Enzymic recognition of amino acids drove the evolution of primordial genetic codes

Jordan Douglas[1,*], Remco Bouckaert[2], Charles W. Carter, Jr [3] , Peter R. Wills[1]

[1]Department of Physics, The University of Auckland, New Zealand
[2]School of Computer Science, The University of Auckland, New Zealand
[3]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, USA
[*]jordan.douglas@auckland.ac.nz

May 4, 2023

Supporting information

Identification of AARS Families

To identify the 33 AARS catalytic domain families, we inferred the phylogenies of Class I and II catalytic domains using BEAST 2 [1]. These analyses, as described in Materials and Methods of the main article, were performed in order to identify families used in subsequent analyses and were based on an alignment of structural elements common to the entire class. These elements are: S1-S5 and H1-H5 for Class I, and S1-S5, H1-H3, SH1 and Motif 1 for Class II. The summary trees are in **Fig. S4** and **Fig. S2** demonstrate a high level of posterior support for the families identified. The families are summarised in **Table S1**. Most families are monophyletic with over 99% support. However, the following families have less than 75% support.

1. Class I ValRS has 67% clade support. The Archaeal form is distinct from the Bacterial and Eukaryotic form. With the remaining 33% support, these two ValRS clades belong to distinct parts of the subclass Ia clade.

2. Class II GlyRS-A is not monophyletic because it contains GlyRS-E. However, the two families combined are monophyletic with over 99% support. The closest GlyRS-A relative to the GlyRS-E family is from the Archaeal *Euryarchaeota* phylum.

