pompeu Fabra, Barcelona, Catalonia 08003, Spain.

* Corresponding authors:
Marco Igor Valencia-Sánchez
Email: igorval@comunidad.unam.mx
Loren Dean Williams
Email: loren.williams@chemistry.gatech.edu

Abstract

Glycyl tRNA aminoacylation is catalyzed by different enzymes in different lineages of the Tree of Life. The archaeal (arcGlyRS) and bacterial (bacGlyRS) versions of glycyl tRNA synthetase are globally different. arcGlyRS is a relatively small protein with two domains: a catalytic domain and a tRNA recognition domain. bacGlyRS is a large, multi-domain protein. We have investigated the complex evolutionary histories of these enzymes. Component domains of bacGlyRS show common ancestry with 1) the catalytic domain of class II tRNA synthetases; 2) the HD domain of the bacterial RNase Y; 3) the body and tail domains of the archaeal CCA-adding enzyme; 5) the anti-codon binding domain of the arginyl tRNA synthetase; and a previously unrecognized domain that we call ATL (Ancient tRNA linkup). The emergence of bacGlyRS in its extant form would have required the recruitment of domains from a pool of existing homologs, possibly at different times, by a process of molecular bricolage. A domain-by-domain analysis reveals that the catalytic and ATL domains are closely related to domains in the universal alanyl tRNA synthetase. Homology of the HD domain to bacteria-specific proteins and homology of the body and tail domains to the archaeal CCA-adding enzyme suggest a later recruitment. The data suggest that bacGlyRS was present at the last bacterial common ancestor and that bacGlyRS and uniAlaRS emerged from a shared core before the last common universal ancestor, suggesting that bacGlyRS and uniAlaRS evolved by a stepwise accumulation of protein domains during the dawn of biology.

Introduction

Translation of mRNAs into coded polypeptides is universal to life (Bowman, et al. 2020). mRNA decoding depends on the ribosome and a group of enzymes called aminoacyl tRNA synthetases (aaRSs) (Burbaum and Schimmel 1991; Woese, et al. 2000; Gomez and Ibba 2020; Giegé and Eriani 2023). aaRSs determine the genetic code by recognizing specific tRNAs and linking them to their cognate amino acids. aaRSs fall into two classes. The catalytic domain of Class I aaRSs is a Rossmann fold while the catalytic domain of Class II aaRSs is an α/β three-layered sandwich. The early evolution of aaRSs has been linked to the origins of the genetic code and to the emergence of the first proteins (Hartman 1995; Ribas de Pouplana and Schimmel 2001; Fournier, et al. 2011; Koonin and Novozhilov 2017; Ribas de Pouplana 2020). Class I and Class II catalytic domains appear to have independent origins in the deep evolutionary past (Eriani, et al. 1990; Moras 1992; Gomez and Ibba 2020).

Over the phylogenetic tree, a given tRNA is typically aminoacylated by a single type of aaRS. However, there are exceptions to this rule. There are two distinct LysRSs (Terada, et al. 2002). The class II LysRS is more abundant in bacteria, while class I LysRS is found predominantly in archaea (Tumbula, et al. 1999).

There are two distinct GlyRSs, which are the subject of this report. An archaeal GlyRS (arcGlyRS) is found in archaea, eukarya, and some bacteria. A bacterial GlyRS (bacGlyRS) is found only in bacteria (Ostrem and Berg 1974; Tang and Huang 2005; Han, et al. 2023). The bacterial and archaeal versions of GlyRS are globally different except for their catalytic domains (Logan, et al. 1995). bacGlyRS and arcGlyRS use different strategies to recognize $tRNA^{Gly}$. Why are there two GlyRSs? How did they originate? Which GlyRS appeared first? Why is the distribution of the two GlyRSs over the phylogenetic tree so patchy?

The bricolage hypothesis suggests that complex proteins can arise and evolve through exaptation of pre-existing domains or modules, which can be combined to produce new functions (Jacob 1977). In this hypothesis, domains can be thought of as evolutionary building blocks that provide information on ancestry. arcGlyRS is relatively small, with a catalytic domain and a tRNA recognition domain. bacGlyRS is larger and more complex. We and others (Dimas-Torres, et al. 2021; Han, et al. 2023) have hypothesized that the component domains of bacGlyRS were serially exapted through bricolage.

If the bricolage hypothesis is correct, homologies of individual domains can be used to help reconstruct the complex histories of multidomain proteins such as bacGlyRS and can help us understand ancestral relationships. Accurate histories require accurate domain delineations. A domain is an independent and stable three-dimensional structure (Rossman and Liljas 1974; Levitt and Chothia 1976; Porter and Rose 2012) and is a common unit of evolution (Ponting, et al. 2000; Ohno 2013). Domains are encoded by genes. Duplicated genes, called paralogs (Fitch 1970), can combine to form a variety of distinct multi-domain proteins.

bacGlyRS contains α and β subunits. The combined data are interpreted here to indicate that there are six component domains in the β subunit of bacGlyRS rather than five domains as proposed previously (Ju, et al. 2021). The domain of bacGlyRS that escaped detection in previous work is extremely informative here about bacGlyRS ancestry, and shows homology to a domain found only in uniAlaRS, a class II aaRS that is universal to all organisms. Sequence and structural comparisons suggest that bacGlyRS is more closely related to uniAlaRS than to arcGlyRS (Smith and Hartman 2015; Valencia-Sánchez, et al. 2016; Dimas-Torres, et al. 2021).

Here, we show that bacGlyRS domains are homologous to universal, bacteria-specific or archaea-specific proteins. The results suggest that bacGlyRS evolved by a stepwise bricolage. Six component domains of bacGlyRS show common ancestry with domains of four proteins that interact with RNAs other than $tRNA^{Gly}$. Component domains of bacGlyRS show sequence similarities with 1) the catalytic domain of uniAlaRS; 2) the ATL domain (see below) of uniAlaRS; 3) a domain of uniArgRS (a universal class I aaRS); 4) the body and tail domains of the archaeal CCA-adding enzyme; and 5) a domain of the bacterial RNase Y (in agreement with previous observations of structural similarity (Ju, et al. 2021)). This conservation of RNA as substrate throughout a series of exaptation steps suggests a complex but accessible mechanism by which RNA-associated proteins are repurposed for new RNA-associated functions.

Results

bacGlyRS and arcGlyRS have a complex evolutionary history.

The phylogenetic partitioning of bacGlyRS and arcGlyRS appears irregular and discontinuous ('patchy') over the extant twigs of the bacterial tree. We examined a broad range of archaeal and bacterial proteomes, some of which are only partial (Figure 1). Many bacterial phyla contain multiple branches of either bacGlyRS or arcGlyRS. Firmicutes, Actinobacteria, Chloroflexi, Proteobacteria and DST, PVC, and FCB groups each contain species with bacGlyRS and species with arcGlyRS. Bacteria containing both bacGlyRS and arcGlyRS are extremely rare. Of the 542 sampled proteomes, arcGlyRS (GenBank: PSL06957.1) and bacGlyRS (GenBank: PSL02860.1) are found together only once, in the bacterium *Haloactinopolyspora alba* (NCBI RefSeq assembly: GCA_003014555.1) isolated from the Dead Sea sediment. Candidate Phyla Radiation (CPR) uniformly contains arcGlyRS; Cyanobacteria uniformly contains bacGlyRS. We do not currently have a compelling model that explains the global or local distribution of bacGlyRS and arcRNA over the bacterial tree.

bacGlyRS sequences are present in early diverging bacteria

The phylogenetic distribution of bacGlyRS suggests that it was present at the Last Bacterial Common Ancestor (LBCA). bacGlyRS is found in early diverging bacterial phyla (Moody, et al. 2022b) such as Firmicutes, Acidobacteria, and Proteobacteria (Figure 1). bacGlyRS is not found in the Candidate Phyla Radiation phylum. The CPR phylum contains putative symbiotic and parasitic bacteria with reduced genomes, large gene losses and extensive lateral gene transfer (Achsel, et al. 2001). The presence of arcGlyRS in CPR proteomes may be due to horizontal gene transfer and not vertical inheritance.

