Ancestral protein topologies draw the rooted bacterial Tree of Life

Jorge-Uriel Dimas-Torres[1], Annia Rodríguez-Hernández[1], Marco Igor Valencia-Sánchez[1], Eduardo Campos-Chávez[1], Victoria Godínez-López[1], Daniel-Eduardo Rodríguez-Chamorro[1], Morten Grøtli[2], Cassandra Fleming[2], Adriana Hernández-González[1], Marcelino Arciniega[1], and Alfredo Torres-Larios[1]*

[1] 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, [2]Department of Chemistry and Molecular Biology, University of Gothenburg, SE-412 96 Gothenburg, Sweden.

Address correspondence to: Alfredo Torres-Larios, PhD; Circuito Exterior s/n, Ciudad Universitaria. Apartado postal 70-243, México D.F. 04510, México. Fax 011-52-55-5622-5630; E-mail:torres@ifc.unam.mx.

Abstract

Aminoacyl tRNA synthetases (aaRSs) are among the proposed set of proteins present in the Last Universal Common Ancestor (LUCA). There are two types of glycil tRNA synthetases (GlyRSs), from which the archaeal-eukaryal type is the one suggested to be present in LUCA. Here we present the crystal structure of a complete bacterial glycyl tRNA synthetase (bacGlyRS) and show that indeed, bacGlyRS carries several structural signals that point it at the origin of all aaRSs. Furthermore, considering the modularity of protein evolution and, using only two sub-domain segments with duplicated ancestral topologies from a single, universal primordial protein corresponding to the RNA binding regions of contemporary bacGlyRS, archaeal CCA adding enzyme (arch-CCAadd) and eukaryotic rRNA processing enzyme (euk-rRNA), we calculated a rooted bacterial ToL that agrees with several phyla relationships unaccounted by the available trees.

Introduction

Among the 355 genes proposed to have been present in LUCA, there are eight that code for aaRSs, enzymes that attach cognate amino acids to their corresponding tRNAs. Glycyl tRNA synthetase (GlyRS) is found between these candidates (Weiss et al., 2018). Unexpectedly, there are two types of GlyRS that do not share a single common ancestor: an α_2 homodimer (eukGlyRS), related to HisRS, ThrRS, ProRS and SerRS and present in eukarya, archea, and some bacteria and an $\alpha_2\beta_2$ heterotetramer (bacGlyRS), related to AlaRS and present in bacteria (Ostrem & Berg, 1974), (Tang & Huang, 2005), (Valencia-Sanchez et al., 2016). Due to its distribution in the three life domains, it is believed that eukGlyRS is the ancestral type that was present in LUCA (Weiss et al., 2018). However,

bacGlyRS is found in bacteria that are deeply rooted, like *Thermotoga* and *Cyanobacteria* (Fournier et al., 2015), (Supplementary Figure 1). Therefore, the questions regarding why there are two GlyRSs and which one may have appeared first remain unanswered. The structure of bacGlyRS may offer further insights into the molecular evolution of aaRSs, their protein domains, and the origins of the genetic code (Koonin & Novozhilov, 2017).

Here we show the three-dimensional structure of a functional, complete bacGlyRS from the thermophilic bacteria *Thermanaerotrix daxensis* (Chloroflexi) (Gregoire et al., 2011) in complex with glycine. The structure was solved by X-ray crystallography and verified in solution by negative staining, electron microscopy experiments. The structure represents the biggest of all aaRSs, but integrates the smallest known catalytic domains (α, around 30 kDa) and the largest tRNA interacting subunit (β, around 70 kDa). It consists of an α_2 dimer, with a β-subunit interacting aside each α-subunit. Regarding homology, the β-subunit shows remote relationships with archaeal CCA-adding enzymes-eukaryotic rRNA processing complex, metal-dependent HD phosphohydrolases and the anticodon binding domain (ACBD) of class I aaRSs, particularly ArgRS and CysRS. All these domains belong to the most ancient groups of enzymes and have sequence identities with bacGlyRS below 10%. Unexpectedly, a model in complex with tRNA shows that the orientation of this substrate would be inverted relative to all class II synthetases, approaching the catalytic domain through the minor groove of the tRNA acceptor stem. The anomalous features found in both the catalytic domain (Valencia-Sanchez et al., 2016) and on the putative complex with tRNA, together with the size of the β-subunit, and the homologies with ancient folds and proteins, demonstrate that bacGlyRS α-subunit represents a primitive aaRSs and that the CCA-adding-rRNAprocessing-like architecture

of the β-subunit represents a multifunctional, primordial non-specific RNA binding protein, making bacGlyRS part of these molecular fossils, ancestors of RNA binders involved in translation and RNA metabolism, providing an example of the coevolution of the ribonucleoprotein (RNP) world (Schimmel, 2011). As these three proteins are related between them by a common, duplicated βαββ topology with an ancient ferredoxin-like fold, we calculated a phylogenetic tree using the universe of archaeal and bacterial proteins, which are most proximal by sequence identity. We thus obtained a rooted tree that agrees with several previous observations, but it also suggests novel relationships between phyla, and it places the most deeply rooted bacteria within the anaerobic, sulfate-reducing *Syntrophobacterales*.

Results

- bacGlyRS is a βααβ complex formed around a central α_2 dimer that is not related to eukGlyRS, forming two different subclasses

We were able to obtain highly purified, naturally fusioned αβ-subunits of GlyRS of *Thermoanaerotrix daxensis* (TdGlyRS, Supplementary Figures 9-11) by means of an His-tag affinity chromatography step and a heat treatment. Negative-staining electron microscopy further confirmed the dimeric nature and the relative orientation of the ensemble (Supplementary Figures 19-20). TdGlyRS binds $tRNA^{Gly}$, with an affinity of 3 μM (Supplementary Figure 17). Using a highly similar protein to TdGlyRS from *Anaerolinea thermophila* (AtGlyRS), we verified that the full, linked enzyme, was able to perform the aminoacylation reaction, that was inhibited by the presence of a transition state analog, glycyl-sulfamoyl adenylate (GSA) (Supplementary Figures 12-13). These findings provide biological relevance to our reported crystal structure.

To discover the architecture of the whole enzyme and to unveil the folding domains of the β-subunit, we solved by molecular replacement using the coordinates of the α-subunit dimer and automatic building, the crystal structure of TdGlyRS in complex with glycine in one monomer at 2.7 Å resolution (Table 1). The electron density map unambiguously showed all features and domains of the 200 kDa enzyme and the bound glycine in one monomer of the biological dimer found in the asymmetric unit (Supplementary Fig. 2).

As previously shown (PDB entry 5F5W), the α-subunit forms a homodimer that also functions as the dimeric interface of the $\alpha_2\beta_2$ ensemble (Figure 1), with 58% of its interface formed by a 97 residues helical region located at the C-terminal part of the subunit and situated on top of the signature antiparallel β-strand of class II synthetases (Supplementary Figure 3). There is no conformational change in the α_2 homodimer in complex with the β-subunit (PDB entry 7LU4) with respect to the structure of the α_2 homodimer alone (PDB entry 5F5W). Surprisingly, the β-subunit does not participate in the stabilization of the heterodimer interface. Instead, each β-subunit only interacts with one α-subunit. The α-β interface is formed mostly by highly conserved residues (Supplementary Figure 4) that are complementary in shape and charge (Supplementary Figure 4), which further supports the geometry of their interaction.

The comparison of the apo monomer of TdGlyRS with the glycine-bound monomer (Supplementary Figure 3) reveals the same conformational changes already reported for the crystal structure of the α-subunit of *Aquifex aeolicus* in complex with an analog of glycine-adenylate; a ~5 Å displacement of an 11-residue region where a Trp forms a cation-π interaction with glycine (Supplementary Figure 3). Thus, the reported crystal structure bacGlyRS represents a productive, biologically meaningful complex. The

presence of glycine bound, and the correlated conformational movement already reported for the binding of GlyAMS (Valencia-Sanchez et al., 2016), shows that the binding of the amino acid alone is enough to promote this movement and that this conserved region, shared with AlaRS, is important for the recognition of the amino acid.

