Bacterial glycyl tRNA synthetase offers glimpses of ancestral protein topologies

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.

Running title

Bacterial glycyl tRNA synthetase as an ancestral protein

Keywords

Aminoacyl tRNA synthetase, protein evolution, multifunctional protein, protein domain, protein-nucleic acid interaction, RNA binding protein, molecular evolution

Abstract

Aminoacyl tRNA synthetases (aaRSs) are among the proposed proteins present in the Last Universal Common Ancestor (LUCA). There are two types of glycyl tRNA synthetases (GlyRSs), from which the archaeal-eukaryal type is the one suggested to be present in LUCA. Here we solved 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, if bacGlyRS is ancestral, it should help to build a reliable Tree of Life (ToL). Given the modular nature of protein evolution, we used only two sub-domain segments with duplicated ancestral topologies, no detected orthologs and an assumed limited horizontal gene transfer (HGT). These motifs correspond to the non-specific RNA binding regions of contemporary bacGlyRS, archaeal CCA-adding enzyme (arch-CCAadd), and eukaryotic rRNA processing enzyme (euk-rRNA). The calculated, rooted bacterial ToL agrees with several phyla relationships unaccounted by the available trees.

Introduction

Among the 355 genes proposed to have been present in LUCA, eight code for aaRSs, enzymes that attach cognate amino acids to their corresponding tRNAs. Glycyl tRNA synthetase (GlyRS) is found between these candidates (1). 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, archaea, and some bacteria and an $\alpha_2\beta_2$ heterotetramer (bacGlyRS), related to AlaRS and present in bacteria (2), (3), (4). Due to its distribution in the three life domains, it is believed that eukGlyRS is the ancestral type that was present in LUCA (1). However, bacGlyRS is found in bacteria that are deeply rooted, like *Thermotoga* and *Cyanobacteria* (5), (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 (6).

Here we show the three-dimensional structure of a functional, complete bacGlyRS from the thermophilic bacteria *Thermanaerotrix daxensis* (Chloroflexi) (7) 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, making an hybrid of class I and II type of tRNA recognition. The atypical features found in both the catalytic domain (4) 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. The CCA-adding-rRNA processing-like architecture of the β-subunit represents a multifunctional, primordial non-specific RNA binding protein. Together with bacGlyRS, they make an example of the coevolution of the ribonucleoprotein (RNP) world (8). These three proteins are related 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, most proximal by sequence identity. We thus obtained a rooted tree that agrees with several previous observations. Still, it also suggests novel relationships between species, 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

To gain further insights into the evolution of bacGlyRS, we solved the crystal structure of the naturally linked αβ-subunits of GlyRS of *Thermoanaerotrix daxensis* (see Supplementary Results). The comparison of the two types of GlyRS (bacterial and eukaryal, 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, bacGlyRS represents a unique 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 βααβ complex formed around a central α_2 dimer (9). In addition, none of the domains shares structural homology between the GlyRSs, except for the core of the catalytic domain, with an 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). Different subclasses must define the difference between the two types of GlyRS. 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 (4).

- 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, Supplementary 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 a 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 model further strengthens that bacGlyRS and AlaRS clearly define a subgroup of aaRSs 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, Supplementary 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) hints that a triplication event was at the origin of this multidomain protein very early in evolution. This hypothesis makes sense because the ferredoxin-like fold is proposed to be very ancient (10), (11), (12), (13) as well as the tRNA acceptor stem (14), (15). 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 (16)

- 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 β-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 4) that yields information consistent with the widely accepted taxonomy, like the overall, Gracilicutes vs. Terrabacteria division (17), the general Firmicutes phylogeny division of Limnochordales, Halanaerobiales, Negativicutes (18), or the placement of Magnetococcales as an Alphaproteobacteria outgroup (19). It also proposes that Methanopyrus is the closest Archaea related to bacteria (20); to divide Deltaproteobacteria into four phyla that do not belong to proteobacteria (21). Epsiloproteobacteria is also suggested to be reclassified as not a Proteobacteria, as proposed in (22). Notably, our proposal implicates that Thermotogales should be considered members of the Firmicutes, as already suggested by (23) and places Fusobacteria very close to the root of the tree, as recently proposed (17).