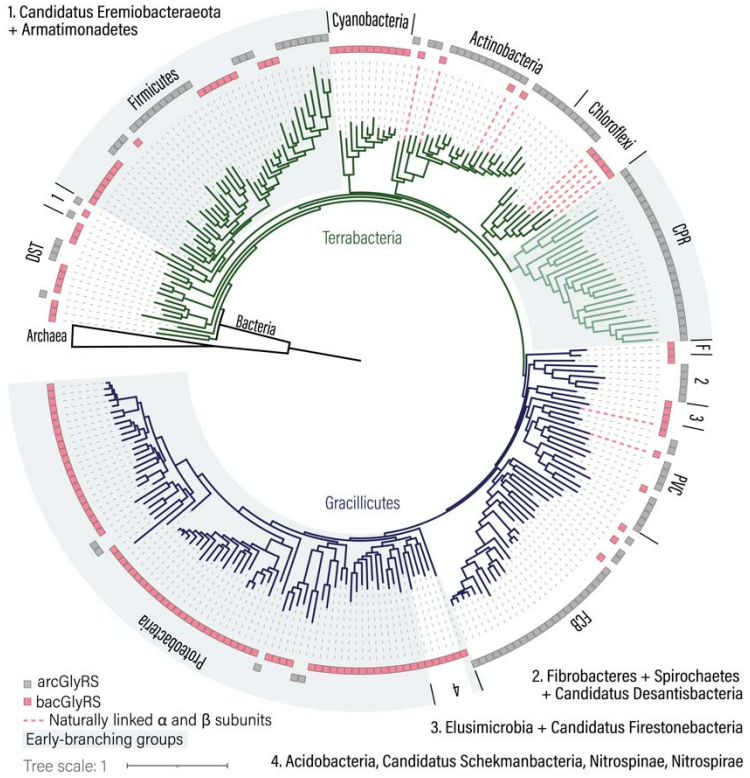

Figure 1. Phylogenetic distribution of the two versions of Glycyl tRNA ligase. The presence of bacGlyRS and arcGlyRS is mapped into an Archaea-Bacteria phylogeny. Pink dotted lines: bacGlyRS with naturally linked α and β subunits form heterodimers instead of heterotetramers. The topology of the tree was obtained from (Moody, et al. 2022a). DST: Deinococcus-Thermus, Synergistetes, Thermotogae, Caldiserica, Coprothermobacterota; CPR: Candidatus Phyla Radiation; PVC: Planctomycetes, Verrucomicrobia, Chlamydiae, Kiritimatiellaeota, Lentisphaerae, Candidatus Desantisbacteria, Candidatus Omnitrophica; FCB: Fibrobacteres, Chlorobi, Bacteroidetes, Gemmatimonadetes, Candidatus Cloacimonetes, Candidatus Fermentibacteria, Candidatus Glassbacteria; F. Fusobacteria.

Protein domains in bacGlyRS are related to RNA-modifying enzymes

Each domain of bacGlyRS has an independent evolutionary history. Six domains in bacGlyRS display strong sequence similarity to domains of other proteins (Figure 2). We determined which proteins are the closest sequence matches to each of the six domains of bacGlyRS (Figure 2D). The catalytic domain shows sequence similarity to the catalytic domain of uniAlaRS (e-value 1.1×10^{-07}). The body and tail domains show sequence similarity to domains of the archaeal CCA-adding enzyme (e-value 1.4×10^{-08}). The HD domain shows sequence similarity to a domain of ribonuclease Y (RNase Y) (e-value 5.7×10^{-08}), and the anti-codon binding domain (ABD) shows sequence similarity to a domain of uniArgRS (e-value 2.9×10^{-17}).

We identified a previously unrecognized domain in bacGlyRS that we call Ancient tRNA linkup domain (ATL) domain. Our characterization of ATL as a distinct domain is consistent with

89 the previous identification of this domain in uniAlaRS (Naganuma, et al. 2009). If ATL is an
90 independent domain in uniAlaRS, then it must also be an independent domain in bacGlyRS. The
91 ATL domains of uniAlaRS and bacGlyRS are similar to each other in sequence and structure
92 (Figure 3A). We infer here from structural considerations that the ATL can fold independently of
93 other elements of bacGlyRS. Previously, the ATL domain was described incorrectly as a
94 dependent element of the HD domain (Ju, et al. 2021; Han, et al. 2023), rather than as an
95 independent domain. The ATL domain of uniAlaRS contains a C-terminal α-helical decoration
96 that is absent from the ATL domain of bacGlyRS.

97 In addition, we observe that the function of the ATL domain has been mischaracterized.
98 The ATL domain in uniAlaRS is called the "putative anticodon-binding domain" in some
99 classification databases (SCOPe ID a.203.1.1, (Chandonia, et al. 2017), and ECOD ID 613.1.1.1
100 (Schaeffer, et al. 2017)). However, neither the ATL domain of uniAlaRS nor the ATL domain of
101 bacGlyRS contacts the anti-codon stem-loop of tRNA (Naganuma, et al. 2014; Kumar, et al.
102 2019). Both interact extensively with the amino acid acceptor stem. This mischaracterization of
103 the ATL domain structure and function has obscured its evolutionary origins.

