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], 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.

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

Many experimental and predicted observations remain unanswered in the current proposed trees of life (ToL). Also, the current trend in reporting phylogenetic data is based in mixing together the information of dozens of genomes or entire conserved proteins. In this work, we consider the modularity of protein evolution and, using only two domains with duplicated ancestral topologies from a single, universal primordial protein corresponding to the RNA binding regions of contemporary bacterial glycyl tRNA synthetase (bacGlyRS), archaeal CCA adding enzyme (arch-CCAadd) and eukaryotic rRNA processing enzyme (euk-rRNA), we propose a rooted bacterial ToL that agrees with several previous observations 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. 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 most bacteria and chloroplasts. Due to its distribution in the three life domains, it is believed that eukGlyRS is the ancestral type that was present in LUCA. However, bacGlyRS is found in bacteria that are deeply rooted, like *Thermotoga* and *Cyanobacteria* (Supplementary Figure 1). Therefore, the

questions regarding why there are two GlyRSs and which one may have appeared first remain unanswered. bacGlyRS is the last major aaRS whose structure was still lacking to complete the collection of aaRSs towards further insights into the molecular evolution of these enzymes, their protein domains, and the origins of the genetic code.

Here we show the three-dimensional structure of a functional, complete bacGlyRS from the thermophilic bacteria *Thermanaerotrix daxensis* (Chloroflexi) 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 archael CCA-adding enzymes-eukaryotic rRNA processing complex, metal-dependent HD phosphohydrolases and the anticodon binding domain (ACDB) 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-Sánchez 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. 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 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 with respect to the structure of the α_2 homodimer alone. 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-Sánchez et al., 2016), shows that the

binding of the amino acid alone is enough to promote this movement and that this 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 and Wierenga, Biochemical Journal, 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 domains is shared between the subclass IId members, bacGlyRS and AlaRS, restraining the homology between these enzymes to the features shared in the catalytic domain (Valencia-Sánchez, 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 them 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) 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 take place in the minor groove of this substrate; that is, it would be a class I-like type of recognition on a class II aaRS. 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. 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): 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), which is of 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, as well as the tRNA acceptor stem.

- 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 babb, ferredoxin-like topologies of archaea-CCAadding and bacterial b-GlyRS draw the rooted bacterial Tree of Life.

With the hypothesis described above in hand, we calculated a phylogenetic tree using the sequences of the available βαββ duplicated topologies of archaea and bacteria domains (Figure 7). Notably, we could use the Archaeal sequences to obtain a rooted tree. Surprisingly, we found that the obtained phylogeny yields information that is very consistent with the widely accepted taxonomy. However, we found several interesting observations:

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-Sánchez 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 subutin, might also indicate a two-stage evolution scenario.

The ACDB represents a bond between class I and II aaRSs. Besides bacGlyRS, no other aaRS has another key domain shared with class I aaRSs. Regarding a linkage with the genetic code, remarkably, according to the amino acid defining and most critical, second codon position of the code, the class I aaRSs that code 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 ACDB (Supplementary Figure x). 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 x).

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

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 bac-GlyRS, 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.

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 providing 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

We propose that clear phylogenetic signal may be extracted from single, carefully chosen ancient protein fragments.