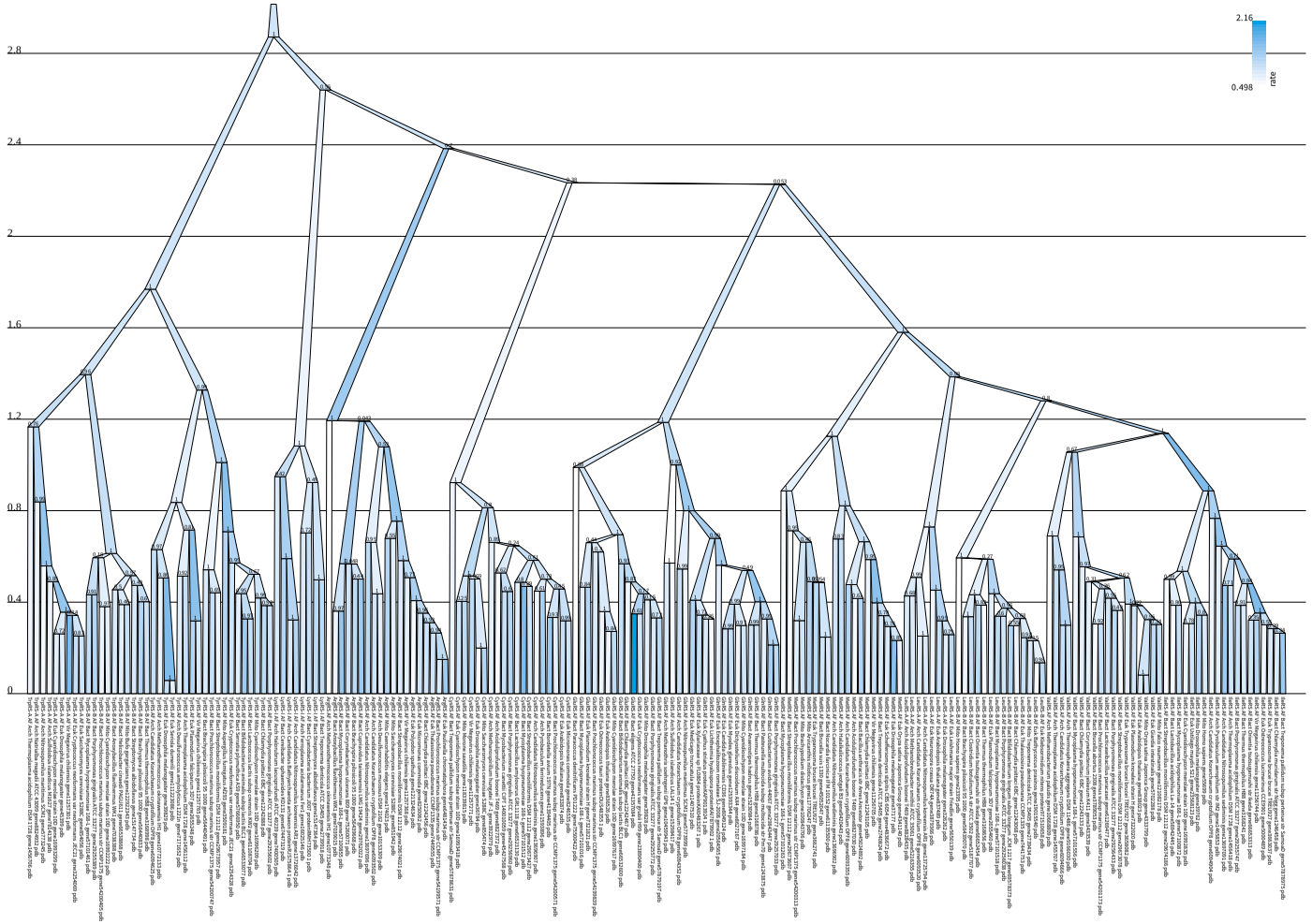

Fig. S1: Phylogeny of the Class I common catalytic domain sequence elements. This summary tree is the maximum clade credibility [2] tree with internal nodes labelled by posterior clade support and node heights are in substitutions per site. Branches are coloured by relative susbsitution rate under the relaxed clock model [3].