The comparison of the two types of GlyRS (Supplementary Figure 5) shows that the organization and relative orientation of the oligomeric ensembles is quite different. In eukGlyRS and most class II aaRSs, besides the catalytic domain, other domains of one monomer are involved in interface contacts to stabilize the oligomer and/or make intersubunit contacts with the tRNA (Supplementary Figure 5), in contrast with bacGlyRS that represents a unique case of homodimeric dependence on merely the catalytic domain, where each β, tRNA-interacting subunit, is completely independent of the homodimer (Figure 1). The bacGlyRS $\alpha_2\beta_2$ heterotetramer would be better defined as a βαaβ complex formed around a central α_2 dimer (Koski et al., 2017). In addition, none of the domains shares structural homology between the GlyRSs, except for the core of the catalytic domain, with extremely limited similarity between the two types of GlyRS. It has been shown that even in this domain, eukGlyRS is more similar to subclass IIa enzymes than to bacGlyRS (Valencia-Sánchez et al., 2016). The difference between the two types of GlyRS must be defined in terms of the different subclasses. Somewhat surprisingly, none of the five domains of the β-subunit is clearly shared between the subclass IId members, bacGlyRS and AlaRS, restraining the homology between these enzymes to the features shared in the catalytic domain (Valencia-Sanchez et al., 2016).

- The homology of bacGlyRS β subunit with tRNA binding domains allows to confidently build a model of the aaRS-tRNA complex

To gain further insight into the evolutionary origin of bacterial GlyRS, we performed structural searches using the algorithm Dali and the DASH database. We gathered the most representative models showing homology for each domain, trimmed the corresponding regions, aligned them using the STAMP algorithm and clustered with Bio3D and BioVinci. We also performed Principal Component Analysis (PCA) on the Cα atoms of the core residues found, according to the algorithm implemented in Bio3D. Surprisingly, the only clear structural homologs that we found (Figure 2, Table 3, Supplementary Figures 6-7) are the class I anticodon binding-domain (ACBD or DALR domain, in particular with ArgRS and CysRS, Supplementary Figure 6), phosphohydrolases possessing the HD domain (Supplementary Figure 7) and the body-tail domains of archaeal tRNA CCA adding enzymes-eukaryotic rRNA processing enzymes (PDB codes 1R89 and 4M5D, respectively).

The homologies found for the β-subunit allowed us to propose a bacGlyRS:tRNA model (Figure 3B), mostly based on the complex of archaeal CCA-adding enzyme with tRNA (Figure 3A). The superposition of the PDB code 1R89 with the coordinates of bacGlyRS, (PDB code 7LU4), renders a putative complex that interacts with the ACBD and leaves the terminal A76 of tRNA within reach to the glycine moiety of the catalytic subunit (Figure 3B, Supplementary Figure 5). Perhaps the most outstanding features of this model are its natural fitting, as given by just placing the template model and, most of all, the resulting relative orientation of the tRNA. Remarkably, if correct, the tRNA recognition would be an hybrid between class I-like type of recognition on a class II aaRS (Supplementary Figure 5). According to their tRNA recognition, atypical for class II aaRS, bacGlyRS and AlaRS could represent the most ancient subclass of aaRSs, as they may illustrate primitive,

somewhat alternate recognition modes given before the class II aaRS:tRNA canonical interaction (Figure 4). This further strengthens the fact that bacGlyRS and AlaRS clearly define a subgroup of aaRSs by themselves (Figure 4).

- bacGlyRS β-subunit presents two architectures: a "recent" α-helical and a putative primitive α+β that may be the result of a triplication event

From the structural analysis of the β-subunit, we defined two different architectures (Figure 2A): an α-helical, comprised of domains related to the anticodon recognition and with clear sequence signatures from the HD and ACBD domains, making this region probably more recent than the second seen architecture (Figure 2A, Table 2), which is an α+β type, and that only shows structural homology with RNA recognition molecules (Figure 6). The finding of a βαββ topology on the three domains comprising the α+β architecture (Figure 5) opens the possibility of a triplication event that gave in part rise to this multidomain protein very early in evolution. This makes sense because the ferredoxin-like fold is proposed to be very ancient (Wolf et al., 1999), (Kim et al., 2012), (Lin et al., 2015), (Raanan et al., 2020) as well as the tRNA acceptor stem (Weiner & Maizels, 1987), (Carter & Wolfenden, 2015). Interestingly, most ribosomal proteins from the large subunit are of the α+β type, with ferredoxin-like βαββ topology, also like the RRM fold, typified by the splicing protein U1A (Rice & Correll, 2008)

- The duplicated βαββ, ferredoxin-like, primitive RNA binding topology present in the three life domains provides a clear coevolution signal of an ancient RNP world.

The duplicated βαββ topology found represents a promiscuous entity, but with a single primitive, non-sequence specific RNA binding activity, when found in tandem. This activity was frozen in time, because although it is found in the three domains of life, it has been

used for single different purposes regarding the binding of ancient RNAs (tRNA acceptor stem and/or rRNA). Thus, we propose that the RNA binding, βαββ tandem topology is an example of the origin of proteins themselves and that their presence in a wide range of different proteins, but with similar ancestral activity, was raised by a single, extremely ancient HGT event, placing the origin of the three domains of life in a very tight window of time. The tandem topology itself may be of pre-LUCA (Last Universal Common Ancestor) origin.

- The ancestral, duplicated βαββ, ferredoxin-like topologies of archaea-CCAadding and bacterial b-GlyRS draw the rooted bacterial Tree of Life.

With the hypothesis described above, we calculated a phylogenetic tree (Figure 7) using the sequences of the available βαββ duplicated topologies of archaea and bacteria domains. Notably, we could use the Archaeal, CCA-adding enzyme sequences to obtain a rooted tree (Figure 7), that yields information consistent with the widely accepted taxonomy, like the overall, Gracilicutes vs Terrabacteria division (Coleman et al., 2021), the general Firmicutes phylogeny division of Limnochordales, Halanaerobiales, Negativicutes (Taib et al., 2020), or the placement of Magnetococcales as an Alpha proteobacteria outgroup (Wang & Luo, 2021). It also proposes that Methanopyrus is the closest Archaea related to bacteria (Long et al., 2020); to divide Deltaproteobacteria in four phyla that do not belong to proteobacteria (Waite et al., 2020). Epsiloproteobacteria is also suggested to be reclassified as not a Proteobacteria, as proposed in (Waite et al., 2017). Notably, our proposal implicates that Thermotogales should be considered members of the Firmicutes, as already suggested by (Zhaxybayeva et al., 2009) and

places Fusobacteria very close to the root of the tree, as recently proposed (Coleman et al., 2021).

Perhaps the two most outstanding features of our proposed tree are: i) the relationship between Cyanobacteria, Chlamydia, Chloroflexi and Actinobacteria. Interestingly, Cyanobacteria and Chloroflexi are two of the seven phyla where reaction-center phototrophy has been reported (Raymond, 2008), (Imhoff et al., 2017), (Ward et al., 2018). Also, the relationship between Cyanobacteria and Chlamydia has been reported (Brinkman et al., 2002), (Ball et al., 2015), (Zeng & Dehesh, 2021), although also disputed (Domman et al., 2015). Finally, Actinobacteria seems to be somewhat related with photosynthesis, as already suggested (Gupta & Khadka, 2016). Remarkably, no other available phylogenetic tree opens the possibility that these four phyla are related; ii) Deltaproteobacteria (sulfur and sulfate reducing bacteria) are placed in the root of the bacterial ToL. It has been proposed that the deepest branches of the phylogenetic tree comprise chemolitotrophic autotrophs. The first autotrophic processes were akin to those now seen in the sulfur and sulfate reducing bacteria (Rosenberg, 2013).

Discussion

- Bacterial GlyRS is at the origin of all aaRSs and non-sequence specific RNA binding activities

Among aaRSs, bacGlyRS represents a unique $\alpha_2\beta_2$ oligomeric arrangement. One of the problems that we faced during the development of this project is that the α and β subunits are easily dissociated. The α-subunit of bacGlyRS is the only known case among aaRSs where the catalytic domain is not covalently attached to other domains. The structure shows the extent of this relative independence, as the α-subunit is exclusively responsible

of the tetrameric assembly and it forms the same, α_2 homodimer that can perform the first step of aminoacylation (Valencia-Sanchez et al., 2016). The α-subunit seems an entity that may have evolved unconstrained from the β-subunit. In contrast, the β-subunit may reflect the restraint coevolution with a substrate, the tRNA, that became bigger with time. The presence of two clearly different types of protein architectures (α and $\alpha+\beta$) in different, potential tRNA recognition stems on the β subunit, might also indicate a two-stage evolution scenario.

The ArgRS-like, bacGlyRS ACBD represents a connection between class I and II aaRSs. Besides bacGlyRS, no other class II aaRS has another key domain shared with class I aaRSs. Regarding a linkage with the genetic code (Figure 8), remarkably, according to the amino acid defining and most critical, second codon position of the code, the class I aaRSs specific for the amino acids belonging to the group U (IleRS, ValRS, LeuRS and MetRS) are clustered together, according to the structure-based sequence alignment of the ACBD (Supplementary Figure 6). Moreover, the ACBDs of bacGlyRS, ArgRS and CysRS are also clustered together, and their cognate amino acids belong to the same group (G) (Supplementary Figure 6, Figure 8).