Perhaps our proposed tree's two most outstanding features 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 (24), (25), (26). Also, the relationship between Cyanobacteria and Chlamydia has been reported (27), (28), (29), although also disputed (30). Finally, Actinobacteria seems to be somewhat related to photosynthesis, as already suggested (31). 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 chemolithotrophic autotrophs. The first autotrophic processes were akin to those seen in the sulfur and sulfate-reducing bacteria (32).

Discussion

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. The α-subunit of bacGlyRS may have appeared as one of the first aaRSs (Supplementary Table 2), even in a pre-LUCA world, where the $\alpha\beta$ architecture of the β-subunit served as a multifunctional, 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. 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 afterward, leaving some domains of the β-subunit (i.e., the tail and body) useful for other enzymes (i.e., the CCA-adding).

Given the relationships of bacGlyRS with AlaRS (catalytic core and primordial aaRS) and ArgRS (primordial ACBD), it is possible to map them on the genetic code table (Figure 5) and obtain a correlation between the emergence of these putative primordial aaRSs with the tRNA acceptor stem code (first code, second codon letter) and the anticodon stem code (second code, first codon letter). Remarkably, it is not possible to make such a relationship with any other aaRS, which further highlights the ancient nature of bacGlyRS. We chose to perform our phylogenetic analysis using the duplicated βαββ topology found because it may represent an example of ancient domain accretion and multiple domain repeats not identified by sequence. Ancient proteins described tend to be a single-domain, so we consider that the duplicated, sub-domain topology offers a better chance of unveiling truly ancient evolutionary events.

Overall, this work proposes (see Highlights in the Supplementary Material) that a 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 multidomain 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.

Experimental procedures

Refer also to Supporting information

Crystallization and structure determination

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). 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 (33), Pointless (34) and reduced with Aimless (35). The structure was solved by molecular replacement using MrBUMP (36), with a search model based on PDB entry 5F5W. The initial model was obtained using the Phenix.Autobuild program of the PHENIX package (37) and was displayed and manually built in COOT (38). Refinement was alternated with the use of the PDB_REDO server (39).

Data availability

The atomic coordinates and structure factors of the αβ bacGlyRS from *T. daxensis* have been deposited in the Protein Data Bank (PDB), code 7LU4.

Supporting information

This article contains supporting information.

Acknowledgments

The authors thank Zdzislaw Wawrzak for help with the crystallographic aspect of this project, Elizabeth Nallely Cabrera González, Georgina E. Espinosa Pérez, Pamela J. Focia and Valerie Tokars for technical guidance.

Author contributions

A.T.L, A.R.H., M.I.V.S, and D.E.R.CH conceived the project. 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.

Funding and additional information

A.T.L. was supported by Universidad Nacional Autónoma de México (PAPIIT-DGAPA) grants IN202416 and IN204820. Use of the Advanced Photon Source 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.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