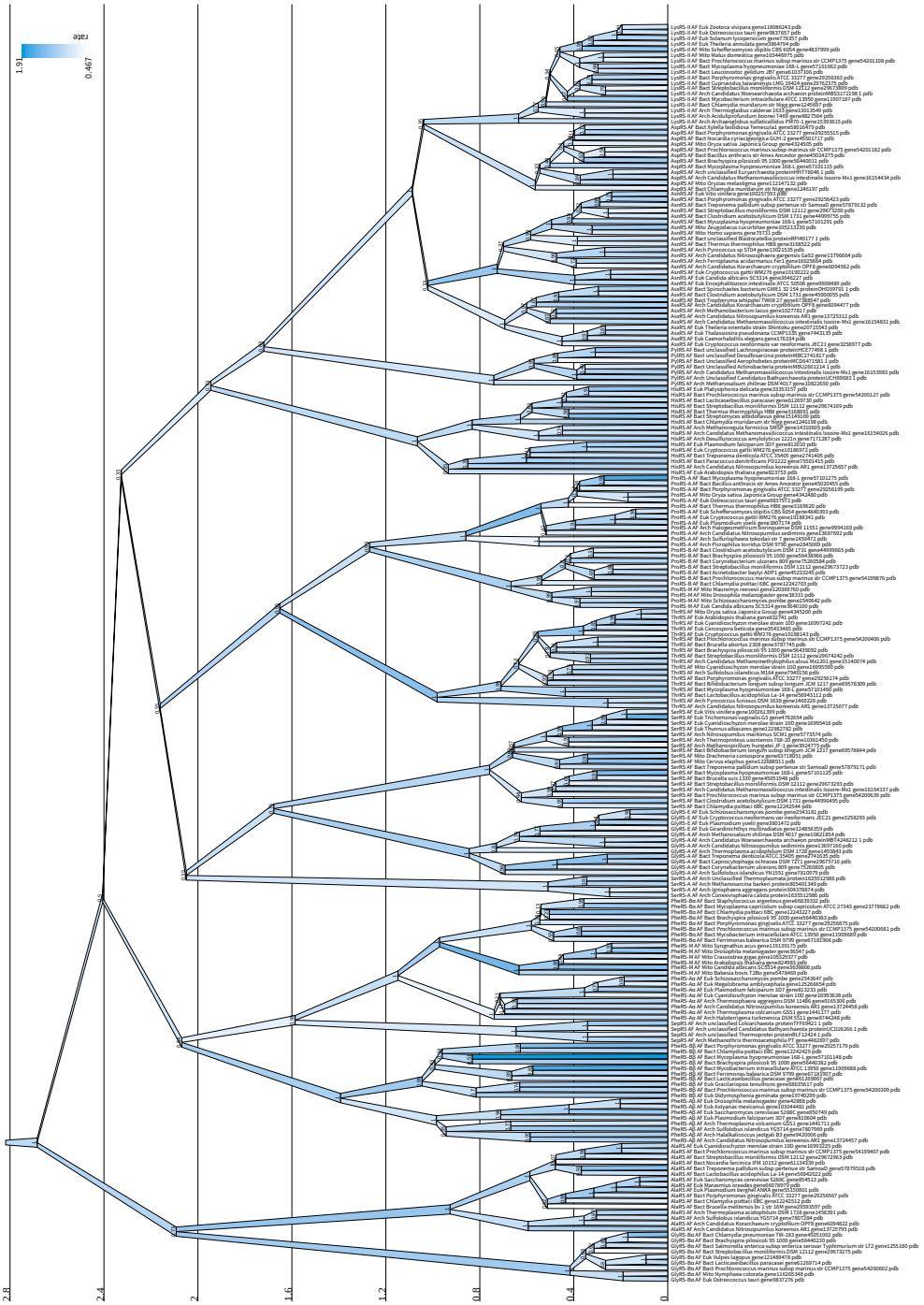

Fig. S2: Phylogeny of the Class II common catalytic domain sequence elements. See **Fig. S4** caption.