Figures

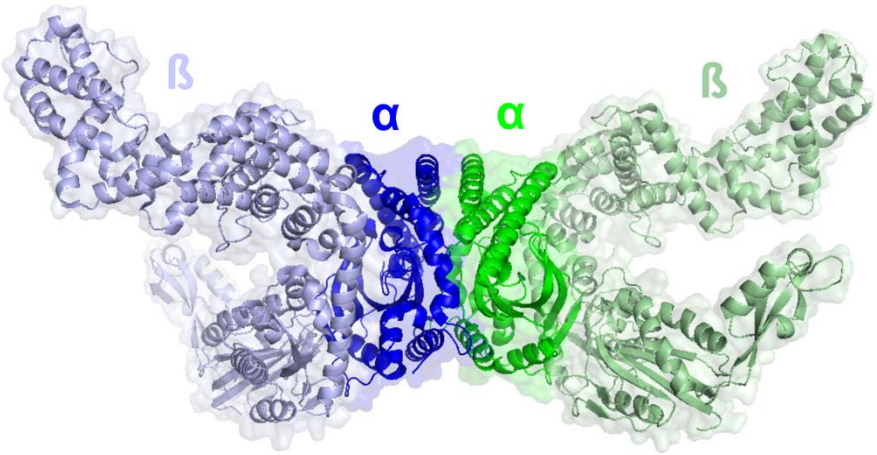

Figure 1. bacGlyRS is a βααβ complex formed around a central α_2 dimer. Each α subunit (300 aa) is flanked by one β subunit (700 aa), making bacGlyRS the biggest known aaRS of all three life domains. Remarkably, the α-subunit represents the smallest known class II catalytic domain, and thus the β-subunit,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

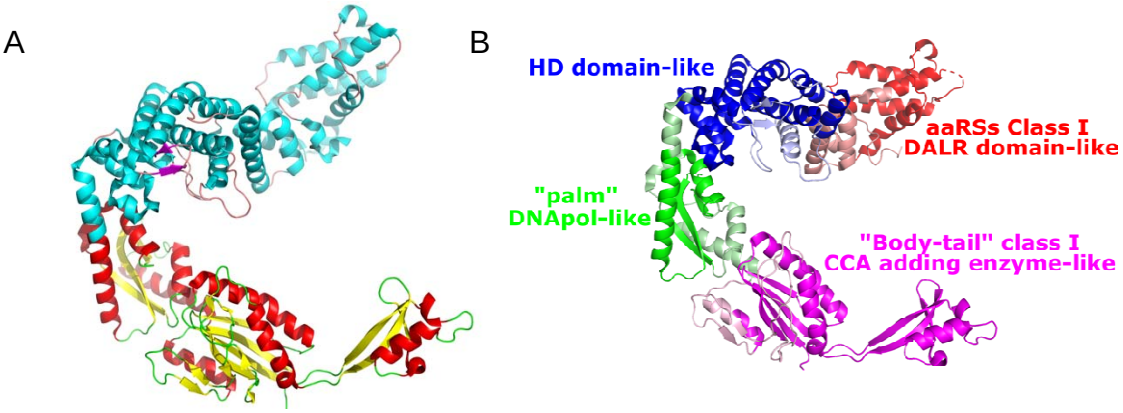

Figure 2. bacGlyRS β-subunit possesses two types of architecture, each one with domains showing homology to tRNA binding proteins specific for the acceptor (ancient) and anticodon stems. Left. In cyan is depicted an almost all α-helical region that defines the HD phosphorylase-like domain (blue in the figure at right) and the tRNA, anticodon binding domain of class I aaRSs (ACDB or DALR domain, red in the figure at right). In yellow and red (at left) is depicted an α+β region that comprises three domains: a linker domain (green in the figure at right) 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.