Figures and figure legends

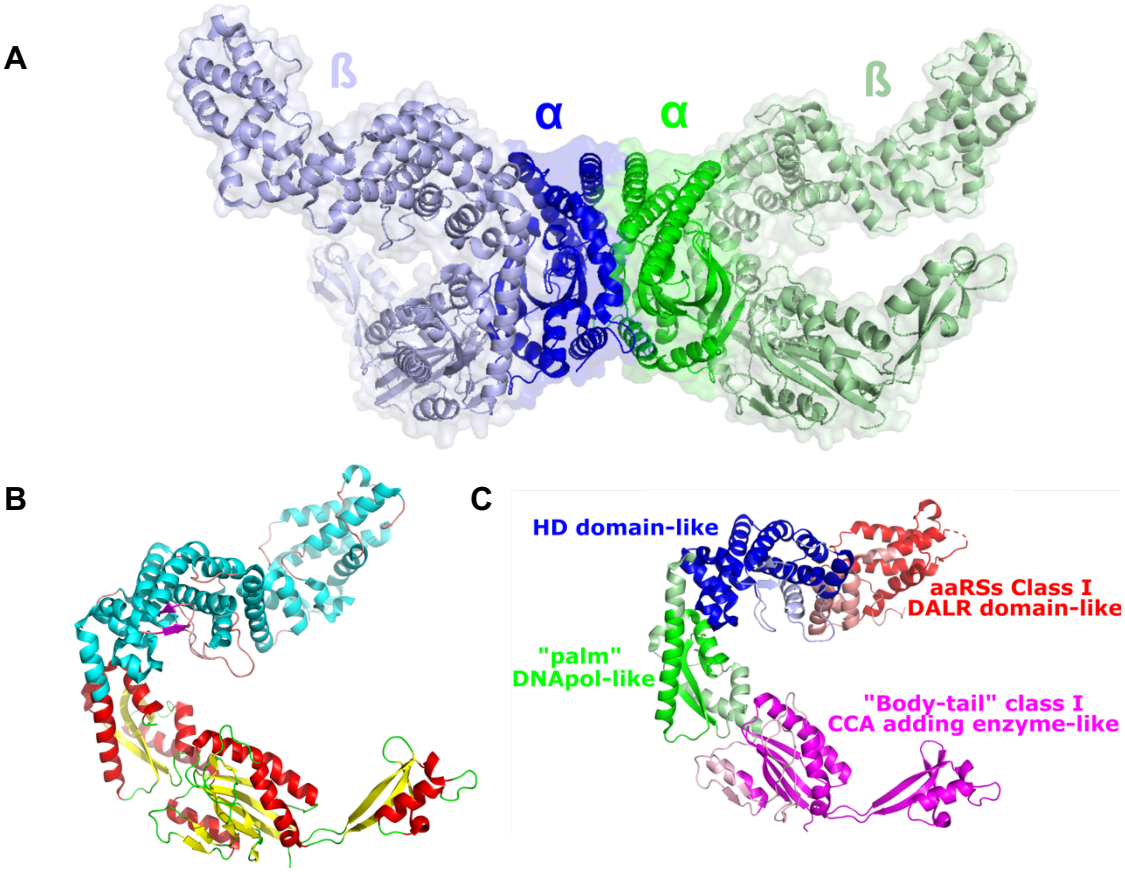

Figure 1. bacGlyRS is a βααβ complex formed around a central α_2 dimer. The β-subunit possesses two types of architecture, each with domains homologous to tRNA binding proteins specific for the acceptor (ancient) and anticodon stems. A. Each α subunit (300 aa) is flanked by one β subunit (700 aa), making 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. **B**. The two architectures of the bacGlyRS β-subunit. In cyan is 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 **C,** respectively. In yellow and red is an α+β region that comprises three domains: a linker domain (green in **C**) 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 **C**). **C**. The five identified domains in the bacGlyRS β-subunit are shown in different colors.