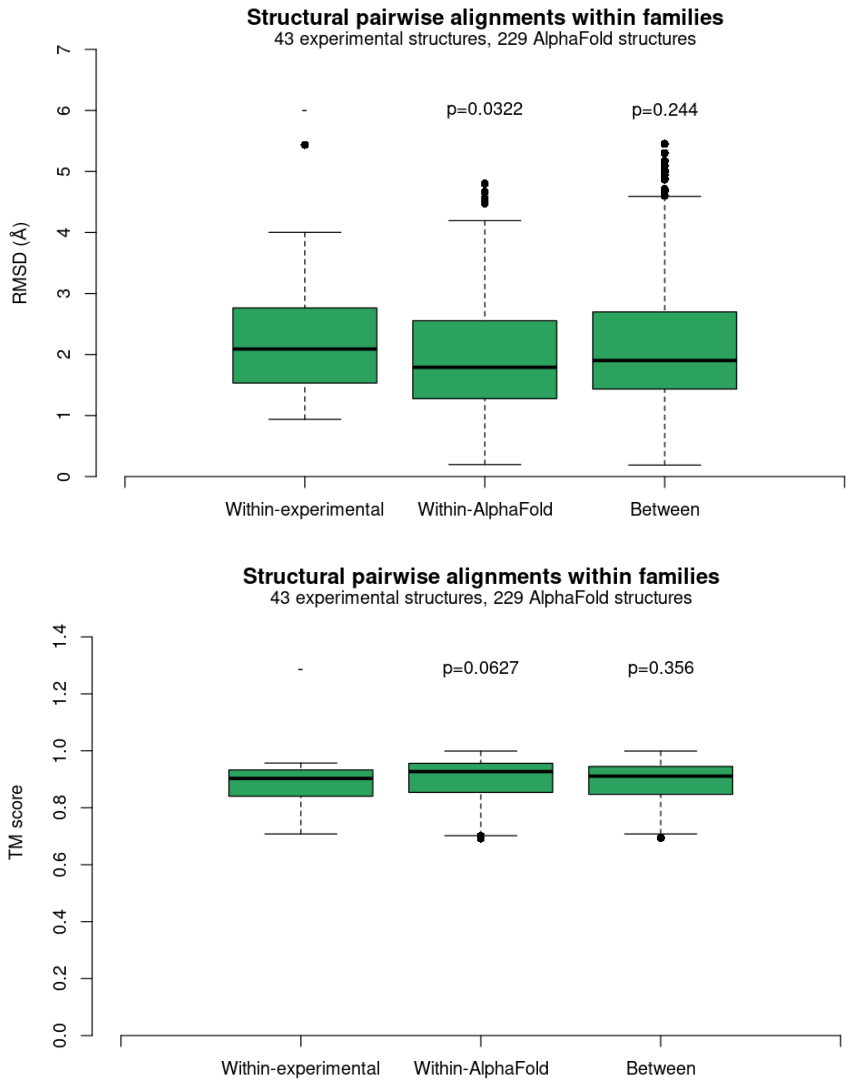

Fig. S3: Distances were calculated from alignments between experimentally solved structures, between AlphaFold structures, or across the two groups, using DeepAlign [4] on pairs of members from the same AARS catalytic domain family. P-values are from two-sided Student t-tests which compares the pairwise distances of each group with those of within-experimental. These results suggest that the AlphaFold structures are generally quite similar to the experimentally determined structures, and that the variation among experimentally determined structures is not significantly different to the variation among AlphaFold structures, or between the two groups. We performed this test on all AARS families with at least 2 experimentally solved structures from at least 2 species. These PDB structures are: ArgRS: 2zuf, 1iq0, 1bs2; CysRS: 1li5, 3c8z; IleRS: 1qu2, 1ile, 7d5c; MetRS: 1rqg, 1pfv, 1woy; TrpRS :3jxe, 3kt3, 1i6m, 3n9i; TyrRS: 1u7d, 1wq4, 2j5b; AsnRS: 1x54, 6pqh, 1asz; AsxRS: 1b8a, 1n9w, 1eov; GluRS-B: 5f5w, 3ufg, 1j5w; HisRS: 1wu7, 4e51, 1htt; PheRS-B :3pco, 1pys; ProRS-B: 5znj,2j3m; PylRS: 2q7e, 2znj; SepRS: 2odr, 2du4; SerRS: 6h9x, 1ser, 1wle; ThrRS: 1evk, 1nyr.