Thus, although our evolutionary proposal (see below) implicates that the ACBD was a later addition to bacGlyRS, the ACBD incorporation should have happened very early during the development of the genetic code. The ACBD, widely spread in most class I aaRSs, probably represents the most ancient domain of this type (Figure 8).

Regarding the tail-body domains, the sequence similarity of the CCA-adding enzyme of class I with the equivalent domains of the bacGlyRS is so low that, at first glance, it is tempting to speculate that they are not homologs, but analogs. However, there are clear

signs of homology in this region: two structural units (domains), not only one, are involved in the comparison; the arrangement and the relative orientation of both domains is the same in both proteins; the connectivity of the secondary structure elements is the same; finally, both proteins share the insertion of large parts of secondary structures to form the body domain (PDB code 1R89). Remarkably, the homology with the tail-body domains is also present with proteins involved in eukaryotic ribosome processing (PDB code 4M5D) and thus, this region is present in the three life domains and represents a connection with general RNA metabolism.

As for the linker domain, although there are issues to find a clear homolog of it, we identified a βαββ topology that is also found in the body domain in tandem with the tail domain; this raises the possibility that a triplication event is the origin of the αβ architecture of bacGlyRS, creating a tandem repeat as it was required. As the sequence conservation between the tail-body-linker domains is extremely low, the structural motifs are highly degenerated and the domains are embellished with other structural motifs, this triplication, if true, must have taken place quite early in evolution. Notably, the βαββ topology found in these domains represents a ferredoxin-like fold, which is considered ancient and central to the origin of metabolism.

GlyRS is supposed to be one of the first aaRSs to evolve. But what motivated the appearance of two GlyRSs? The structure of bacGlyRS offers a possible explanation, where the α-subunit of bacGlyRS may have appeared as one of the first aaRSs (Table 2), even in a pre-LUCA world, where the αβ architecture of the β-subunit served as a multi-functional, helical-RNA binding/recognition module. Perhaps this recognition was based on size only, not sequence, as it is for the CCA-adding enzyme. As the genetic

code and the tRNA became more sophisticated, other domains, including the ACBD, were created or recruited in bacteria, and at this time another, full GlyRS, smaller and including both the catalytic and the anticodon recognition regions (eukGlyRS), rapidly evolved in archaea from ProRS, or another class IIa aaRS, that substituted the primordial, α-bacGlyRS and was inherited, post-LUCA, to some bacterial species through horizontal gene transfer (HGT). It is possible that archaea inherited from LUCA the primordial bacGlyRS to quickly lose it afterwards, leaving some domains of the β-subunit (i.e. the tail and body), useful for other enzymes (i.e. the CCA-adding).

The alternative hypothesis, a relatively recent appearance of the bacGlyRS, requiring the development of a completely novel, multidomain protein, or the HGT of the αβ architecture from archaea, with an atypical tRNA recognition, provided that the canonical tRNA interaction pattern was already present in class II aaRSs, and then the concomitant loss of the much shorter, single chain eukGlyRS, seems less likely on the current structural scenario.

Moreover, given the relationships of bacGlyRS with AlaRS (catalytic core and primordial aaRS) and ArgRS (ACBD), it is possible to map them on the genetic code table (Figure 8) and obtain a correlation between the emergence of these first putative aaRSs with the tRNA acceptor stem code (first code) and the anticodon stem code (second code). Remarkably, it is not possible to make such a relationship with any other aaRS, which further highlights the ancient nature of bacGlyRS.

Thus, regarding bacGlyRS, our work defines the structure of the full bacterial GlyRS $\alpha_2\beta_2$ heterotetramer, completely delineates the differences between the two types of GlyRS,

proposes a model of the tRNA recognition process, and provides insight that poses $\alpha_2\beta_2$ GlyRS at a landmark on the origin of aaRSs and other proteins involved in RNA metabolism, thus offering perhaps the first clear example of the coevolution of the Ribonucleoproteic (RNP) world. Further work is needed to corroborate the tRNA interactions and the chemical details of the aminoacylation reaction.

- The rooted bacterial Tree of Life (ToL) can be inferred from ancestral topologies

In general, performing a phylogenetic analysis with any aaRS is very complicated because the homologies with other aaRSs. That is, the problem of correctly resolving orthologs, paralogs and horizontal gene transfer (HGT) is very difficult to solve. bacGlyRS offers a unique window to analyze these ancient proteins, because due to its localized nature in bacterial species, the sequences selected correspond to unambiguous orthologous genes. There are several choices to perform a phylogenetic analysis on bacGlyRS: 1) on the α-subunit, catalytic domain; the interactions of aaRSs with the cognate amino acid and ATP occur in this domain. There are two different folds of this domain that define the two classes of aaRSs, I and II (Eriani et al., 1990). This class-defining domain is considered the ancient tRNA synthetase, and bacGlyRS has the shortest sequence among the class II aaRSs (Smith & Hartman, 2015). Thus, bacGlyRS catalytic domain is the simplest by far among aaRSs. Moreover, the helical repeats that cover it may represent primordial oligomerization topologies. In addition, the evolution rate of the whole α-subunit may be completely independent form the β-subunit. Indeed, the α-subunit is much more conserved; a signal that the β-subunit is more ancient. The β-subunit presents two types of multidomain organizations based on two of the most basal folds: α and $\alpha+\beta$, both of which may have originated from multiple duplication events

(Caetano-Anolles et al., 2009). Or we may choose to analyze 2) the α-helical, anticodon binding domain (ACBD). Regarding the interactions of aaRSs with tRNA, these are mainly done in specific ACBDs. There are three different ACBDs known, being the most expanded the α-helical domain of class I aaRSs. Interestingly, ArgRS, which ACBD is the closest related to bacGlyRS, is the most deeply rooted class I enzyme (Perona & Hadd, 2012). Option 3) the duplicated βαββ ancient protein topology, which in principle sits the ancient tRNA TpsiC-loop (this work and (Ju et al., 2021)), which sequence is highly conserved and defines the limit of the acceptor stem (reviewed in Tamaki, 2018, Frontier Genetics). We chose to perform our phylogenetic analysis using this topology, because we think that it represents an example of ancient domain accretion and of multiple domain repeats not identified by sequence. Ancient proteins described tend to be single domain, so we consider that the duplicated, sub-domain topology offers a better chance to unveil truly ancient evolutionary events. It should be noted that most known, currently accepted phylogenies are made with either a single or few genes from whole, multidomain genes/proteins (Philippe et al., 2011) or using phylogenomics approaches (Zhu et al., 2019), (Coleman et al., 2021). These approaches do not consider the different evolving rates of protein domains and, moreover, of protein domain segments. HGT and orthology determination biases are also common in bacterial phylogenies (Superson et al., 2019). With our proposed approach, we believe that several of these issues may be overcome. Surprisingly, the derived tree agrees very well with several available trees and other available evidence. It also supports previously suggested phyla relationships, and even places the root very close to Fusobacteria, recently proposed as the most ancestral phyla (Coleman et al., 2021). Thus, we believe that this classification is supported not only by

our data, but also from several, previously scattered pieces of evidence and we recommend dividing the bacteria domain into four superphylum according to the obtained tree (Figure 7): Desulfophyla, Firmicutes, Cyanobacteryota and Proteobacteria.

Overall, this work proposes (see Highlights in the Supplementary Material) that clear phylogenetic signal may be extracted from single, carefully chosen ancient protein fragments. It illustrates the coevolution of a ribonucleoprotein world, in which protein ancestral topologies were duplicated to generate multi-domain proteins able to bind promiscuous RNAs. It also represents a step towards deciphering the grammar of primordial domain combinations and how the very first proteins emerged.