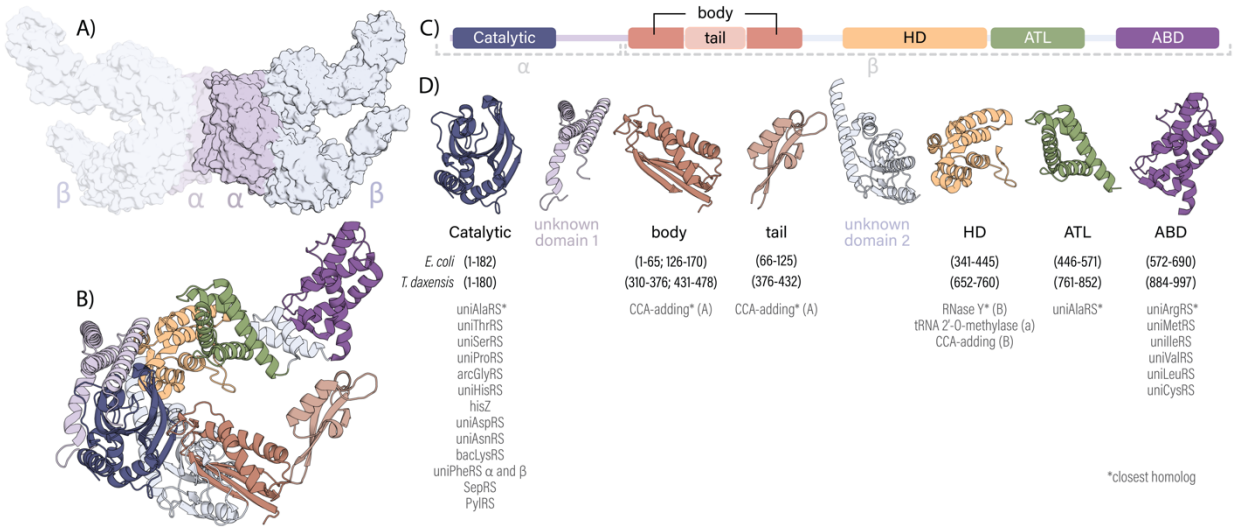

Figure 2. Domain organization of bacGlyRS. Structure of bacGlyRS of *Thermoanaerotrix daxensis* with naturally linked α and β-subunits (PDB: 7LU4). (A) Surface representation of bacGlyRS indicating the region that corresponds to the α- (light purple) and β- (light blue) subunits in other bacteria. (B) A ribbon representation of bacGlyRS shows structural units in different colors: dark blue, catalytic domain; light purple, unknown domain 1; dark orange, body domain; light orange, tail domain; light blue, unknown domain 2; yellow, HD domain; green, ATL domain; and dark purple, ABD domain. (C) Domains and structural units of bacGlyRS. (D) Proteins containing domains with high-scoring sequence similarity by bidirectional BLAST.

104

105 bacGlyRS and uniAlaRS homology is observed in key functional domains

106 The primary functions of bacGlyRS and uniAlaRS are conducted by two domains that are
107 conserved between them. The primary functions of aaRSs are to recognize tRNAs and amino
108 acids and to covalently join them together. The catalytic and ATL domains, which carry out
109 these essential functions are homologous between bacGlyRS and uniAlaRS.

110 The ATL domains of uniAlaRS and bacGlyRS recognize tRNAs by variations of a common
111 mechanism. $tRNA^{Gly}$ and $tRNA^{Ala}$ are recognized by uniAlaRS and bacGlyRS by interactions of
112 ATL with the acceptor stem (Naganuma, et al. 2009; Naganuma, et al. 2014; Han, et al. 2023).
113 The tRNA acceptor stem is the first seven base pairs of tRNA, the unpaired nucleotide at
114 position 73, and the 3'-CCA tail. In species containing bacGlyRS, the acceptor stems of $tRNA^{Gly}$
115 and $tRNA^{Ala}$ differ in three pairs and in position 73. These three base pairs and position 73 are
116 signatures (Penev, et al. 2021), meaning that they are highly conserved within each group,
117 $tRNA^{Gly}$ or $tRNA^{Ala}$, but differ between the two groups. The specific signatures of $tRNA^{Gly}$ are
118 base pairs C2:G71, G3:C70, G4:C69, and U73. The signatures of $tRNA^{Ala}$ are G2:C71, G3:U70,
119 C4:G69, and A73. In uniAlaRS, the signatures include the wobble base pair G3:U70, which is the
120 main identity determinant for $tRNA^{Ala}$ aminoacylation by uniAlaRS (McClain, et al. 1991;
121 Naganuma, et al. 2009; Naganuma, et al. 2014).

122 The ATL domain has adapted to distinguish between features of the amino acid acceptor
123 stems of $tRNA^{Gly}$ and $tRNA^{Ala}$ (Figure 3). We have identified four distinct ATL adaptations,
124 allowing a common fold to recognize different tRNAs. To discriminate between the two
125 different tRNAs, the ATL domain (i) slightly shifts global position relative to the tRNA, (ii) shifts
126 local position by conformation change, (iii) changes sequence, and (iv) acquires α-helical
127 decorations. A comparison of the global positions of ^{bacGly}ATL and ^{Ala}ATL relative to the tRNAs
128 reveals a modest shift of 4.1 Å. In addition, the root mean square deviation (RMSD) after direct
129 superimpositions of ^{bacGly}ATL to ^{Ala}ATL quantitates local changes in conformation and is 3.1 Å
130 (Figure 3A). ATL undergoes both global and local changes in position.

131 The global pose of the ATL domain is generally conserved between bacGlyRS and
132 uniAlaRS. To characterize local positional shifts, we have followed the protocol used for tRNA
133 and have reindexed the ATL residues so that a common index (subscripted here) indicates the

134 same column in the sequence alignment. $^{bacGly}ATL_{10}$ indicates the amino acid at the 10th
135 alignment position of ATL domain of bacGlyRS while $^{Ala}ATL_{10}$ indicates the amino acid of the ATL
136 domain of uniAlaRS at the same alignment position (Figure 3F). Differences in local positioning
137 of amino acids within ATL are seen by differences in ATL indexes of a common interaction. For
138 example, local positional shifts are seen in $^{bacGly}ATL_{29}$ and $^{Ala}ATL_{28}$ interaction with the G1 base
139 and the discriminator positions of the tRNAs (U73 in $tRNA^{Gly}$ and A73 in $tRNA^{Ala}$) and by
140 $^{bacGly}ATL_{35}$ and $^{Ala}ATL_{34}$ interaction with C72 and the discriminator. Base pair G1:C72 is
141 conserved in both $tRNA^{Gly}$ and $tRNA^{Ala}$. By contrast, there is no shift at index 26. Both $^{bacGly}ATL_{26}$
142 and $^{Ala}ATL_{26}$ interact with position 70, which is a C in $tRNA^{Gly}$ and a U in $tRNA^{Ala}$. The ATL domain
143 interacts with the base of the discriminator position (U73 in $tRNA^{Gly}$, and A73 in $tRNA^{Ala}$).

144 The ^{bacGly}ATL interaction with tRNA appears intrinsically less stringent than that of the
145 ^{Ala}ATL domain. The ^{bacGly}ATL domain contacts nucleotides of only one strand in the region of
146 C68, C69 and C70. It does not contact the opposing strand (G3, G4 or G5, Figure 3D). By
147 contrast, ^{Ala}ATL targets both strands of $tRNA^{Ala}$, interacting with both bases of base-pairs
148 G3:U70 and C4:G69 in part by using the ^{Ala}ATL-specific C-terminal decoration (Figure 3E).

149 The global pose of the catalytic domain is radically different between bacGlyRS and
150 uniAlaRS. In uniAlaRS, the catalytic and ATL domains are directly adjacent in sequence and
151 structure. By contrast, in bacGlyRS, the HD and BT domains are inserted between the catalytic
152 and ATL domains. The HD domain shifts and rotates, by 130°, the catalytic domain of bacGlyRS
153 compared to that of uniAlaRS (Figure 3B). The acceptor arm of $tRNA^{Gly}$ bends to accommodate
154 the rotations (Figure 3C). To accommodate the change in the position of the catalytic domain of
155 bacGlyRS, the conformation of four unpaired nucleotides adjusts, such that the position of the
156 terminal adenosine is shifted by 17 Å in the bacGlyRS complex compared to in the uniAlaRS
157 complex.

158 The conformation of the CCA tail and the position of the terminal adenosine appear to
159 be unique in bacGlyRS compared to other aaRSs. The CCA tail of $tRNA^{Gly}$ bends away from the
160 anti-codon in the complex with bacGlyRS. All other class II aaRSs bend the CCA tail in the
161 opposite direction, towards the anticodon (Figure S1) or maintain the relaxed linear structure
162 (Mohan, et al. 2009). bacGlyRS shifts the terminal adenosine toward the minor groove. The

position of the terminal adenosine is shifted by 19 Å in the tRNA complex of bacGlyRS compared to in PheRS (Figure S1).