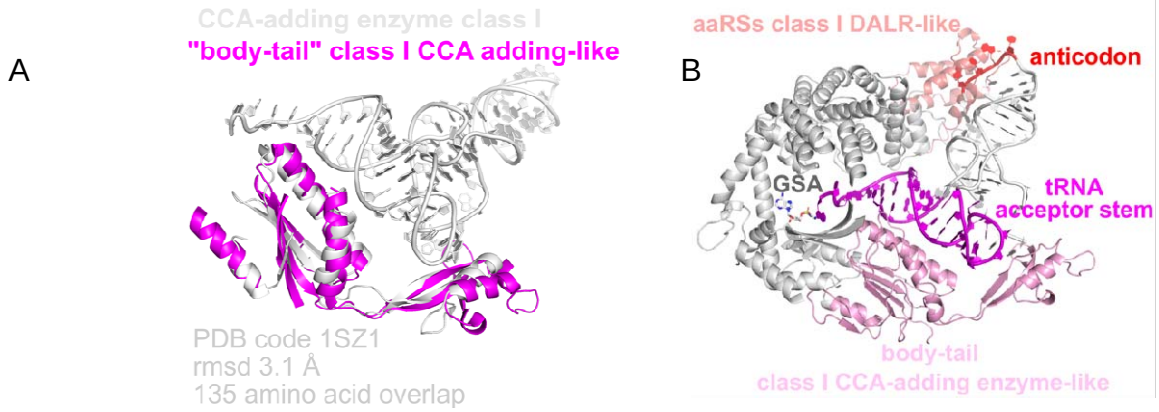

Figure 3. archaeal CCA-adding enzyme provides a template to build a clash-free model of the bacGlyRS:tRNA:GSA ternary complex. Left. Superposition of the body-tail domains of the CCA-adding enzyme:tRNA complex (white) on the equivalent domains of bacGlyRS (magenta). Right. Resulting model of the superposition 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.

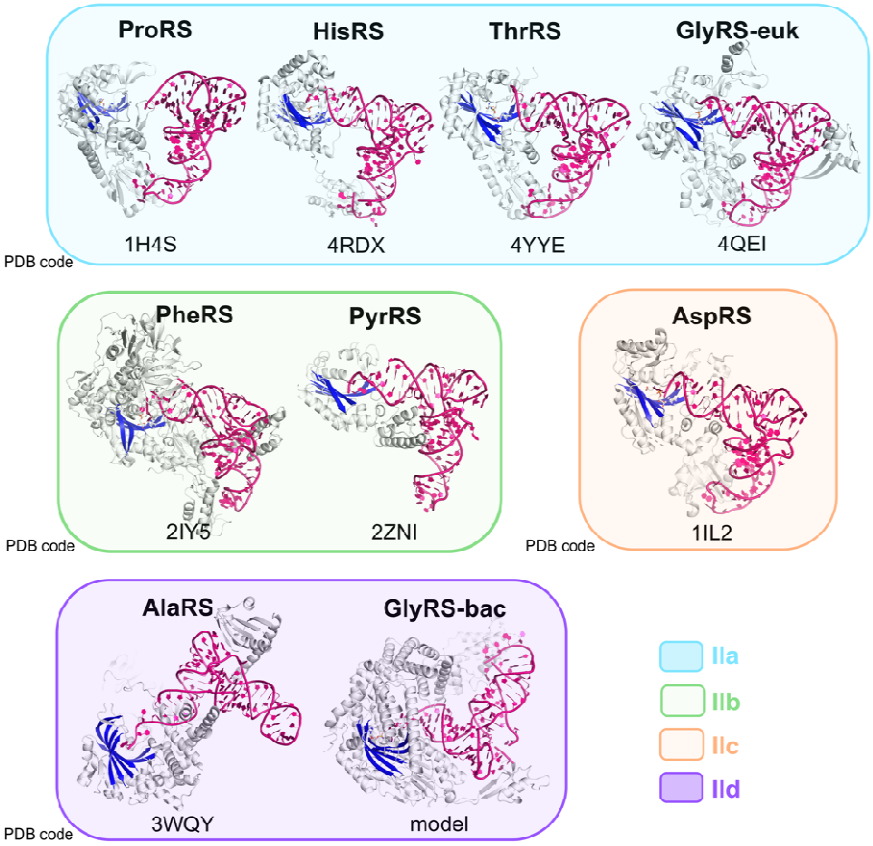

Figure 4. bacGlyRS and AlaRS clearly define a subclass among class II aaRS. Besides their remarkable similarities at the catalytic domain (Valencia-Sánchez, 2016), bacGlyRS and AlaRS would

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.