Materials and methods

Refer also to Supplementary Methods

Crystallization and structure determination

The plasmid containing the coding sequence of αβTdGlyRS, with the Cys residues mutated to Ser to avoid protein aggregation and a TEV protease cleavage site-His_6 tag at the carboxyl terminus, was ordered from the company Atum DNA2.0, with the sequence optimized for production in *Escherichia coli*. The protein was purified by affinity to Ni^{2+} and a heat treatment at 72°C for 30 min. Crystals of abTdGlyRS were obtained by sitting-drop vapor diffusion. We obtained a highly reproducible crystallization condition, with very nice crystals that diffracted poorly (control condition; 320 mM ammonium phosphate, 1% dextran sodium sulfate Mr 4,000 (Molecular Dimensions) and 100 mM glycine pH 8.5). These crystals grew within 4 hours and we used them in a cross-seeding protocol to obtain crystals from a condition difficult to reproduce: we collected the crystals from 5 to 7 wells in 50 uL of mother liquor and crushed them by vortexing for 2 minutes

in the presence of a glass bead. We verified the crushing on the optical microscope and the seeds were diluted from 10 to up to 10 million fold. Finally, the crystallization test was performed by placing 1.5 µL of protein [36 mg/mL], 1 uL of the corresponding dilution and 1.5 uL of mother liquor (550 mM magnesium formate, 1% dextran sodium sulfate M_r 4,000, 100 mM glycine pH 8.5, 3% sucrose, 100 mM sodium acetate pH 4.6). The crystals were dehydrated for one day substituting the mother liquor with the cryoprotective solution, which was prepared by mixing 625 µL sucrose at 70%, 200 µL magnesium formate 1M, 50 µL sodium acetate 1M pH 4.6, 25 µL Gly 2 M pH 8.5 and 100 µL dextran sodium sulfate Mr 5,000 at 30%. Diffraction data were collected at 100 K using a wavelength of 0.9785 Å at the Life Sciences Collaborative Access Team (LS-CAT) 21-ID-D beamline at the Advanced Photon Source (Argonne National Laboratory, Argonne, Illinois, USA). Data were indexed and processed with xia2 (https://xia2.github.io/) that used XDS (Kabsch, 2010), Pointless (Evans, 2011) and reduced with Aimless (Evans & Murshudov, 2013). The structure was solved by molecular replacement using MrBUMP (Keegan & Winn, 2008), with a final search model based on the structure of the dimeric α-subunit bacGlyRS from *Aquifex aeolicus* (PDB entry 5F5W). A dimer is presented in the asymmetric unit, with a solvent percentage in the crystals of 75%. The initial model was obtained using the Phenix.Autobuild program of the PHENIX package (Liebschner et al., 2019) and was displayed and manually built in COOT (Emsley & Cowtan, 2004). Refinement was alternated with manual building/refinement in COOT, PHENIX and the PDB_REDO server (Joosten et al., 2014). Figures were prepared with Pymol (The PyMOL Molecular Graphics System, Version 2.1, Schrödinger, LLC).

Structural and sequence analysis

We used a combination of the outputs of both the Dali Server (http://ekhidna2.biocenter.helsinki.fi/dali/ (Holm & Laakso, 2016) and DASH (Database of Aligned Structural Homologs https://sysimm.org/dash/) (Rozewicki et al., 2019)) to look for structural homologs of the five domains of the bacGlyRS β-subunit. We prepared a collection of the trimmed models (PDB codes) comprising only the relevant possible homolog domains and we calculated an overall multiple structural alignment using the STAMP algorithm (Russell & Barton, 1992) as implemented in MultiSeq in VMD (Roberts et al., 2006). The aligned structures were used in Bio3D (Grant et al., 2006) to define an invariant core with the minimal structural variance among all the protein structures. This core of Cα atoms was the basis for the RMSD matrix (plotted in BioVinci), dendrogram calculation and Principal Component Analysis (PCA). The structures were superimposed onto the core and the variances shown by PC1 and PC2 were used to define the different subgroups.

For the cases of the body-tail domains of bacGlyRS and archaeal CCA-adding enzyme, we performed structure-based multiple sequence alignments. All the sequences obtained from a BLAST Protein search on the refseq_select database of the body-tail domains were analyzed. The output sequences were aligned with the ClustalW plug-in in MultiSeq. A profile alignment was calculated considering the structural alignment of corresponding domains of the models (PDB codes 1R89, residues 277-379 and 7LU4, residues 316-432). For tree calculation we used IQ-TREE 2 (Minh et al., 2020), which is suggested to find the highest likelihood scores among ML trees (Zhou et al., 2018). The sequences, log-file with the parameters used, the derived tree and the file used to obtain the color code figure in iToL (Letunic & Bork, 2021) are available in the Supplementary Material.

Figures

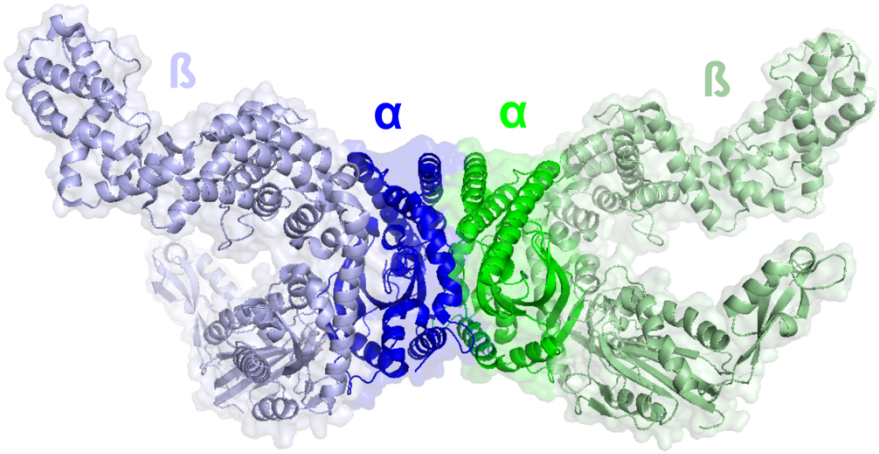

Figure 1. bacGlyRS is a βααβ complex formed around a central α_2 dimer. Each α subunit (300 aa) is flanked by one β subunit (700 aa), making bacGlyRS the biggest known aaRS of all three life domains (see Supplementary Discussion). Remarkably, the α-subunit represents the smallest known class II catalytic domain, and thus the β-subunit, which is easily dissociated *in vitro*, provides most of the molecular size of the enzyme.

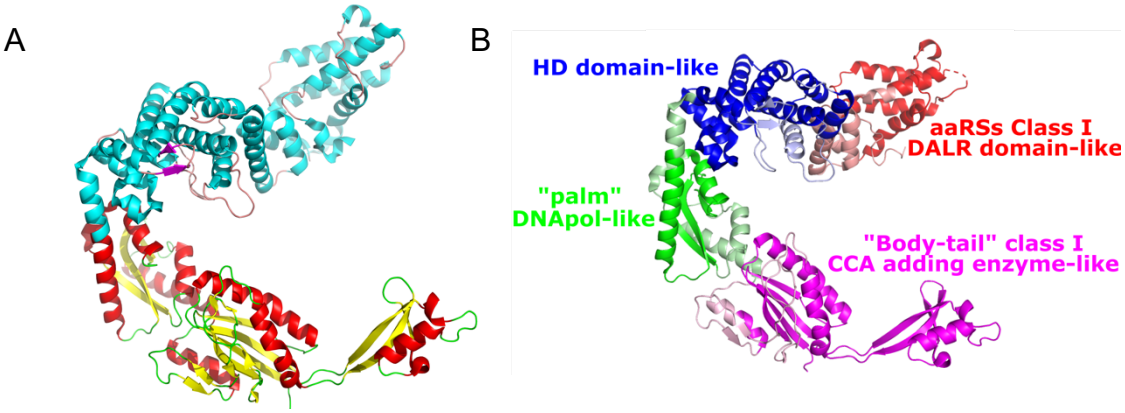

Figure 2. bacGlyRS β-subunit possesses two types of architecture, each one with domains homologous to tRNA binding proteins specific for the acceptor (ancient) and anticodon stems. (**A**) The two architectures of the bacGlyRS β-subunit. In cyan are depicted an almost all α-helical region that defines the HD phosphorylase-like domain and the tRNA anticodon binding domain of class I aaRSs (ACBD or DALR), these domains are shown in blue and red in **B,** respectively. In yellow and red is an α+β region that comprises three domains: a linker domain (green in **B**) with no clear homology to any protein, although the closest structural match is with the palm domain of DNA polymerases, and two domains with clear homology to class I, tRNA CCA adding enzymes from archaea (pink in **B**). (**B**) The five identified domains in the bacGlyRS β-subunit shown in different colors.