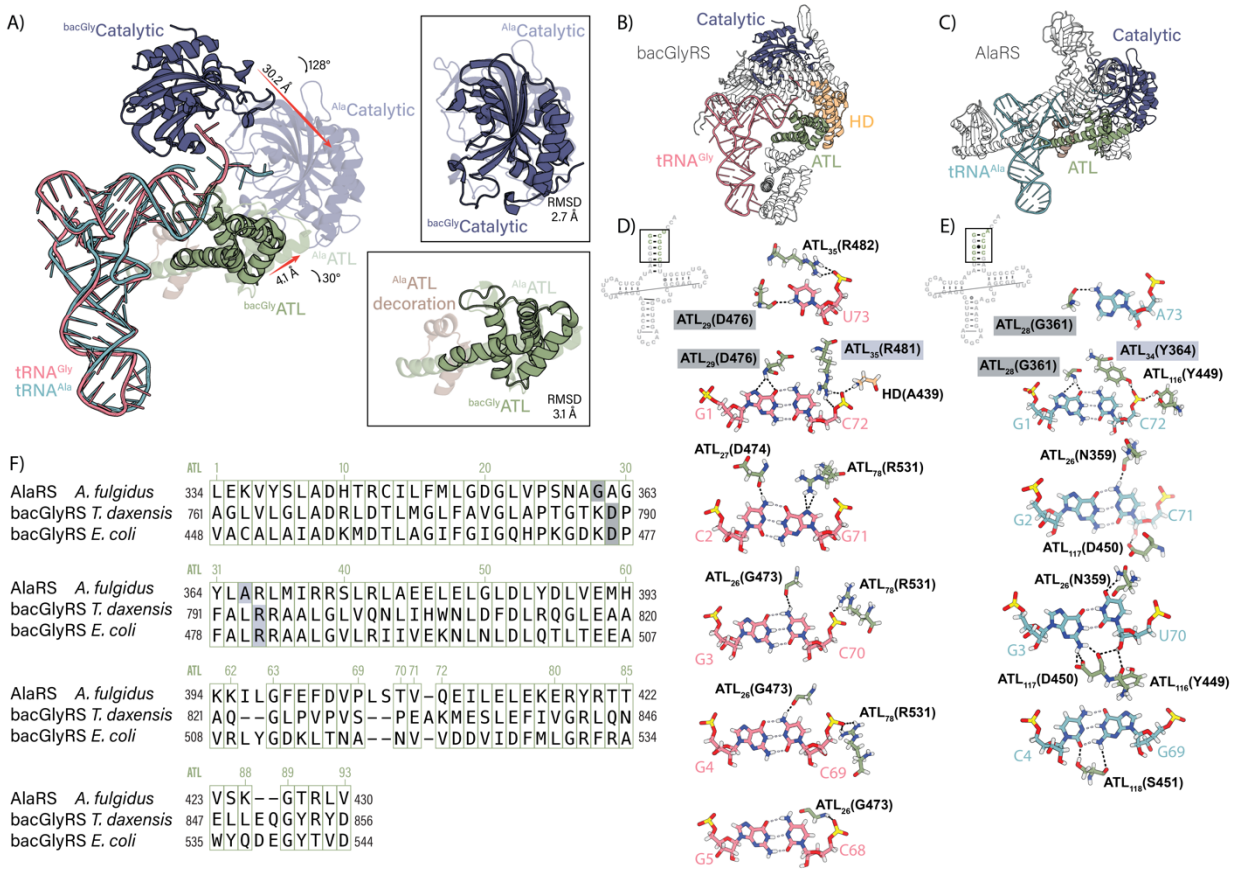

Figure 3. Structures of bacGlyRS and uniAlaRS showing regions of homology. (A) Comparison of the relative position of the catalytic and ATL domains of bacGlyRS (PDB: 7YSE) and uniAlaRS with respect to their cognate tRNA molecules. Insets: Structure superimposition of the catalytic and ATL domains of bacGlyRS and uniAlaRS. (D) Hydrogen bonds between the ATL domain of bacGlyRS and its cognate tRNA. (E) Hydrogen bonds between the ATL domain of uniAlaRS and its cognate tRNA. bacGlyRS of *Escherichia coli* in complex with tRNA. uniAlaRS of *Archaeoglobus fulgidus* in complex with tRNA (PDB: 3WQY).

The catalytic and ATL domains of bacGlyRS resemble archaeal uniAlaRS homologs

We formulated a tentative model in which bacGlyRS originated via bacterial duplication of a two-domain catalytic-ATL uniAlaRS module. This model predicts that bacGlyRS component domains would be most similar to bacterial uniAlaRS homologs. Therefore, we investigated the similarity of bacGlyRS domains to archaeal and bacterial sequences.

A composite multiple-sequence alignment (MSA) containing both bacterial and archaeal uniAlaRS sequences reveals lineage-specific signature positions and insertions (Figure S2). The sequences of bacterial uniAlaRS and archaeal uniAlaRS are distinctive. This type of pattern in a

174 composite MSA is called a block structure and is observed in many universal aaRS and other
175 proteins of the translation system (Woese, et al. 2000; Vishwanath, et al. 2004). Signature
176 positions correspond to columns in a composite MSA that are conserved within groups but are
177 different between groups (Penev, et al. 2021).

178 We compared single-sequences using the program CLANS (Altschul, et al. 1990; Frickey
179 and Lupas 2004) and profiles using the program HH-align (Steinegger, et al. 2019). Profiles are
180 calculated from multiple-sequence alignments and reflect position-specific sequence variability.

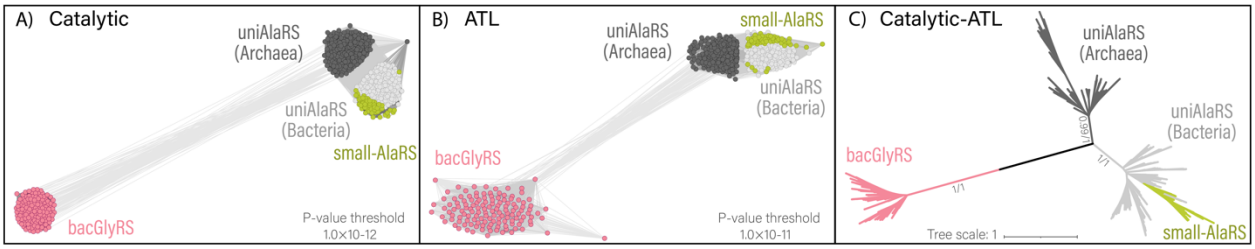

Figure 4. Sequence analysis of bacGlyRS protein domain homologs. Cluster map based on pairwise sequence similarities. Nodes represent sequences; connections between nodes indicate matches between pairs of sequences with a *P*-value lower than the threshold. (A) Pairwise sequence similarity between the catalytic domains of bacGlyRS and uniAlaRS. (B) Pairwise sequence similarity between the ATL domains of bacGlyRS and uniAlaRS. (C) Maximum-likelihood tree of catalytic and ATL domains including bacGlyRS and uniAlaRS sequences rooted at the midpoint. small-AlaRS are sequences from an anomalous subgroup of uniAaRS found in some archaea and bacteria.