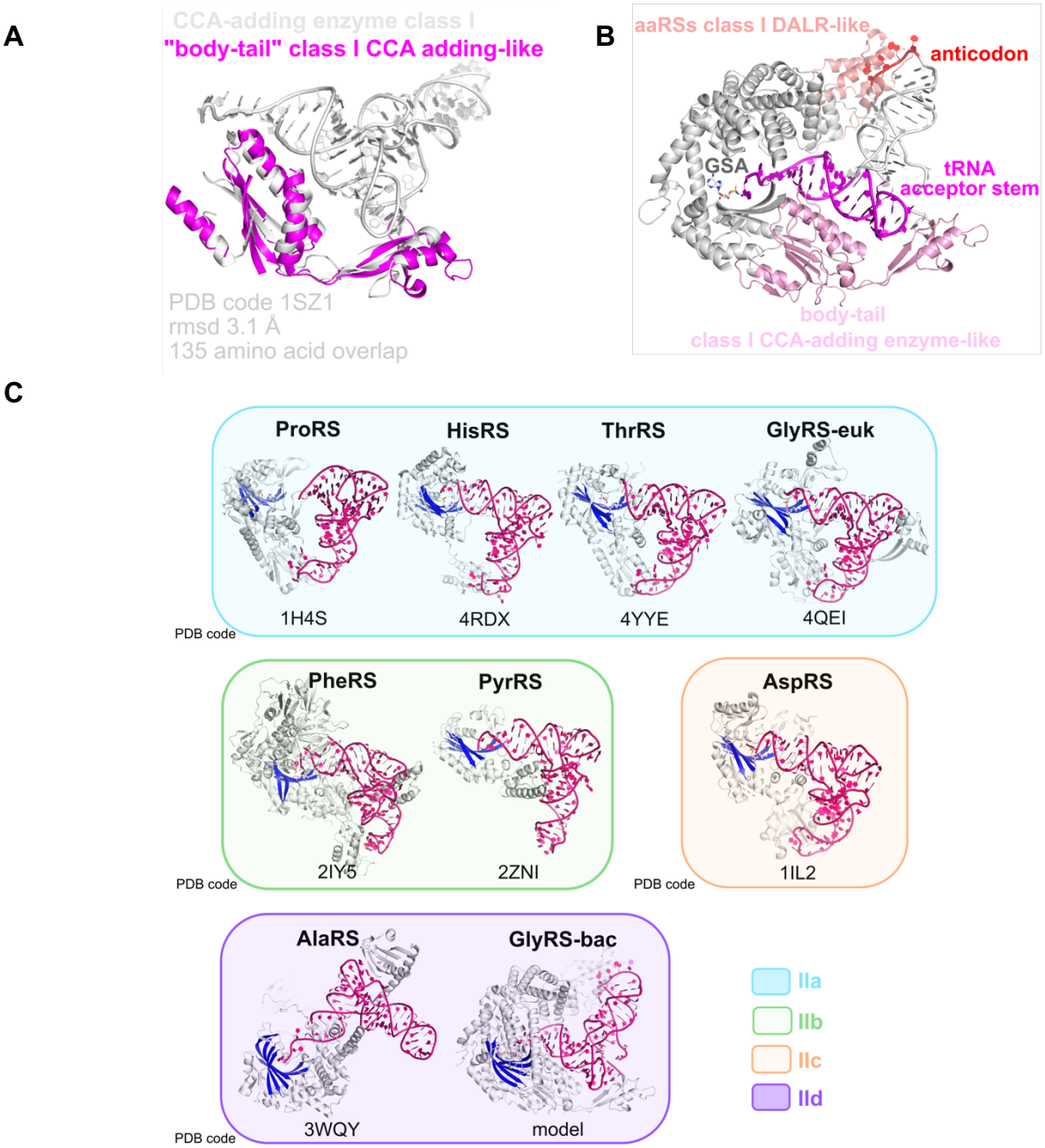

Figure 2. The archaeal CCA-adding enzyme provides a template to build a clash-free model of the bacGlyRS:tRNA:GSA ternary complex. The complex shows that bacGlyRS and AlaRS clearly define a subclass among class II aaRS A. Superposition of the body-tail domains of the CCA-adding enzyme:tRNA complex (white) 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. **C.** 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 regarding all other class II aaRSs (see also Supplementary Figure 5). The relative orientation of the catalytic domain (depicted in part in blue) is the same for all models.