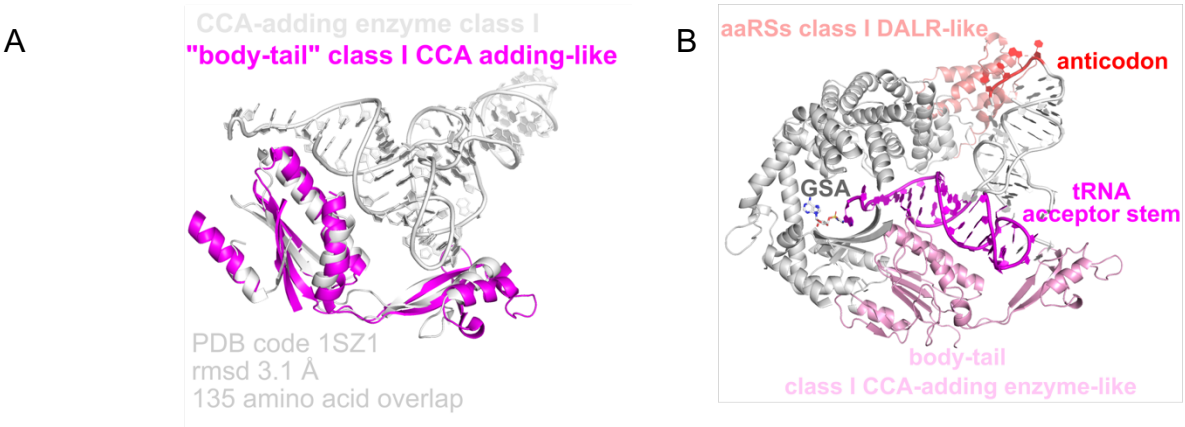

Figure 3. The archaeal CCA-adding enzyme provides a template to build a clash-free model of the bacGlyRS:tRNA:GSA ternary complex. A. Superposition of the body-tail domains of the CCA-adding enzyme:tRNA complex (white) on the equivalent domains of bacGlyRS (magenta). **B.** Resulting model of the superposition at left but depicted on the whole monomer of bacGlyRS and using the α-subunit model (PDB code 5F5W) to depict the GSA (glycyl-sulfamoyl-adenylate) molecule on the catalytic cavity. The tRNA model is within 6 Å reach of all three regions (body-tail, ACBD and GSA) and thus fits perfectly on bacGlyRS.

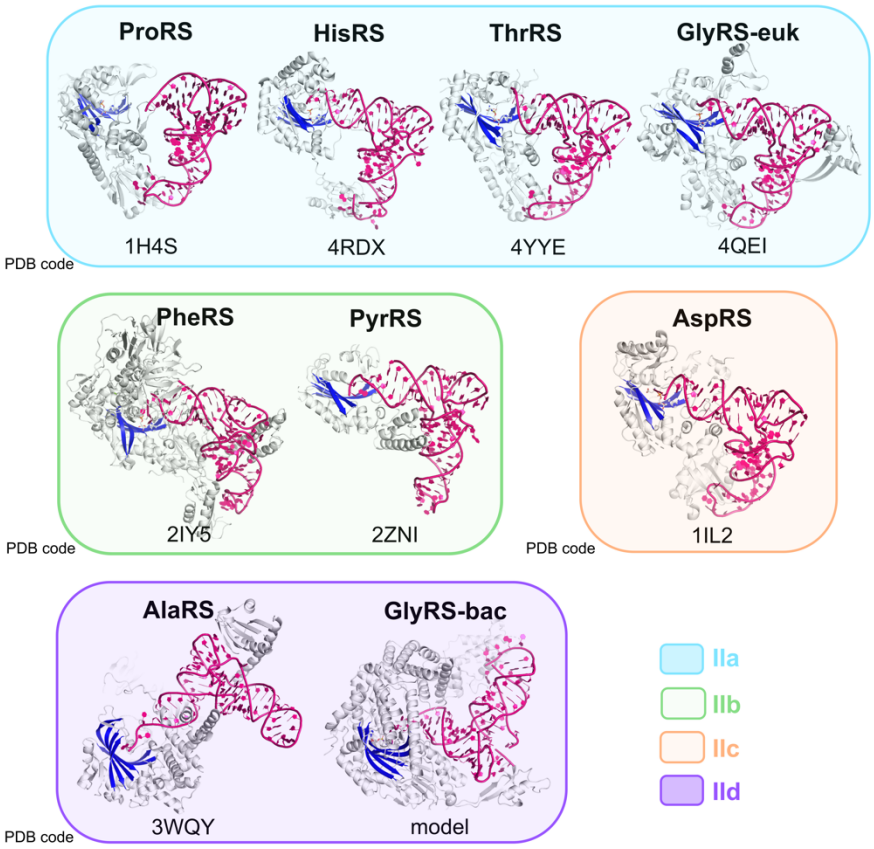

Figure 4. bacGlyRS and AlaRS clearly define a subclass among class II aaRS. Besides their remarkable similarities at the catalytic domain (Valencia-Sánchez, 2016), bacGlyRS and AlaRS (magenta, subclass IId) would recognize their cognate tRNA in a completely different orientation with respect to all other class II aaRSs. The relative orientation of the catalytic domain (depicted in part in blue) is the same for all models.

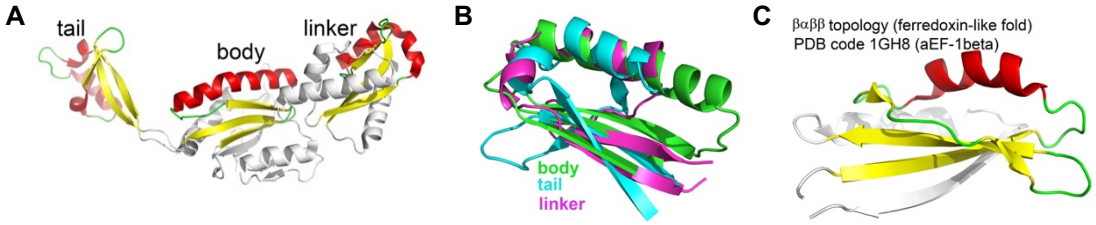

Figure 5. The body-tail-linker domains of bacGlyRS β-subunit may be the result of a triplication event. A. The linker, body and tail domains of bacGlyRS β-subunit show a repeated βαββ topology that can be roughly superposed, as shown in **B.** This topology has homology with the ancient and widespread ferredoxin fold, exemplified in **C** by the PDB code 1GH8 that represents an archaeal elongation factor. This is the closest homology of the linker domain with any known protein. That is, it is more similar to other segments of the same chain than to other proteins. Regarding the body domain, it is remarkable that the protein chain traces the βαββ topologies of the body and tail domains and then forms the rest of the body domain with large parts of secondary structures (Supplementary Figure 8) (Arumugam et al., 2013). We also may refer to the body-tail topology as a protein tandem repeat of class V (Kajava, 2012).

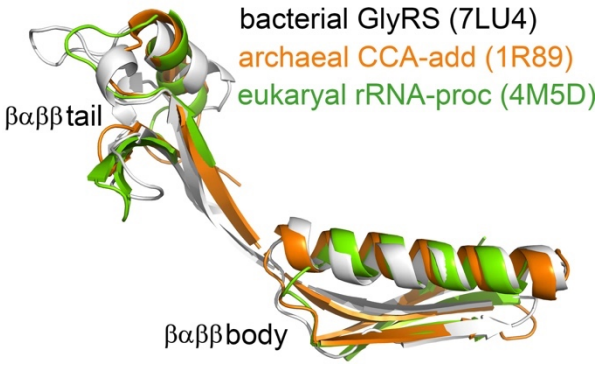

Figure 6. The duplicated βαββ, ferredoxin-like, primitive RNA binding topology present in the three life domains provides a clear coevolution signal of an ancient RNP world. Superposition of the duplicated βαββ topologies of bacGlyRS, archCCAadd and eukrRNAproc, all representing ancient protein topologies binding to ancient RNA motifs from the three domains of life. The homology between bacGlyRS and archCCAadd can be marginally detected by sequence similarity, which is around 10%. No other tandem topology homolog to this one has been found in nature, which shows a limited horizontal gene transfer (HGT). The fact that this is a duplicated, not single topology, reassures its uniqueness and highlights its phylogenetic signal.

Figure 7. The ancestral, duplicated βαββ, ferredoxin-like topologies of archaea-CCAadding and bacterial β-GlyRS draw the rooted bacterial Tree of Life. The expected limited Horizontal Gene Transfer (HGT) of the duplicated βαββ topology of bacGlyRS led to this proposition, rooted on the archaeal βαββ topology of archaeal CCA-adding enzyme. Several deeply rooted phyla belong to what was previously known as Deltaproteobacteria. We propose four superphylum, with Desulfophyla as the root. Fusobacterium, recently proposed to be close to the root, remains within this group. The tree was calculated with IQ-TREE v.2.1.3. (Minh et al., 2020), with no sequence outliers in the composition test. The diagram was drawn with iTOL. The tree and the sequences from which it was derived are available in the Supplementary Material.