181 The catalytic and ATL domains of bacGlyRS resemble archaeal homologs more closely
182 than bacterial homologs (Figures 4A and B). Therefore, our tentative model in which bacGlyRS
183 originated by duplication of a bacterial catalytic-ATL uniAlaRS module is not supported by the
184 data (Figure 5D). A cluster map of sequences from the catalytic domains of bacGlyRS and
185 uniAlaRS at a P-value threshold of 10^{-12} shows connections between bacGlyRSs and archaeal
186 uniAlaRSs. The cluster map lacks connections between bacGlyRSs and bacterial uniAlaRSs
187 (Figure 4A). Similarly, a cluster map of sequences from the ATL domains of bacGlyRS and
188 uniAlaRS at a P-value threshold of 10^{-11} shows connections between bacGlyRSs and archaeal
189 uniAlaRSs. This cluster map lacks connections between bacGlyRSs and bacterial uniAlaRSs
190 (Figure 4B). Thus, the catalytic and ATL domains of bacGlyRS resemble archaeal homologs and
191 not bacterial homologs.

192 The catalytic and ATL domains represent an evolutionary and functional core

193 The canonical uniAlaRS has five component domains; (i) the catalytic domain, which
194 aminoacylates the tRNA, (ii) the ATL domain, which recognizes $tRNA^{Ala}$, (iii) a cradle loop barrel

195 domain, (iv) an editing domain and (v) a C-terminal domain (C-Ala) that binds to the tRNA elbow
196 formed by the D- and T- loops (Swairjo, et al. 2004; Naganuma, et al. 2009) (Figures 3C and S1).
197 We identified a subgroup of uniAlaRSs that have only three domains: catalytic, ATL, and editing
198 (Figures 4 and S1). uniAlaRSs in this subgroup lack the cradle loop barrel domain and the C-Ala
199 domain of canonical uniAlaRSs (Figure S2). We call this anomalous subgroup small-AlaRS. small-
200 AlaRS proteins have bacteria-specific insertions, suggesting that small-AlaRS sequences were
201 transferred to some archaea via horizontal gene transfer. The maximum likelihood tree of
202 catalytic and ATL domain sequences supports a horizontal transfer hypothesis of small-AlaRS to
203 some archaea (Figure 4C and S3). We identified small-AlaRS sequences in the proteomes of CPR
204 bacteria, Candidatus Thermoplasmatota, and DPANN, and Asgard groups. The distribution of
205 small-AlaRS sequences, and recently discovered viral mini-AlaRS sequences (Antika, et al. 2023),
206 which lack the cradle loop barrel, the editing and the C-Ala domains of canonical cellular
207 uniAlaRSs, support the hypothesis of a functional core composed of the catalytic and ATL
208 domain.
209 The combined data support a model in which bacGlyRS and uniAlaRS emerged before
210 LUCA from a shared ancestor formed by a catalytic and ATL domains (Figure 5A and B). Other
211 possibilities, which appear less parsimonious, put the emergence of the bacGlyRS core after the
212 divergence of LUCA. Possible scenarios are: 1) formation of the proto-bacGlyRS in the bacterial
213 lineage independently from the uniAlaRS catalytic-ATL core (Figure 5C); and 2) formation of the
214 proto-bacGlyRS by partial or complete duplication of the uniAlaRS gene in the archaeal lineage,
215 followed by a lateral gene transfer of the proto-bacGlyRS from archaea to the bacterial lineage,
216 and 3) formation of the proto-bacGlyRS by partial or complete duplication of the uniAlaRS gene
217 in the bacterial lineage (Figure 5D).
218

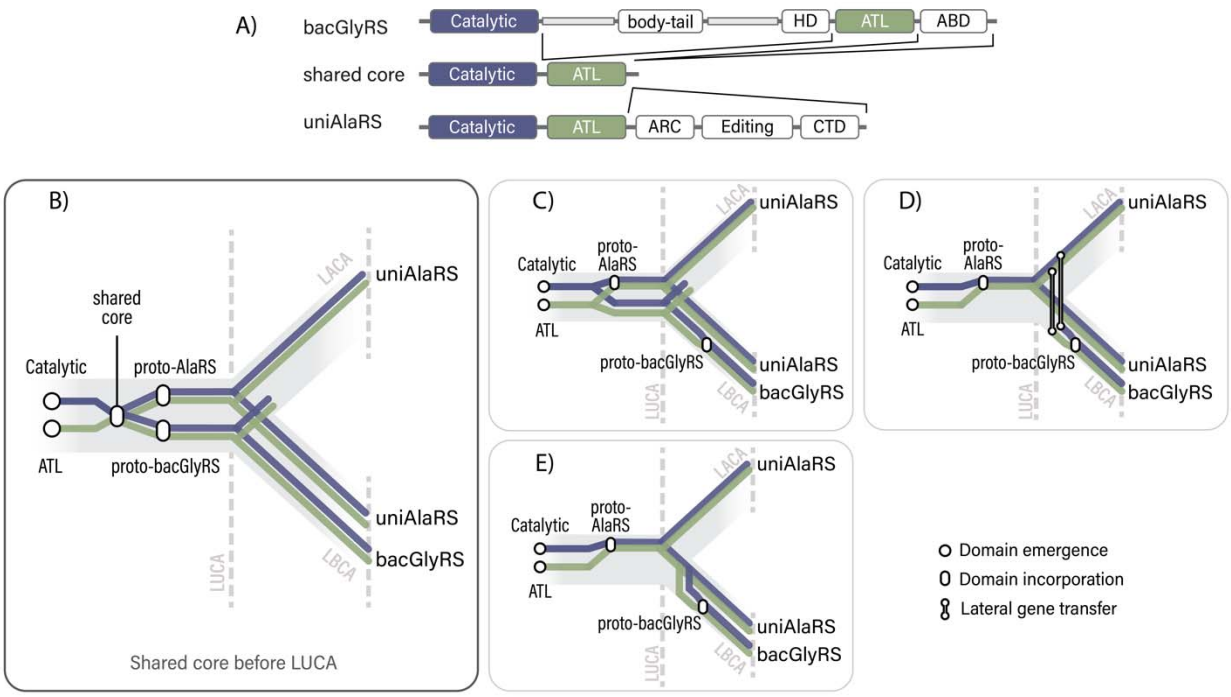

Figure 5. Emergence of bacGlyRS through molecular bricolage. (A) Comparative diagram of the multidomain arrangements of bacGlyRS and uniAlaRS. (B) Emergence of the common ancestor of bacGlyRS and uniAlaRS before the divergence of LUCA. Duplication of the common ancestor gives rise to the proto-uniAlaRS and proto-bacGlyRS. The ancestor of bacGlyRS is lost in archaea. (C) Emergence and duplication of the catalytic and ATL domains at LUCA. Before the divergence between archaea and bacteria, one copy of the catalytic domain and one copy of the ATL domain form the uniAlaRS core, the paralogs of the catalytic and ATL domains continue as single-domain proteins. Secondary loss of the copy of the catalytic and ATL domains in archaea. Formation of the bacGlyRS core in bacteria. (D and E) Evolution of the core of uniAlaRS before the divergence of LUCA. (D) The catalytic and ATL domains are horizontally transferred from the archaeal to the bacterial lineage and formation of the bacGlyRS core. (E) Duplication of the catalytic and ATL domains from uniAlaRS in bacteria and formation of the bacGlyRS core.