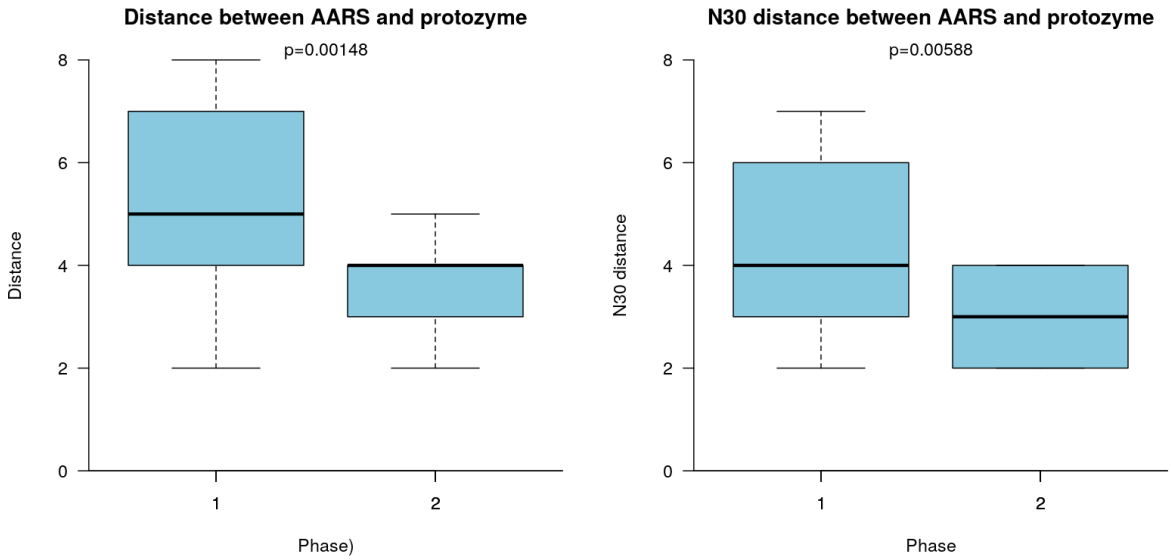

Fig. S4: Distance between each AARS and the protozyme, according to **Fig. 4** of the main article. Left: the distance is equal to the number of modules which were inserted or deleted to assemble the extant AARS from the protozyme. For example, LeuRS-A has a distance of 8. Right: the N30 distance is the same but it discounts any insertion modules which have an average length less than 30 amino acids (i.e., the zinc finger, the small interface, and the 2a hairpin). The N30 distance of LeuRS-A is 7. p-values are the result of a one sided Student t-test with a null hypothesis that the phase I and II amino acids are activated by AARS which are equally close to the protozyme (i.e., the same degree of structural primitivity). Both distance metrics suggest that the more recently occurring AARS are indeed more primitive.

Class	**Subclass**	**Family (PDB structure)**	P_{S}	P_{SM}
I	a	IleRS (7d5c), LeuRS-A (1wkb), LeuRS-B (2v0c), MetRS (1pfv), ValRS (1gax)	0.70	0.72
	b	GlnRS (1qtq), GluRS (1j09), GlxRS (3aii)	0.80	0.82
	c	TrpRS (1i6m), TyrRS (1wq4)	0.98	0.97
	d	ArgRS (1bs2)		
	e	LysRS-I (1irx)		
	f	CysRS (1li5)		
II	a	GlyRS-A (1ati), GlyRS-E (2q5h), ProRS-A (1nj8), ProRS-B (2j3m), ProRS-M, SerRS (1wle), SerRS-A (2cjb), ThrRS (1nyr)	0.97	0.80
	b	AsnRS (1asz), AspRS (1c0a), AsxRS (1eov), LysRS-II (3e9h)	0.99	0.99
	c	PheRS-M (3cmq) , PheRS-Aα (3l4g), PheRS-Bα (3pco), SepRS (2odr), HisRS (1htt)	0.37	0.78
	d	AlaRS (3hxv), GlyRS-B (1j5w)	0.94	0.98
	e	PylRS (2q7e)		

Table S1: AARS catalytic domain families and an example of an experimentally solved structure. There are no solved structures for ProRS-M. The estimated clade supports P are shown for subclasses with more than 1 member, where P_{S} was calculated using only sequence information, and P_{SM} using both sequences and insertion modules. The inclusion of singletons (such as ArgRS and LysRS-I) into other subclasses was found to significantly reduce their supports and therefore they are best identified as singletons.

Parameter	**Prior**	**Posterior (Class I)**	**Posterior (Class I)**
Family tree height	Yule prior	1.92 (1.58,2.31)	2.51 (2.16,2.87)
Yule birth rate	LN(-0.5,1)	0.787 (0.364,1.22)	0.679 (0.348,0.968)
Relaxed clock standard deviation	Gamma(4,0.05)	0.447 (0.348,0.569)	0.409 (0.338,0.526)
Amino acid substitution rate	1	1	1
Insertion module birth rate	LN(-5,2)	0.0411 (8.85e-06,0.18)	0.0295 (5.43e-06,0.121)
Insertion module death rate	LN(-7.3,2)	0.00465 (5.92e-07,0.0169)	0.00334 (5.77e-06,0.0147)
Substitution model indicator	Uniform({ 0,1,...,14 })	13 {13,13}	13 {13,13}
Using gamma rate heterogeneity?	Uniform([0,1])	1 [1,1]	1 [1,1]
Estimating invariant proportion?	Uniform([0,1])	1 [1,1]	1 [1,1]
Gamma rate heterogeneity shape	Exponential(1)	2.06 (1.94,2.19)	2.04 (1.88,2.25)
Proportion of invariant sites	Beta(1,4)	0.00502 (0.00138,0.00896)	0.0122 (0.00453,0.0192)

Table S2: Prior and posterior distributions under the Sequence+IM model. Prior distributions: LN: LogNormal distribution with parameters μ and σ; gamma and beta distributions with parameters α and β; exponential distribution with mean; discrete uniform distribution over the specified elements. The gamma shape also has a minimum value of 0.1, and the most commonly estimated subsitution model by OBAMA was model 13 – the VT model [5].