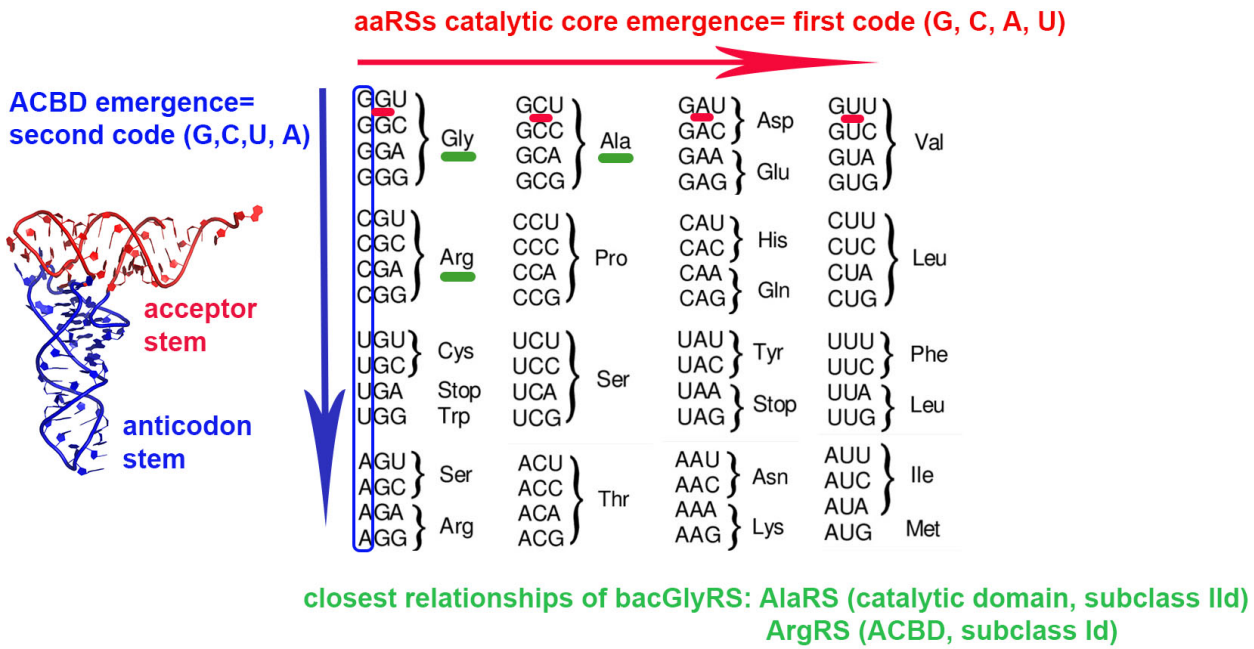

Figure 8. The genetic code table reflects the age of primordial aaRSs and their evolutionary reading of the genetic code (second base first, first base second). The "first, or operational code" (Schimmel et al., 1993) corresponds to the second letter of the codon (Rodin & Rodin, 2006), and it is suggested to lie on the acceptor stem of the tRNA (ancient). A complex between an aaRS catalytic core and the βαββ body-tail domains may have carried out the primitive aminoacylation task. The "second code" corresponds to the first letter of the codon and would lie on the anticodon stem of the tRNA. This proposal is in agreement with an almost concomitant appearance of class I/II aaRSs and with previous hypothesis on the emergence of aaRSs (Smith & Hartman, 2015). The first, primitive code, could have been specific for Gly, Ala, Arg and Pro.

Tables

Table 1 Data collection and refinement statistics

	αβ-TdGlyRS-Gly
Data collection	
Space group	C2
Cell dimensions	
a, *b*, *c* (Å)	238.58, 127.86, 133.06
α, *β*, *γ* (°)	90.0, 103.6, 90.0
Resolution (Å)	37.62-2.50 (2.54-2.50) *
R_{sym} or R_{merge}	0.126 (3.32)
$I/\sigma I$	14.6 (1.5)
Completeness (%)	99.2 (99.1)
Redundancy	17.1 (15.8)
Mean (I) half-set correlation $CC_{(1/2)}$	0.999 (0.452)
Wilson B-factor ($Å^2$)	72.9
Refinement	
Resolution (Å)	35.74-2.50 (2.53-2.50)
No. reflections	132944 (4301)
R_{work}/ R_{free}	0.207/0.228 (0.363/0.406)
No. atoms	
Protein	14710
Ligand/ion	5
B-factors	
Protein	81.0
Ligand/ion	76.7
R.m.s deviations	
Bond lengths (Å)	0.004
Bond angles (º)	0.586

One crystal was used.
*Highest resolution shell is shown in parentheses.

Table 2. bacGlyRS as an ancestral protein

Feature	Evidence
Coevolution with the primordial code (second codon position appeared first-coded in acceptor stem, first codon position appeared second-coded in anticodon).	α: Catalytic domain groups with AlaRS (subclass IId); Ala, as Gly, is a key aa defined by the second codon position (Ala, Gly, Asp, Val are coded by the primordial code C, G, A, U). β: ACBD groups with ArgRS and CysRS, corresponding to amino acids coded by G on the second codon position and C/A, U on the first codon position..
tRNA age (acceptor stem first).	Triplicated βαββ (ancient, ferredoxin-like protein topology) with recognition motif for tRNA acceptor, the most ancient tRNA domain.
tRNA size (gene duplication event to generate AC stem).	β-subunit seems to have grown as the tRNA became larger (first architecture: α+β for acceptor; then α for anticodon).
Independent subunits/function (aa-specific and tRNA-promiscuous binding may have evolved with different constraints/timelines).	βααβ subunit organization (central α_2 homodimer, specific aa-tRNA recognition; β, tRNA recognition by size, not sequence-specific).
Multifunctionality (in LUCA, a few proteins may have performed several related functions).	Body-tail RNA recognition module present in three different functions (aminoacylation, CCA polymerization and ribosomal processing).
Widespread protein domains (the more widespread in organisms and functions, the more ancient).	Present in the three domains of life: α-Core of catalytic domain (all class II aaRSs-catalysis) β-ACBD (seven class I aaRSs-binding) β-Body-tail (three RNA processing enzymes-binding) β-HD-like domain (phosphohydrolases-catalysis)
Homology (if the homologs are not easily detected, then the protein may be more ancient).	All domains have borderline sequence similarities (8-13%) with their closest structural neighbors.
Non-canonical universal aaRS elements (the less canonical, the less recent).	α: atypical sequence of motifs 1 and 2 on the catalytic domain, leading to non-canonical α-α dimer stabilization and amino group recognition of the cognate amino acid. β: atypical orientation of ACBD relative to the catalytic domain, leading to a putative inverted tRNA orientation with respect to aaRSs of class II. Hybrid; includes key elements of both classes of aaRSs; β: class I AC tRNA recognition; α: class II aa recognition.
Ancient topology repeats or peptide tandem repeats (proteins are thought to have originated from peptide/motif repeats).	Six α-helices in α-subunit (top of catalytic domain; ensemble for monomer:monomer interactions). Sixteen α-helices in β-subunit (HD-like domain and ACBD). Three βαββ motifs in β-subunit (body-tail-linker domains).

Table 3. Name and classification of bacGlyRS domains

subunit	name	CATH topology	CATH superfamily	SCOPe fold	SCOPe family	Pfam family	ECOD/PASS2.6
α 188-286	oligom module	Up-down bundle	1.20.58.180 (PDB 5f5w)			PF02091	
β 313-372 436-523	body	Alpha-beta plaits	3.30.70.590 (PDB 1r89)	Ferredoxin-like	d.58.16.2		X: alpha-beta plaits H: (4m5d, 6rxv (U3)), (1q79, 2q66 (polyA)), (1r89 (CCA-add))
β 378-432	tail	Alpha-beta plaits	3.30.70.1550 (PDB 1r89)	Ferredoxin-like	d.58.16.2		
β 526-641	linker	Alpha-beta plaits	3.30.70.60 (PDB 2yy3)				
β 647-774	HD-like	Orthogonal bundle	1.10.3210.10 (PBD 4mcw)	HDdomain/ PDEase-like	a.211.1.1	PF01966	Alpha complex topology. X: PDEase-like
β 833-992	ACBD	Orthogonal bundle	1.10.730.10 (PDB 1bs2)	ACBD aaRSs class I	a.27.1.1	PF05746	Alpha bundles X: ACBD class I

Data availability

The atomic coordinates and structure factors of the αβ, bacterial glycyl tRNA synthetase from *Thermanaerothrix daxensis* in the complex with glycine have been deposited in the Protein Data Bank (PDB) with the accession code 7LU4.