Discussion

The presence of bacGlyRS in a wide range of bacterial species suggests that it existed in the Last Bacterial Common Ancestor (LBCA). The presence of bacGlyRS in Terrabacteria and Gracilicutes, the two major bacterial groups (Figure 1) supports this early origin. However, the distribution of bacGlyRS and arcGlyRS in the bacterial tree appears patchy and irregular. Currently, we do not have a model that fully explains the discontinuous distribution over evolutionary history. The dominance of arcGlyRS in CPR proteomes suggests that bacteria with reduced genomes prefer a smaller enzyme.

The main functions of aaRSs are to recognize tRNAs and amino acids, and to perform aminoacylation. bacGlyRS is one of the largest aaRSs and is usually composed of two subunits, with a catalytic domain in the α subunit and several additional recognition domains in the β subunit. In most bacteria (Figure 1) bacGlyRS is a heterotetramer (Gomez and Ibba 2020). The

231 two subunits are joined in a single polypeptide chain in some bacteria (Figure 1); bacGlyRS with
232 naturally linked α and β subunits form dimers (Figure 2A).

233 A domain-by-domain analysis reveals evolutionary relationships between bacGlyRS
234 recognition domains and four different RNA-modifying proteins (Figure 2). The catalytic domain
235 was exapted (Frenkel-Pinter, et al. 2022) by bacGlyRS to perform aminoacylation. The ATL
236 domain was exapted to recognize the acceptor-stem of the $tRNA^{Gly}$. These two domains are
237 closely related to domains in the universal uniAlaRS. The HD domain was exapted apparently to
238 interact with the acceptor stem of $tRNA^{Gly}$ (Han, et al. 2023) and its presence influences the
239 interaction of bacGlyRS and the CCA tail of $tRNA^{Gly}$ (discussed below). The HD domain in
240 bacGlyRS is closely related to the bacterial RNase Y. The body-tail domains were exapted by
241 bacGlyRS for interacting with the tRNA acceptor stem, these two domains are also found
242 performing an equivalent function in the archaeal CCA-adding enzyme (Schimmel and Yang
243 2004; Tomita, et al. 2004). The ABD domain was exapted by bacGlyRS to interact with the anti-
244 codon loop and is homologous to a domain in the class I uniArgRS.

245 Here we provide evidence that ATL is an independent domain rather than a dependent
246 component of the HD domain of bacGlyRS as previously proposed. This discovery sheds light on
247 the evolutionary history of aaRSs. Homology between the catalytic and ATL domains of
248 bacGlyRS and uniAlaRS suggests deep ancestry of their core functional domains. It is likely that
249 an ancestral two-domain aminoacyl tRNA synthetase gave rise to two separate enzymes, one
250 for glycine and another for alanine. This model is consistent with sequence and structural
251 similarities between the core domains of bacGlyRS and uniAlaRS.

252 The conservation of the ATL domain pose in uniAlaRS and bacGlyRS is consistent with a
253 model in which the ATL domain is ancestral to both uniAlaRS and bacGlyRS (Fig. 5B). We have
254 identified four adaptations of the ATL, allowing a common fold to recognize different tRNAs.
255 The ATL domain (i) slightly shifts global position relative to the acceptor stem of the tRNA, (ii)
256 shifts local positions of amino acid backbone and sidechains by conformation change, (iii)
257 changes sequence, and (iv) acquires a C-terminal decoration in the uniAlaRS that is absent from
258 the bacGlyRS.

259 bacGlyRS and uniAlaRS use different strategies to achieve final recognition of their
260 cognate tRNAs (Figure 3). ATL domains of both enzymes have similar interactions with the
261 discriminator base and the G1:C72 base pair which is conserved in $tRNA^{Gly}$ and $tRNA^{Ala}$.
262 uniAlaRS uses the ATL domain to interact intensively with the acceptor stem, employing the C-
263 terminal decoration to assist in targeting both strands of the acceptor step. The ATL domain in
264 bacGlyRS lacks C-terminal decoration and interacts less intensively with the acceptor stem. In
265 the bacGlyRS complex, the ATL domain only contacts nucleotides of one strand of the acceptor
266 stem. To compensate, bacGlyRS interacts with other regions of the tRNA. bacGlyRS but not
267 uniAlaRS exapted additional recognition domains through a process of bricolage.

268 The conformation of the CCA tail and the position of the terminal adenosine appear to
269 be unique in bacGlyRS compared to other aaRSs. In uniAlaRS, the catalytic and ATL domains are
270 directly adjacent in sequence and structure. tRNA interacts with the CCA tail of uniAlaRS is a
271 relaxed state, which is reasonably similar to the structure in the free tRNA (Westhof, et al.
272 1988). Insertion of the HD domain in bacGlyRS caused a translation and rotation of the catalytic
273 domain relative to the ATL domain (Figure 3B). To accommodate this translation/rotation, the
274 position of the terminal adenosine shifts by 24 Å. The acceptor-arm of $tRNA^{Gly}$ in the bacGlyRS
275 complex bends acutely in the direction of the major groove (Figure 3C).

276 The sequences of bacGlyRS and archaeal uniAlaRS have distinctive features. The
277 catalytic and ATL domains of bacGlyRS more closely resemble archaeal homologs than bacterial
278 homologs (Figures 4A and B). The combined results suggest that bacGlyRS and uniAlaRS share a
279 common ancestor composed of the catalytic and ATL domains. The catalytic and ATL domains
280 represent an evolutionary and functional core. A model in which bacGlyRS originated by
281 duplication of a bacterial catalytic-ATL uniAlaRS module is not supported by the data (Figure
282 5D).

283 In a simple but attractive model, insertion of the body-tail and HD domains (or possibly just the
284 HD domain) into a proto-AlaRS gene formed of a novel enzyme, a proto-bacGlyRS. In this model, the
285 interaction between ATL domain and the tRNA remained relatively unchanged as the catalytic domain
286 was displaced and the CCA tail bends towards the minor groove. The proto-bacGlyRS enzyme underwent
287 adaptations that conferred a new amino acid specificity (Gly) and a new tRNA identity ($tRNA^{Gly}$). It is
288 possible that the new tRNA was better suited to the CCA tail orientation imposed by the HD insertion.

Our observations here reflect deep evolutionary history. Glycine and alanine appear to be ancient amino acids, both prebiotically available and biologically important. They are among the most abundant amino acids produced in spark discharge experiments (Parker, et al. 2011) and are found in chondrite asteroids (Botta, et al. 2002). We propose that bacGlyRS and uniAlaRS emerged from a shared core before LUCA (Figure 5A and B), from catalytic and ATL domains that had separate and distinct origins. These prebiotic amino acids share the first and third codon positions, and the proteins that aminoacylate $tRNA^{Gly}$ and $tRNA^{Ala}$ share sequence and structure relationships. It appears we are seeing evidence of an ancestral protein with non-specific catalytic activity that may have been able to link either glycine or alanine to an RNA, presumably an ancestor of tRNA. The deep relationship between alanine and glycine is seen in their shared chemistry and biology.

Methods

Sequence similarity search

A multiple sequence alignment (MSAs) of bacGlyRS sequences was calculated for a representative set of archaea and bacteria. The MSA was trimmed to the domain boundaries as defined by the structure analysis. The MSAs of the domains were converted to profile Hidden Markov Models using the HHsuite version 3.3.0 (Achsel, et al. 2001). A profile-profile similarity search on the Evolutionary Classification of Domains (ECOD) database was performed with HHsearch using the profiles of bacGlyRS domains as queries. Target domains in ECOD yielding HHsearch probabilities greater than 60% over more than 50 aligned columns were retrieved to serve as queries of a second search on a sequence database containing the proteomes of the representative set of bacterial and archaeal species. The second search was performed using jackhmmer from the HMMER suite version 3.3.1 (Achsel, et al. 2001), with default parameters (five iterations). The ECOD database only contains domains that are directly annotated on proteins with experimentally determined structures. The second search aims to identify sequence similarity in proteins without experimentally determined structures. Sequences passing the threshold values (E-value $< 1\times10^{-3}$, query coverage 80%) were retrieved and trimmed to the region of sequence similarity.