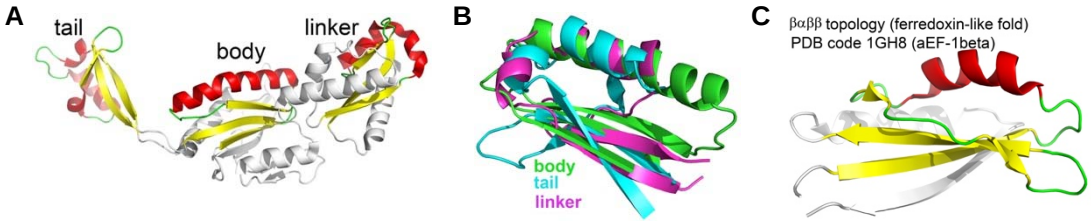

Figure 5. The body-tail-linker domains of bacGlyRS β-subunit may be the result of a triplication event. A. The linker, body and tail domains of bacGlyRS β-subunit show a repeated βαββ topology that can be roughly superposed, as shown in B. This topology has homology with the ancient and widespread ferredoxin 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.

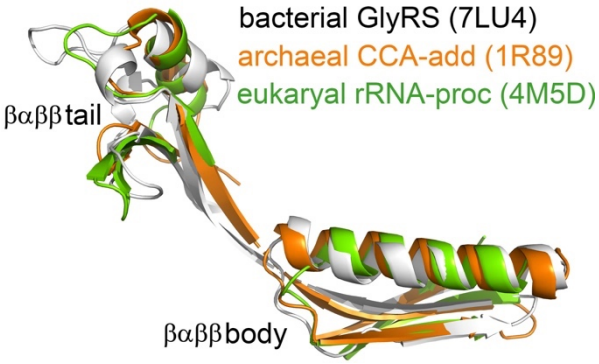

Figure 6. The duplicated βαββ, ferredoxin-like, primitive RNA binding topology present in the three life domains provides a clear coevolution signal of an ancient RNP world. Superposition of the duplicated βαββ topologies of bacGlyRS, archCCAadd and eukrRNAproc, all representing ancient protein topologies 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., 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.

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	α: Catalytic domain groups with AlaRS (subclass IId); Ala, as Gly, is

(second codon position appeared first-coded in acceptor stem, first codon position appeared second-coded in anticodon).	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-ACDB (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 ACDB). 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

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. A.T.L dedicates this work to Elizabeth Nallely Cabrera González, Prof. Ruy Pérez-Montfort and Prof. Enrique Merino-Pérez, IBT-UNAM, for their constant support and guidance. The authors thank Elizabeth Nallely Cabrera González for scientific and 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. 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. wrote the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

References

Organisms that possess the dimeric GlyRS

EggNOG Database COG0423 http://eggnog5.embl.de/#/app/home

Thermanaerothrix daxensis {Gregoire, 2011 #14}

1r89 description

The body domain represents an structural outlier because it has inserted large part of its secondary structures (Table 1) {Arumugam, 2013 #5}

Domains description

The HD domain may be an ancestor of phosphohydrolases {Aravind, 1998 #27}. The ancestral enzymes in nucleotide metabolism were probably phosphotransferases transferring P-containing groups with a phosphate group (EC 2.7.4), hydrolases acting on P-containing acid anhydrides (EC 3.6) {Caetano-Anolles, 2007 #29}

Evolution of multidomain proteins

Importance of domain combination and of domain-centric functional classification {Vogel, 2004 #40}, {Naveenkumar, 2019 #41}.

There is a link between internal duplication and the evolution of multi-domain proteins {Nacher, 2010 #43}, {Forslund, 2019 #47}.

3D protein domain structures and protein domain architecture are the molecular fossils of evolution and HGT in proteins is limited {Aziz, 2021 #46}, {Kurland, 2015 #42}.

We also may refer to the body-tail topology as a protein tandem repeat of class V {Kajava, 2012 #44}.