Acknowledgements

Dino Moras (IGBMC, France) wishes to make a full endorsement of the results of this paper and the resulting conclusions. A.T.L. was supported by Fundación Marcos Moshinsky, Consejo Nacional de Ciencia y Tecnología (CONACyT), grants CB2015-253323, PDCPN2014-247543 (2012-2018 Government Administration) and Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica, Universidad Nacional Autónoma de México (PAPIIT-DGAPA) grants IN202416 and IN204820. D.E.R.CH. and J.U.D.T received a grant from CONACyT to perform graduate studies and data in this work is part of his PhD and MSc dissertation, respectively, in the Posgrado en Ciencias Bioquímicas, Universidad Nacional Autónoma de México. Use of the Advanced Photon

Source, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science by Argonne National Laboratory, was supported by the U.S. DOE under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor, Grant 085P1000817. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work used resources of the Northwestern University Structural Biology facility, which is generously supported by NCI CCSG P30 CA060553 grant awarded to the Robert H. Lurie Comprehensive Cancer Center. A.T.L dedicates this work to Elizabeth Nallely Cabrera González (E.N.C.G), Prof. Alfonso Mondragón, Prof. Ruy Pérez-Montfort, Prof. Anne-Catherine Dock-Bregeon and Prof. Enrique Merino-Pérez, IBT-UNAM, for their constant scientific support and guidance. The authors thank Zdzislaw Wawrzak (APS) for help/assistance with the crystallographic aspect of this project, E.N.C.G., Pamela J. Focia and Valerie Tokars for technical guidance, Prof. John J Perona for the gift of the gene coding for the tRNA nucleotidyl transferase and Georgina E. Espinosa Pérez, MSc, for assistance in crystallographic data collection.

Author contributions

A.T.L, A.R.H., M.I.V.S and D.E.R.CH conceived the project. A.T. L. conceived and designed the experiments. J.U.D.T and V.G.L. purified and crystallized TdGlyRS. A.R.H performed the activity experiments. A.T.L. collected crystallographic data and solved the structure of TdGlyRS. E.C.CH. performed and analyzed the EM negative staining experiments. M.G. and C.F. synthesized the GSA compound. E.C.CH. and V.G.L. performed the affinity measurements. M.A., D.E.R.CH., A.H.G. and A.T.L. performed the

refinement of the model and the structural and sequence analysis. A.T.L. and A.R.H. wrote the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

References

Arumugam, G., Nair, A. G., Hariharaputran, S., & Ramanathan, S. (2013). Rebelling for a reason: protein structural "outliers". *PLoS One, 8*(9), e74416. https://doi.org/10.1371/journal.pone.0074416

Ball, S. G., Colleoni, C., Kadouche, D., Ducatez, M., Arias, M. C., & Tirtiaux, C. (2015, Jun-Jul). Toward an understanding of the function of Chlamydiales in plastid endosymbiosis. *Biochim Biophys Acta, 1847*(6-7), 495-504. https://doi.org/10.1016/j.bbabio.2015.02.007

Brinkman, F. S., Blanchard, J. L., Cherkasov, A., Av-Gay, Y., Brunham, R. C., Fernandez, R. C., Finlay, B. B., Otto, S. P., Ouellette, B. F., Keeling, P. J., Rose, A. M., Hancock, R. E., Jones, S. J., & Greberg, H. (2002, Aug). Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. *Genome Res, 12*(8), 1159-1167. https://doi.org/10.1101/gr.341802

Caetano-Anolles, G., Wang, M., Caetano-Anolles, D., & Mittenthal, J. E. (2009, Feb 1). The origin, evolution and structure of the protein world. *Biochem J, 417*(3), 621-637. https://doi.org/10.1042/BJ20082063

Carter, C. W., Jr., & Wolfenden, R. (2015, Jun 16). tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proc Natl Acad Sci U S A, 112*(24), 7489-7494. https://doi.org/10.1073/pnas.1507569112

Coleman, G. A., Davin, A. A., Mahendrarajah, T. A., Szantho, L. L., Spang, A., Hugenholtz, P., Szollosi, G. J., & Williams, T. A. (2021, May 7). A rooted phylogeny resolves early bacterial evolution. *Science, 372*(6542). https://doi.org/10.1126/science.abe0511

Domman, D., Horn, M., Embley, T. M., & Williams, T. A. (2015, Mar 11). Plastid establishment did not require a chlamydial partner. *Nat Commun, 6*, 6421. https://doi.org/10.1038/ncomms7421

Emsley, P., & Cowtan, K. (2004, Dec). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr, 60*(Pt 12 Pt 1), 2126-2132. https://doi.org/10.1107/S0907444904019158

Eriani, G., Delarue, M., Poch, O., Gangloff, J., & Moras, D. (1990, Sep 13). Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature, 347*(6289), 203-206. https://doi.org/10.1038/347203a0

Evans, P. R. (2011, Apr). An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr, 67*(Pt 4), 282-292. https://doi.org/10.1107/S090744491003982X

Evans, P. R., & Murshudov, G. N. (2013, Jul). How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr, 69*(Pt 7), 1204-1214. https://doi.org/10.1107/S0907444913000061

Fournier, G. P., Andam, C. P., & Gogarten, J. P. (2015, Apr 22). Ancient horizontal gene transfer and the last common ancestors. *BMC Evol Biol, 15*, 70. https://doi.org/10.1186/s12862-015-0350-0

Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A., & Caves, L. S. (2006, Nov 1). Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics, 22*(21), 2695-2696. https://doi.org/10.1093/bioinformatics/btl461

Gregoire, P., Fardeau, M. L., Joseph, M., Guasco, S., Hamaide, F., Biasutti, S., Michotey, V., Bonin, P., & Ollivier, B. (2011, Nov). Isolation and characterization of Thermanaerothrix daxensis gen. nov., sp. nov., a thermophilic anaerobic bacterium pertaining to the phylum "Chloroflexi", isolated from a deep hot aquifer in the Aquitaine Basin. *Syst Appl Microbiol, 34*(7), 494-497. https://doi.org/10.1016/j.syapm.2011.02.004

Gupta, R. S., & Khadka, B. (2016, Feb). Evidence for the presence of key chlorophyll-biosynthesis-related proteins in the genus Rubrobacter (Phylum Actinobacteria) and its implications for the evolution and origin of photosynthesis. *Photosynth Res, 127*(2), 201-218. https://doi.org/10.1007/s11120-015-0177-y

Holm, L., & Laakso, L. M. (2016, Jul 8). Dali server update. *Nucleic Acids Res, 44*(W1), W351-355. https://doi.org/10.1093/nar/gkw357

Imhoff, J. F., Rahn, T., Kunzel, S., & Neulinger, S. C. (2017). Photosynthesis Is Widely Distributed among Proteobacteria as Demonstrated by the Phylogeny of PufLM Reaction Center Proteins. *Front Microbiol, 8*, 2679. https://doi.org/10.3389/fmicb.2017.02679

Joosten, R. P., Long, F., Murshudov, G. N., & Perrakis, A. (2014, Jul 1). The PDB_REDO server for macromolecular structure model optimization. *IUCrJ, 1*(Pt 4), 213-220. https://doi.org/10.1107/S2052252514009324

Ju, Y., Han, L., Chen, B., Luo, Z., Gu, Q., Xu, J., Yang, X. L., Schimmel, P., & Zhou, H. (2021, Aug 14). X-shaped structure of bacterial heterotetrameric tRNA synthetase suggests cryptic prokaryote functions and a rationale for synthetase classifications. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkab707

Kabsch, W. (2010, Feb). Xds. *Acta Crystallogr D Biol Crystallogr, 66*(Pt 2), 125-132. https://doi.org/10.1107/S0907444909047337

Kajava, A. V. (2012, Sep). Tandem repeats in proteins: from sequence to structure. *J Struct Biol, 179*(3), 279-288. https://doi.org/10.1016/j.jsb.2011.08.009

Keegan, R. M., & Winn, M. D. (2008, Jan). MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallogr D Biol Crystallogr, 64*(Pt 1), 119-124. https://doi.org/10.1107/S0907444907037195

Kim, J. D., Rodriguez-Granillo, A., Case, D. A., Nanda, V., & Falkowski, P. G. (2012). Energetic selection of topology in ferredoxins. *PLoS Comput Biol, 8*(4), e1002463. https://doi.org/10.1371/journal.pcbi.1002463

Koonin, E. V., & Novozhilov, A. S. (2017, Nov 27). Origin and Evolution of the Universal Genetic Code. *Annu Rev Genet, 51*, 45-62. https://doi.org/10.1146/annurev-genet-120116-024713

Koski, M. K., Anantharajan, J., Kursula, P., Dhavala, P., Murthy, A. V., Bergmann, U., Myllyharju, J., & Wierenga, R. K. (2017, Feb 20). Assembly of the elongated collagen prolyl 4-hydroxylase alpha2beta2 heterotetramer around a central alpha2 dimer. *Biochem J, 474*(5), 751-769. https://doi.org/10.1042/BCJ20161000

Letunic, I., & Bork, P. (2021, Jul 2). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res, 49*(W1), W293-W296. https://doi.org/10.1093/nar/gkab301

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J., & Adams, P. D. (2019, Oct 1). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol, 75*(Pt 10), 861-877. https://doi.org/10.1107/S2059798319011471