We broadly describe the phylogenetic distribution of the best hits according to the number of sampled proteomes containing homologous sequences as follows: universal (presence in > 50% of archaea and >50% of bacteria); archaea-specific (presence in >50% of archaea and <10% of bacteria); bacteria-specific (presence in < 10% of archaea and <50% of bacteria); lineage-specific within archaea (presence in <50% of archaea and <50% of bacteria); lineage-specific within bacteria (presence in <10% of archaea and <50% of bacteria sampled); or of complex distribution (presence in 10-50% of archaea and 10-50% of bacteria).

Cluster analysis

The query MSAs and trimmed target sequences retrieved from the second sequence similarity search were clustered with CLANS (Frickey and Lupas 2004) based on pairwise sequence similarities. To find the best sequence match of each query domain, the cut-offs of the clustering were adjusted to find the P-value for which the group of query sequences shows sequence relationships with only one other group of orthologs. The subset of best matches was further clustered to reveal possible differences between archaeal and bacterial sequences. Bacterial and archaeal uniAlaRS sequences displayed differential similarity to bacGlyRS sequences on the cluster analysis. To corroborate the clustering analysis results, additional analyses were performed at the level of profiles.

Pairwise profile analysis: HHalign

An MSA containing orthologous sequences of uniAlaRS was further divided to include only archaeal sequences or only bacterial sequences. The MSAs of archaeal and bacterial sequences of uniAlaRS and the MSA of bacGlyRS were trimmed to the catalytic and ATL domains. The MSA of the catalytic domain of bacGlyRS was aligned to the MSAs of the catalytic domains of archaeal and bacterial uniAlaRS sequences using HHalign. The HHalign scores of these MSA-MSA comparisons agree with the differential pairwise similarities observed in CLANS between archaeal and bacterial sequences.

ML tree

The composite MSA of the catalytic domain containing sequences from bacGlyRS and archaeal sequences from uniAlaRS was concatenated to the composite MSA of the ATL domain.

The concatenated MSA was trimmed with trimAl (Capella-Gutiérrez, et al. 2009). A Maximum Likelihood tree was calculated using the trimmed MSA using PhyML (Guindon, et al. 2010) and SMS model selection tool (Lefort, et al. 2017) on the Montpellier Bioinformatics Platform. Branch support corresponds to a Bayesian-like transformation of aLRT (aBayes) (Anisimova, et al. 2011).

Data Availability

Sequence alignments and CLANS input files associated with this manuscript have been deposited in the FigShare repository DOI: 10.6084/m9.figshare.23269007.

Acknowledgements

The authors dedicate this work to the memory of Professor ATL who passed away on September 9th, 2021. Professor ATL worked on the structural determination of the bacterial glycyl tRNA synthetase and performed the initial structural analysis. The authors thank Jessica Bowman, Eric Smith and Arturo Becerra for discussions. This work was funded by the National Aeronautics and Space Administration grant 80NSSC18K1139. CAC was sponsored by the National Aeronautics and Space Administration (NASA) through a contract with ORAU.
Author contributions: Conceptualization, ATL, MIVS and CAC. Methodology, CAC. Formal analysis CAC, MA, AHG, MIVS and LDW. Writing – Original Draft, CAC and LDW. Writing – Review & Editing, LRP and MIVS.

References

Achsel T, Stark H, Lührmann R. 2001. The Sm Domain Is an Ancient RNA-Binding Motif with Oligo(U) Specificity. Proc Natl Acad Sci USA 98:3685-3689.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. J Mol Biol 215:403-410.

Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-Based Approximation Schemes. Syst Biol 60:685-699.

Antika TR, Chrestella DJ, Tseng YK, Yeh YH, Hsiao CD, Wang CC. 2023. A Naturally Occurring Mini-Alanyl-tRNA Synthetase. Commun Biol 6:314.

Botta O, Glavin DP, Kminek G, Bada JL. 2002. Relative Amino Acid Concentrations as a Signature for Parent Body Processes of Carbonaceous Chondrites. Orig Life Evol Biosph 32:143-163.

Bowman JC, Petrov AS, Frenkel-Pinter M, Penev PI, Williams LD. 2020. Root of the Tree: The Significance, Evolution, and Origins of the Ribosome. Chem Rev 120:4848-4878.

Burbaum JJ, Schimmel P. 1991. Structural Relationships and the Classification of Aminoacyl-tRNA Synthetases. J Biol Chem 266:16965-16968.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. Trimal: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. Bioinformatics 25:1972-1973.

Chandonia JM, Fox NK, Brenner SE. 2017. Scope: Manual Curation and Artifact Removal in the Structural Classification of Proteins - Extended Database. J Mol Biol 429:348-355.
Dimas-Torres J, Rodríguez-Hernández A, Valencia-Sánchez M, Campos-Chávez E, Godínez-López V, Rodríguez-Chamorro D, Grøtli M, Fleming C, Hernández-González A, Arciniega M. 2021. Bacterial Glycyl tRNA Synthetase Offers Glimpses of Ancestral Protein Topologies.
Eriani G, Delarue M, Poch O, Gangloff J, Moras D. 1990. Partition of tRNA Synthetases into Two Classes Based on Mutually Exclusive Sets of Sequence Motifs. Nature 347:203-206.
Fitch WM. 1970. Distinguishing Homologous from Analogous Proteins. Syst Zool 19:99-113.
Fournier GP, Andam CP, Alm EJ, Gogarten JP. 2011. Molecular Evolution of Aminoacyl tRNA Synthetase Proteins in the Early History of Life. Orig Life Evol Biosph 41:621-632.
Frenkel-Pinter M, Petrov AS, Matange K, Travisano M, Glass JB, Williams LD. 2022. Adaptation and Exaptation: From Small Molecules to Feathers. J Mol Evol 90:166-175.
Frickey T, Lupas A. 2004. Clans: A Java Application for Visualizing Protein Families Based on Pairwise Similarity. Bioinformatics 20:3702-3704.
Giegé R, Eriani G. 2023. The tRNA Identity Landscape for Aminoacylation and Beyond. Nucleic Acids Res 51:1528-1570.
Gomez MAR, Ibba M. 2020. Aminoacyl-tRNA Synthetases. RNA 26:910-936.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of Phyml 3.0. Syst Biol 59:307-321.
Han L, Luo Z, Ju Y, Chen B, Zou T, Wang J, Xu J, Gu Q, Yang X-L, Schimmel P. 2023. The Binding Mode of Orphan Glycyl-tRNA Synthetase with tRNA Supports the Synthetase Classification and Reveals Large Domain Movements. Science Advances 9:eadf1027.
Hartman H. 1995. Speculations on the Evolution of the Genetic Code Iv the Evolution of the Aminoacyl-tRNA Synthetases. Orig Life Evol Biosph 25:265-269.
Jacob F. 1977. Evolution and Tinkering. Science 196:1161-1166.
Ju Y, Han L, Chen B, Luo Z, Gu Q, Xu J, Yang XL, Schimmel P, Zhou H. 2021. X-Shaped Structure of Bacterial Heterotetrameric tRNA Synthetase Suggests Cryptic Prokaryote Functions and a Rationale for Synthetase Classifications. Nucleic Acids Res 49:10106-10119.
Koonin EV, Novozhilov AS. 2017. Origin and Evolution of the Universal Genetic Code. Annu Rev Genet 51:45-62.
Kumar A, Åqvist J, Satpati P. 2019. Principles of tRNA(Ala) Selection by Alanyl-tRNA Synthetase Based on the Critical G3·U70 Base Pair. ACS Omega 4:15539-15548.
Lefort V, Longueville JE, Gascuel O. 2017. Sms: Smart Model Selection in Phyml. Mol Biol Evol 34:2422-2424.
Levitt M, Chothia C. 1976. Structural Patterns in Globular Proteins. Nature 261:552-558.
Logan D, Mazauric M, Kern D, Moras D. 1995. Crystal Structure of Glycyl-tRNA Synthetase from Thermus Thermophilus. EMBO J 14:4156-4167.
McClain WH, Foss K, Jenkins R, Schneider J. 1991. Four Sites in the Acceptor Helix and One Site in the Variable Pocket of tRNA (Ala) Determine the Molecule's Acceptor Identity. Proc Natl Acad Sci USA 88:9272-9276.
Mohan S, Hsiao C, VanDeusen H, Gallagher R, Krohn E, Kalahar B, Wartell RM, Williams LD. 2009. Mechanism of RNA Double Helix-Propagation at Atomic Resolution. J Phys Chem B 113:2614-2623.