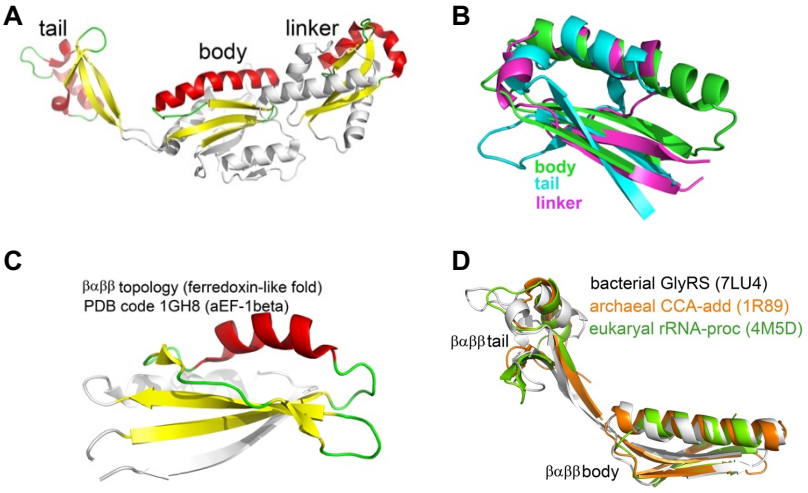

Figure 3. The body-tail-linker domains of the bacGlyRS β-subunit may be the result of a triplication event. The duplicated βαββ, ferredoxin-like, primitive RNA binding topology in the three life domains provides a clear coevolution signal of an ancient RNP world. A. The linker, body, and tail domains of bacGlyRS β-subunit show a repeated βαββ topology that can be roughly superposed, as shown in **B.** This topology is homolog of the ancient and widespread ferredoxin fold, exemplified in **C** by the PDB code 1GH8, representing an archaeal elongation factor. This topology 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) (40). We also may refer to the body-tail topology as a protein tandem repeat of class V (41). **D.** 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, around 10%. No other tandem topology homolog has been found in nature, which shows a limited horizontal gene transfer (HGT). This duplicated, not single, topology, reassures its uniqueness and highlights its phylogenetic signal.