Lin, Y. R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A. F., Montelione, G. T., & Baker, D. (2015, Oct 6). Control over overall shape and size in de novo designed proteins. *Proceedings of the National Academy of Sciences of the United States of America, 112*(40), E5478-E5485. https://doi.org/10.1073/pnas.1509508112

Long, X., Xue, H., & Wong, J. T. (2020). Descent of Bacteria and Eukarya From an Archaeal Root of Life. *Evol Bioinform Online, 16*, 1176934320908267. https://doi.org/10.1177/1176934320908267

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020, May 1). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol, 37*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015

Ostrem, D. L., & Berg, P. (1974, Mar 26). Glycyl transfer ribonucleic acid synthetase from Escherichia coli: purification, properties, and substrate binding. *Biochemistry, 13*(7), 1338-1348. https://doi.org/10.1021/bi00704a006

Perona, J. J., & Hadd, A. (2012, Nov 6). Structural diversity and protein engineering of the aminoacyl-tRNA synthetases. *Biochemistry, 51*(44), 8705-8729. https://doi.org/10.1021/bi301180x

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T., Manuel, M., Worheide, G., & Baurain, D. (2011, Mar). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol, 9*(3), e1000602. https://doi.org/10.1371/journal.pbio.1000602

Raanan, H., Poudel, S., Pike, D. H., Nanda, V., & Falkowski, P. G. (2020, Mar 31). Small protein folds at the root of an ancient metabolic network. *Proc Natl Acad Sci U S A, 117*(13), 7193-7199. https://doi.org/10.1073/pnas.1914982117

Raymond, J. (2008, Feb). Coloring in the tree of life. *Trends Microbiol, 16*(2), 41-43. https://doi.org/10.1016/j.tim.2007.11.003

Rice, P. A., & Correll, C. C. (2008). *Protein-nucleic acid interactions : structural biology*. RSC Pub.

Roberts, E., Eargle, J., Wright, D., & Luthey-Schulten, Z. (2006, Aug 16). MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics, 7*, 382. https://doi.org/10.1186/1471-2105-7-382

Rodin, S. N., & Rodin, A. S. (2006, Jun). Origin of the genetic code: first aminoacyl-tRNA synthetases could replace isofunctional ribozymes when only the second base of codons was established. *DNA Cell Biol, 25*(6), 365-375. https://doi.org/10.1089/dna.2006.25.365

Rosenberg, E. (2013). *The prokaryotes -prokaryotic physiology and biochemistry : prokaryotic physiology and biochemistry* (4th ed.). Springer. Contributor biographical information http://www.loc.gov/catdir/enhancements/fy1317/2012955034-b.html
Publisher description http://www.loc.gov/catdir/enhancements/fy1317/2012955034-d.html
Table of contents only http://www.loc.gov/catdir/enhancements/fy1317/2012955034-t.html

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019, Jul 2). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res, 47*(W1), W5-W10. https://doi.org/10.1093/nar/gkz342

Russell, R. B., & Barton, G. J. (1992, Oct). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins, 14*(2), 309-323. https://doi.org/10.1002/prot.340140216

Schimmel, P. (2011, Mar). The RNP bridge between two worlds. *Nat Rev Mol Cell Biol, 12*(3), 135. https://doi.org/10.1038/nrm3061

Schimmel, P., Giege, R., Moras, D., & Yokoyama, S. (1993, Oct 1). An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci U S A, 90*(19), 8763-8768. https://doi.org/10.1073/pnas.90.19.8763

Smith, T. F., & Hartman, H. (2015, Nov 30). The evolution of Class II Aminoacyl-tRNA synthetases and the first code. *FEBS Lett, 589*(23), 3499-3507. https://doi.org/10.1016/j.febslet.2015.10.006

Superson, A. A., Phelan, D., Dekovich, A., & Battistuzzi, F. U. (2019, Oct 1). Choice of species affects phylogenetic stability of deep nodes: an empirical example in Terrabacteria. *Bioinformatics, 35*(19), 3608-3616. https://doi.org/10.1093/bioinformatics/btz121

Taib, N., Megrian, D., Witwinowski, J., Adam, P., Poppleton, D., Borrel, G., Beloin, C., & Gribaldo, S. (2020, Dec). Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol, 4*(12), 1661-1672. https://doi.org/10.1038/s41559-020-01299-7

Tang, S. N., & Huang, J. F. (2005, Feb 28). Evolution of different oligomeric glycyl-tRNA synthetases. *FEBS Lett, 579*(6), 1441-1445. https://doi.org/10.1016/j.febslet.2005.01.045

Valencia-Sanchez, M. I., Rodriguez-Hernandez, A., Ferreira, R., Santamaria-Suarez, H. A., Arciniega, M., Dock-Bregeon, A. C., Moras, D., Beinsteiner, B., Mertens, H., Svergun, D., Brieba, L. G., Grotli, M., & Torres-Larios, A. (2016, Jul 8). Structural

Insights into the Polyphyletic Origins of Glycyl tRNA Synthetases. *J Biol Chem, 291*(28), 14430-14446. https://doi.org/10.1074/jbc.M116.730382

Waite, D. W., Chuvochina, M., Pelikan, C., Parks, D. H., Yilmaz, P., Wagner, M., Loy, A., Naganuma, T., Nakai, R., Whitman, W. B., Hahn, M. W., Kuever, J., & Hugenholtz, P. (2020, Nov). Proposal to reclassify the proteobacterial classes Deltaproteobacteria and Oligoflexia, and the phylum Thermodesulfobacteria into four phyla reflecting major functional capabilities. *Int J Syst Evol Microbiol, 70*(11), 5972-6016. https://doi.org/10.1099/ijsem.0.004213

Waite, D. W., Vanwonterghem, I., Rinke, C., Parks, D. H., Zhang, Y., Takai, K., Sievert, S. M., Simon, J., Campbell, B. J., Hanson, T. E., Woyke, T., Klotz, M. G., & Hugenholtz, P. (2017). Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed Reclassification to Epsilonbacteraeota (phyl. nov.). *Front Microbiol, 8*, 682. https://doi.org/10.3389/fmicb.2017.00682

Wang, S., & Luo, H. (2021, Jun 3). Dating Alphaproteobacteria evolution with eukaryotic fossils. *Nat Commun, 12*(1), 3324. https://doi.org/10.1038/s41467-021-23645-4

Ward, L. M., Hemp, J., Shih, P. M., McGlynn, S. E., & Fischer, W. W. (2018). Evolution of Phototrophy in the Chloroflexi Phylum Driven by Horizontal Gene Transfer. *Front Microbiol, 9*, 260. https://doi.org/10.3389/fmicb.2018.00260

Weiner, A. M., & Maizels, N. (1987, Nov). tRNA-like structures tag the 3' ends of genomic RNA molecules for replication: implications for the origin of protein synthesis. *Proc Natl Acad Sci U S A, 84*(21), 7383-7387. https://doi.org/10.1073/pnas.84.21.7383

Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V., & Martin, W. F. (2018, Aug). The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet, 14*(8), e1007518. https://doi.org/10.1371/journal.pgen.1007518

Wolf, Y. I., Brenner, S. E., Bash, P. A., & Koonin, E. V. (1999, Jan). Distribution of protein folds in the three superkingdoms of life. *Genome Res, 9*(1), 17-26. https://www.ncbi.nlm.nih.gov/pubmed/9927481

Zeng, L., & Dehesh, K. (2021, Feb 26). The eukaryotic MEP-pathway genes are evolutionarily conserved and originated from Chlaymidia and cyanobacteria. *BMC Genomics, 22*(1), 137. https://doi.org/10.1186/s12864-021-07448-x

Zhaxybayeva, O., Swithers, K. S., Lapierre, P., Fournier, G. P., Bickhart, D. M., DeBoy, R. T., Nelson, K. E., Nesbo, C. L., Doolittle, W. F., Gogarten, J. P., & Noll, K. M. (2009, Apr 7). On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc Natl Acad Sci U S A, 106*(14), 5865-5870. https://doi.org/10.1073/pnas.0901260106

Zhou, X., Shen, X. X., Hittinger, C. T., & Rokas, A. (2018, Feb 1). Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. *Mol Biol Evol, 35*(2), 486-503. https://doi.org/10.1093/molbev/msx302

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., Al-Ghalith, G. A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J. B., Huang, S., Salam, N., Jiao, J. Y., Wu, Z., Xu, Z. Z., Cantrell, K., Yang, Y., Sayyari, E., Rabiee, M., Morton, J. T., Podell, S., Knights, D., Li, W. J., Huttenhower, C., Segata, N., Smarr, L., Mirarab, S., & Knight, R. (2019, Dec 2). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun, 10*(1), 5477. https://doi.org/10.1038/s41467-019-13443-4