Moody ER, Mahendrarajah TA, Dombrowski N, Clark JW, Petitjean C, Offre P, Szöllősi GJ, Spang A, Williams TA. 2022a. An Estimate of the Deepest Branches of the Tree of Life from Ancient Vertically Evolving Genes. eLife 11:e66695.

Moody ERR, Mahendrarajah TA, Dombrowski N, Clark JW, Petitjean C, Offre P, Szöllősi GJ, Spang A, Williams TA. 2022b. An Estimate of the Deepest Branches of the Tree of Life from Ancient Vertically Evolving Genes. eLife 11.

Moras D. 1992. Structural and Functional Relationships between Aminoacyl-tRNA Synthetases. Trends Biochem Sci 17:159-164.

Naganuma M, Sekine S, Chong YE, Guo M, Yang XL, Gamper H, Hou YM, Schimmel P, Yokoyama S. 2014. The Selective tRNA Aminoacylation Mechanism Based on a Single G•U Pair. Nature 510:507-511.

Naganuma M, Sekine S, Fukunaga R, Yokoyama S. 2009. Unique Protein Architecture of Alanyl-tRNA Synthetase for Aminoacylation, Editing, and Dimerization. Proc Natl Acad Sci USA 106:8489-8494.

Ohno S. 2013. Evolution by Gene Duplication: Springer Science & Business Media.

Ostrem DL, Berg P. 1974. Glycyl Transfer Ribonucleic Acid Synthetase from Escherichia coli: Purification, Properties, and Substrate Binding. Biochemistry 13:1338-1348.

Parker ET, Cleaves HJ, Dworkin JP, Glavin DP, Callahan M, Aubrey A, Lazcano A, Bada JL. 2011. Primordial Synthesis of Amines and Amino Acids in a 1958 Miller H2S-Rich Spark Discharge Experiment. Proc Natl Acad Sci USA 108:5526-5531.

Penev PI, Alvarez-Carreño C, Smith E, Petrov AS, Williams LD. 2021. Twincons: Conservation Score for Uncovering Deep Sequence Similarity and Divergence. PLoS Comput Biol 17:e1009541.

Ponting CP, Schultz J, Copley RR, Andrade MA, Bork P. 2000. Evolution of Domain Families. Advances in protein chemistry 54:185-244.

Porter LL, Rose GD. 2012. A Thermodynamic Definition of Protein Domains. Proc Natl Acad Sci USA 109:9420-9425.

Ribas de Pouplana L. 2020. The Evolution of Aminoacyl-tRNA Synthetases: From Dawn to LUCA. In. The Enzymes: Elsevier. p. 11-37.

Ribas de Pouplana L, Schimmel P. 2001. Aminoacyl-tRNA Synthetases: Potential Markers of Genetic Code Development. Trends Biochem Sci 26:591-596.

Rossman MG, Liljas A. 1974. Letter: Recognition of Structural Domains in Globular Proteins. J Mol Biol 85:177-181.

Schaeffer RD, Liao Y, Cheng H, Grishin NV. 2017. ECOD: New Developments in the Evolutionary Classification of Domains. Nucleic Acids Res 45:D296-d302.

Schimmel P, Yang XL. 2004. Two Classes Give Lessons About Cca. Nat Struct Mol Biol 11:807-808.

Smith TF, Hartman H. 2015. The Evolution of Class II Aminoacyl-tRNA Synthetases and the First Code. FEBS Lett 589:3499-3507.

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation. BMC Bioinformatics 20:473.

Swairjo MA, Otero FJ, Yang X-L, Lovato MA, Skene RJ, McRee DE, de Pouplana LR, Schimmel P. 2004. Alanyl-tRNA Synthetase Crystal Structure and Design for Acceptor-Stem Recognition. Mol Cell 13:829-841.

Tang SN, Huang JF. 2005. Evolution of Different Oligomeric Glycyl-tRNA Synthetases. FEBS Lett 579:1441-1445.

Terada T, Nureki O, Ishitani R, Ambrogelly A, Ibba M, Söll D, Yokoyama S. 2002. Functional Convergence of Two Lysyl-tRNA Synthetases with Unrelated Topologies. Nat Struct Biol 9:257-262.

Tomita K, Fukai S, Ishitani R, Ueda T, Takeuchi N, Vassylyev DG, Nureki O. 2004. Structural Basis for Template-Independent RNA Polymerization. Nature 430:700-704.

Tumbula D, Vothknecht UC, Kim H-s, Ibba M, Min B, Li T, Pelaschier J, Stathopoulos C, Becker H, Söll D. 1999. Archaeal Aminoacyl-tRNA Synthesis: Diversity Replaces Dogma. Genetics 152:1269-1276.

Valencia-Sánchez MI, Rodríguez-Hernández A, Ferreira R, Santamaría-Suárez HA, Arciniega M, Dock-Bregeon A-C, Moras D, Beinsteiner B, Mertens H, Svergun D, Brieba LG, Grøtli M, Torres-Larios A. 2016. Structural Insights into the Polyphyletic Origins of Glycyl tRNA Synthetases. J Biol Chem 291:14430-14446.

Vishwanath P, Favaretto P, Hartman H, Mohr SC, Smith TF. 2004. Ribosomal Protein-Sequence Block Structure Suggests Complex Prokaryotic Evolution with Implications for the Origin of Eukaryotes. Mol Phylogenet Evol 33:615-625.

Westhof E, Dumas P, Moras D. 1988. Restrained Refinement of Two Crystalline Forms of Yeast Aspartic Acid and Phenylalanine Transfer RNA Crystals. Acta Crystallogr A 44 (Pt 2):112-123.

Woese CR, Olsen GJ, Ibba M, Söll D. 2000. Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process. Microbiol Mol Biol Rev 64:202-236.

365