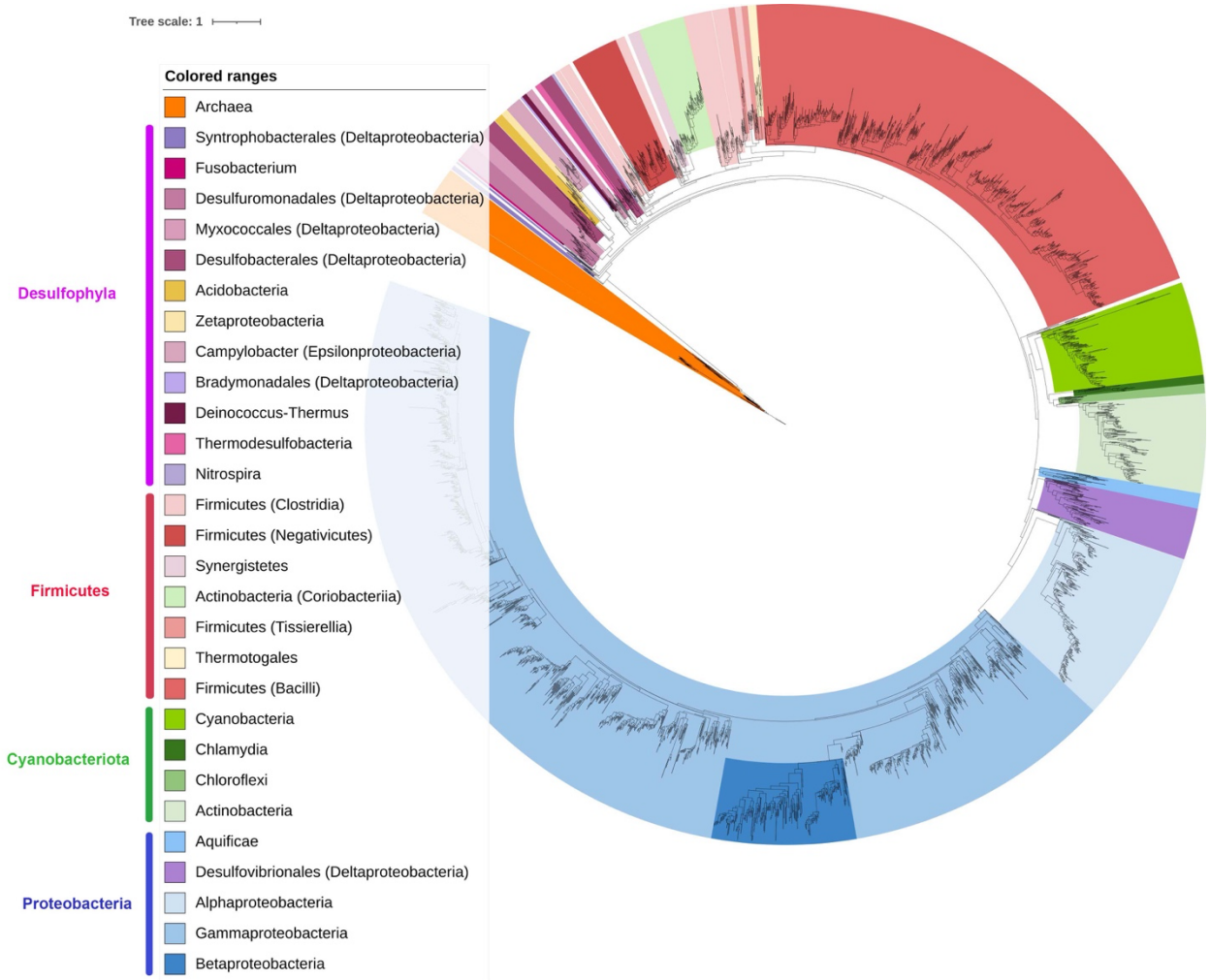

Figure 4. The ancestral, duplicated βαββ, ferredoxin-like topologies of archaea-CCAadding and bacterial β-GlyRS draw the rooted bacterial Tree of Life. The expected limited Horizontal Gene Transfer (HGT) of the duplicated βαββ topology of bacGlyRS led to this proposition, rooted in the archaeal βαββ topology of archaeal CCA-adding enzyme. Several deeply rooted phyla belong to what was previously known as Deltaproteobacteria. We propose four superphyla, 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. (42), 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.