Primordial protein folds and repeats

Reuse of sub-domain motifs {Heizinger, 2021 #15}, {Kolodny, 2021 #16}

Expansion of protein domain repeats {Bjorklund, 2006 #18}

Vocabulary of ancient peptides at the origin of folded proteins {Alva, 2015 #28}

Origin of a folded repeat protein from an intrinsically disordered ancestor {Zhu, 2016 #30}

Proteins possibly originated from small molecule, nucleic acid binding modules {Caetano-Anolles, 2009 #45}

Proposals for the evolution of protein architectures and protein domain repeats {Wang, 2009 #31}, {Bukhari, 2013 #34}. However, a clear example is lacking.

βαββ topology

βαββ unit found in the ferredoxin-like fold (Fig 3A) {Lin, 2015 #1}. The ferredoxin fold is proposed to be one of the oldest in nature {Kim, 2012 #3}. In addition, ferredoxin-like domains are the second most common in archaea and bacteria {Wolf, 1999 #4}

Most ribosomal proteins from the large subunit is of the α + β type with βαββ topology similar to the RRM typified by the splicing protein U1A (page 156, Chapter 7) {Rice, 2008 #2}

We tried to find the closest structural neighbors to the babb topology found in b-GlyRS (Supp Fig x, PCA analysis). The closest homolog is the body-tail domains of CCAadd enzyme of type II

Issues in phylogenetics. More is not enough {Philippe, 2011 #36}

All know phylogenies are made with either a single or few genes from whole, multidomain genes/proteins. This approach does not consider the different evolving rates of proteins. HGT, orthology determination, inaccurate evolutionary models and compositional biases are also common in bacterial phylogenies {Superson, 2019 #38}.

Phylum relationships

Thermotogales should be considered members of the Firmicutes {Zhaxybayeva, 2009 #6}

Chlamydia and cyanobacteria are related {Brinkman, 2002 #7}, {Ball, 2015 #37} (for). {Domman, 2015 #8} (against). The eukaryotic MEP pathway evolved from Chlamydia and Cyanobacteria {Zeng, 2021 #9}. Notably, our proposed domains are not related to phototrophy.

Four of the seven phyla where reaction-center phototrophy has been reported are present in our tree (Cyanobacteria, Chloroflexi, Firmicutes and Proteobacteria) {Raymond, 2008 #11}, {Imhoff, 2017 #13}, {Ward, 2018 #10}. Interestingly, Cyanobacteria appears to be close to Chloroflexi. In addition, Actinobacteria seems to be somewhat related with photosynthesis, as already suggested {Gupta, 2016 #12}. Heliobacterium (Firmicutes-Clostridia) has the simplest known photosynthetic apparatus but seems to be contemporary to Cyanobacteria. Together, these two phyla (representing the two types of different known reaction centers) may be at the origin of phototrophy.

Alphaproteobacteria phylogeny uses Magnetococcales as an outgroup {Wang, 2021 #17}.

Previous remarkable trees and reclassifications

Deltaproteobacteria is divided in four phyla that do not belong to proteobacteria {Waite, 2020 #23}. Epsiloproteobacteria is also proposed to be reclassified {Waite, 2017 #35}.

Evolutionary proximity between domains Bacteria and Archaea (the tree including the presence of bacGlyRS was taken from this reference) {Zhu, 2019 #19}.

A rooted phylogeny resolves early bacterial evolution {Coleman, 2021 #20}

Firmicutes phylogeny (our tree is in agreement!!). Divide in Limnochordales, Halanaerobiales, Negativicutes {Taib, 2020 #21} (look at their whole phylogeny)

ML derived trees

IQ-TREE finds the highest likelihood scores {Zhou, 2018 #22}

Deltaproteobacteria (sulfur and sulfate reducing bacteria and Archaea) in the root of the 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 (pag 449) {Rosenberg, 2013 #39}.

Archaea-first and coevolution

Hinted in {Kim, 2012 #33}. Postulated in {Caetano-Anolles, 2014 #25}. Reviewed in {Staley, 2018 #24}. Methanopyrus as closer to bacteria in {Long, 2020 #26}