Class I Insertion Module Phylogenies

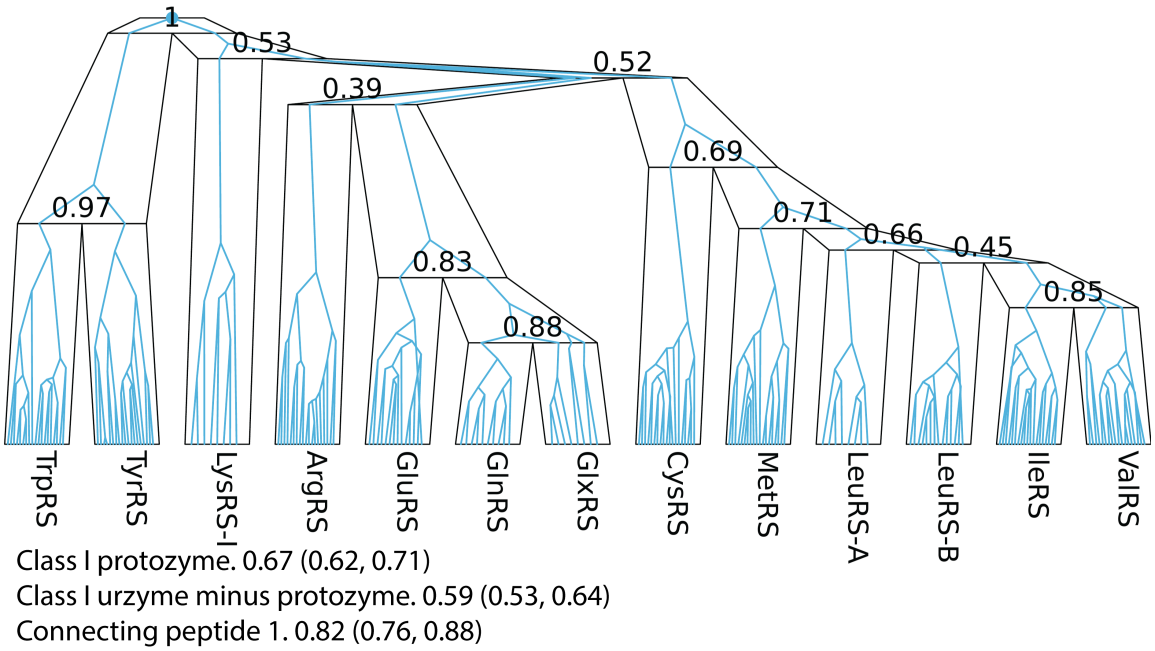

Fig. S5: Phylogeny of the three structural modules common to Class I and their estimated relative amino acid substitution rates (and 95% credible interval). Family tree internal nodes are labelled by posterior support. All of the consequent Class I modules are assumed to have a phylogeny which is a clade of the coloured phylogeny above.