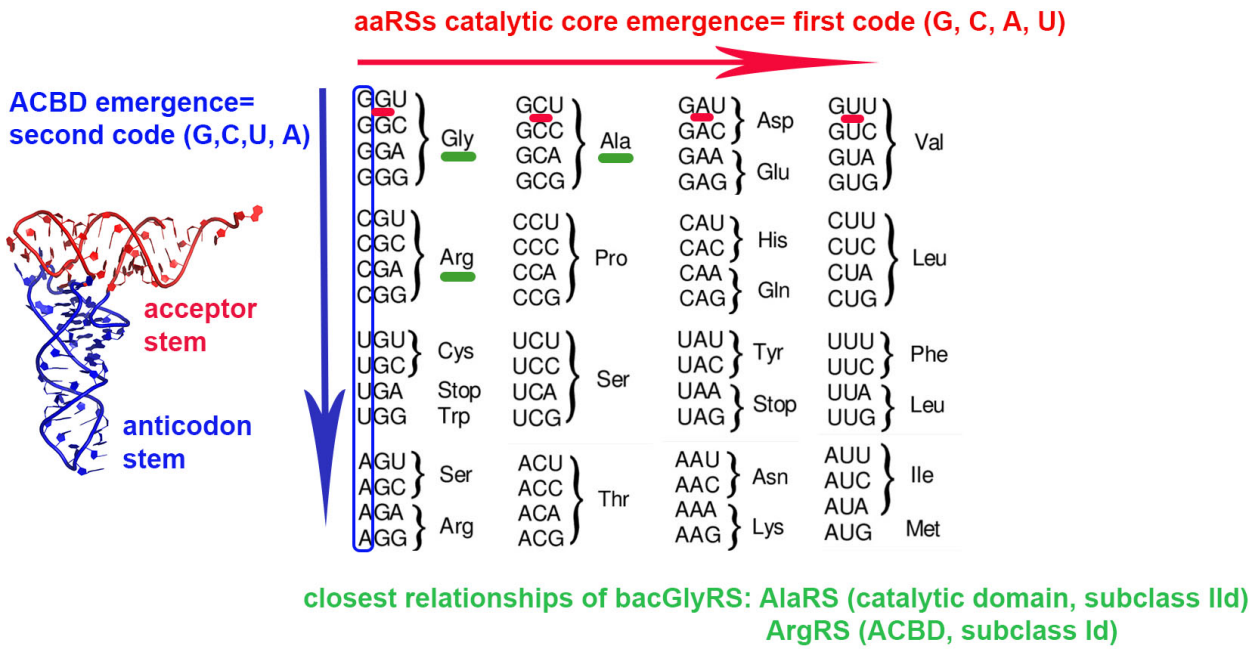

Figure 5. The genetic code table reflects the age of primordial aaRSs and their evolutionary reading (second base first, first base second). The "first, or operational code" (43) corresponds to the second letter of the codon (44), and it is suggested to lie on the acceptor stem of the tRNA (ancient). 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 agrees with an almost concomitant appearance of class I/II aaRSs and previous hypotheses on the emergence of aaRSs (45). The first, primitive code, could have been specific for Gly, Ala, Arg, and Pro.

References

1. Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V., and Martin, W. F. (2018) The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet* **14**, e1007518
2. Ostrem, D. L., and Berg, P. (1974) Glycyl transfer ribonucleic acid synthetase from Escherichia coli: purification, properties, and substrate binding. *Biochemistry* **13**, 1338-1348
3. Tang, S. N., and Huang, J. F. (2005) Evolution of different oligomeric glycyl-tRNA synthetases. *FEBS Lett* **579**, 1441-1445
4. 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., and Torres-Larios, A. (2016) Structural Insights into the Polyphyletic Origins of Glycyl tRNA Synthetases. *J Biol Chem* **291**, 14430-14446
5. Fournier, G. P., Andam, C. P., and Gogarten, J. P. (2015) Ancient horizontal gene transfer and the last common ancestors. *BMC Evol Biol* **15**, 70
6. Koonin, E. V., and Novozhilov, A. S. (2017) Origin and Evolution of the Universal Genetic Code. *Annu Rev Genet* **51**, 45-62
7. Gregoire, P., Fardeau, M. L., Joseph, M., Guasco, S., Hamaide, F., Biasutti, S., Michotey, V., Bonin, P., and Ollivier, B. (2011) 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**, 494-497
8. Schimmel, P. (2011) The RNP bridge between two worlds. *Nat Rev Mol Cell Biol* **12**, 135
9. Koski, M. K., Anantharajan, J., Kursula, P., Dhavala, P., Murthy, A. V., Bergmann, U., Myllyharju, J., and Wierenga, R. K. (2017) Assembly of the elongated collagen prolyl 4-hydroxylase alpha2beta2 heterotetramer around a central alpha2 dimer. *Biochem J* **474**, 751-769
10. Wolf, Y. I., Brenner, S. E., Bash, P. A., and Koonin, E. V. (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res* **9**, 17-26
11. Kim, J. D., Rodriguez-Granillo, A., Case, D. A., Nanda, V., and Falkowski, P. G. (2012) Energetic selection of topology in ferredoxins. *PLoS Comput Biol* **8**, e1002463
12. Lin, Y. R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A. F., Montelione, G. T., and Baker, D. (2015) Control over overall shape and size in de novo designed proteins. *P Natl Acad Sci USA* **112**, E5478-E5485
13. Raanan, H., Poudel, S., Pike, D. H., Nanda, V., and Falkowski, P. G. (2020) Small protein folds at the root of an ancient metabolic network. *Proc Natl Acad Sci U S A* **117**, 7193-7199
14. Weiner, A. M., and Maizels, N. (1987) 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**, 7383-7387

15. Carter, C. W., Jr., and Wolfenden, R. (2015) tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proc Natl Acad Sci U S A* **112**, 7489-7494
16. Rice, P. A., and Correll, C. C. (2008) *Protein-nucleic acid interactions : structural biology*, RSC Pub., Cambridge
17. Coleman, G. A., Davin, A. A., Mahendrarajah, T. A., Szantho, L. L., Spang, A., Hugenholtz, P., Szollosi, G. J., and Williams, T. A. (2021) A rooted phylogeny resolves early bacterial evolution. *Science* **372**
18. Taib, N., Megrian, D., Witwinowski, J., Adam, P., Poppleton, D., Borrel, G., Beloin, C., and Gribaldo, S. (2020) Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol* **4**, 1661-1672
19. Wang, S., and Luo, H. (2021) Dating Alphaproteobacteria evolution with eukaryotic fossils. *Nat Commun* **12**, 3324
20. Long, X., Xue, H., and Wong, J. T. (2020) Descent of Bacteria and Eukarya From an Archaeal Root of Life. *Evol Bioinform Online* **16**, 1176934320908267
21. 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., and Hugenholtz, P. (2020) 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**, 5972-6016
22. 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., and Hugenholtz, P. (2017) Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed Reclassification to Epsilonbacteraeota (phyl. nov.). *Front Microbiol* **8**, 682
23. 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., and Noll, K. M. (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc Natl Acad Sci U S A* **106**, 5865-5870
24. Raymond, J. (2008) Coloring in the tree of life. *Trends Microbiol* **16**, 41-43
25. Imhoff, J. F., Rahn, T., Kunzel, S., and Neulinger, S. C. (2017) Photosynthesis Is Widely Distributed among Proteobacteria as Demonstrated by the Phylogeny of PufLM Reaction Center Proteins. *Front Microbiol* **8**, 2679
26. Ward, L. M., Hemp, J., Shih, P. M., McGlynn, S. E., and Fischer, W. W. (2018) Evolution of Phototrophy in the Chloroflexi Phylum Driven by Horizontal Gene Transfer. *Front Microbiol* **9**, 260
27. 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., and Greberg, H. (2002) Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. *Genome Res* **12**, 1159-1167
28. Ball, S. G., Colleoni, C., Kadouche, D., Ducatez, M., Arias, M. C., and Tirtiaux, C. (2015) Toward an understanding of the function of Chlamydiales in plastid endosymbiosis. *Biochim Biophys Acta* **1847**, 495-504

29. Zeng, L., and Dehesh, K. (2021) The eukaryotic MEP-pathway genes are evolutionarily conserved and originated from Chlaymidia and cyanobacteria. *BMC Genomics* **22**, 137
30. Domman, D., Horn, M., Embley, T. M., and Williams, T. A. (2015) Plastid establishment did not require a chlamydial partner. *Nat Commun* **6**, 6421
31. Gupta, R. S., and Khadka, B. (2016) 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**, 201-218
32. Rosenberg, E. (2013) *The prokaryotes -prokaryotic physiology and biochemistry : prokaryotic physiology and biochemistry*, 4th ed., Springer, New York
33. Kabsch, W. (2010) Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132
34. Evans, P. R. (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr* **67**, 282-292
35. Evans, P. R., and Murshudov, G. N. (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* **69**, 1204-1214
36. Keegan, R. M., and Winn, M. D. (2008) MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallogr D Biol Crystallogr* **64**, 119-124
37. 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., and Adams, P. D. (2019) Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877
38. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-2132
39. Joosten, R. P., Long, F., Murshudov, G. N., and Perrakis, A. (2014) The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213-220
40. Arumugam, G., Nair, A. G., Hariharaputran, S., and Ramanathan, S. (2013) Rebelling for a reason: protein structural "outliers". *PLoS One* **8**, e74416
41. Kajava, A. V. (2012) Tandem repeats in proteins: from sequence to structure. *J Struct Biol* **179**, 279-288
42. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534
43. Schimmel, P., Giege, R., Moras, D., and Yokoyama, S. (1993) An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci U S A* **90**, 8763-8768
44. Rodin, S. N., and Rodin, A. S. (2006) 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**, 365-375

45. Smith, T. F., and Hartman, H. (2015) The evolution of Class II Aminoacyl-tRNA synthetases and the first code. *FEBS Lett* **589**, 3499-3507