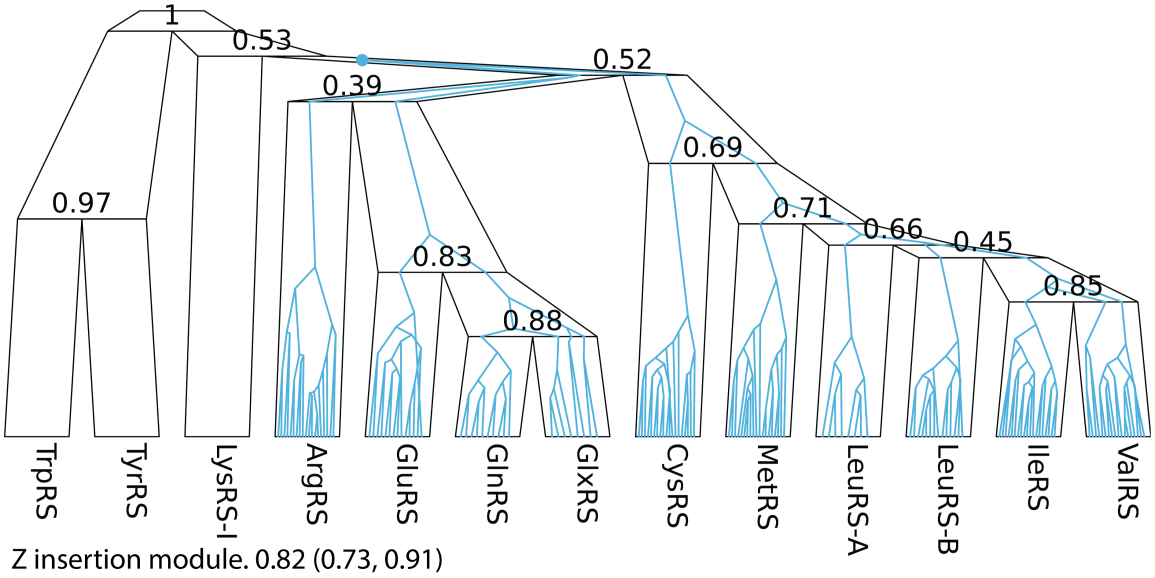

Fig. S6: Phylogeny of the Z insertion module. The root of the coloured tree is the last common ancestor of CP2 after its origin.

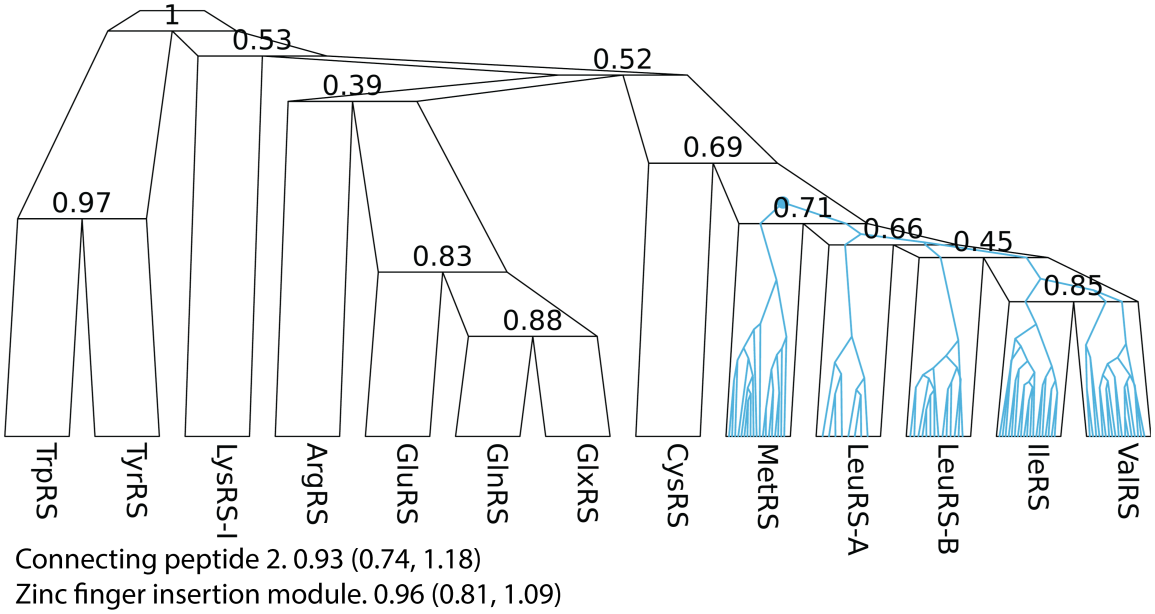

Fig. S7: Phylogeny of CP2 and the zinc finger insertion modules.

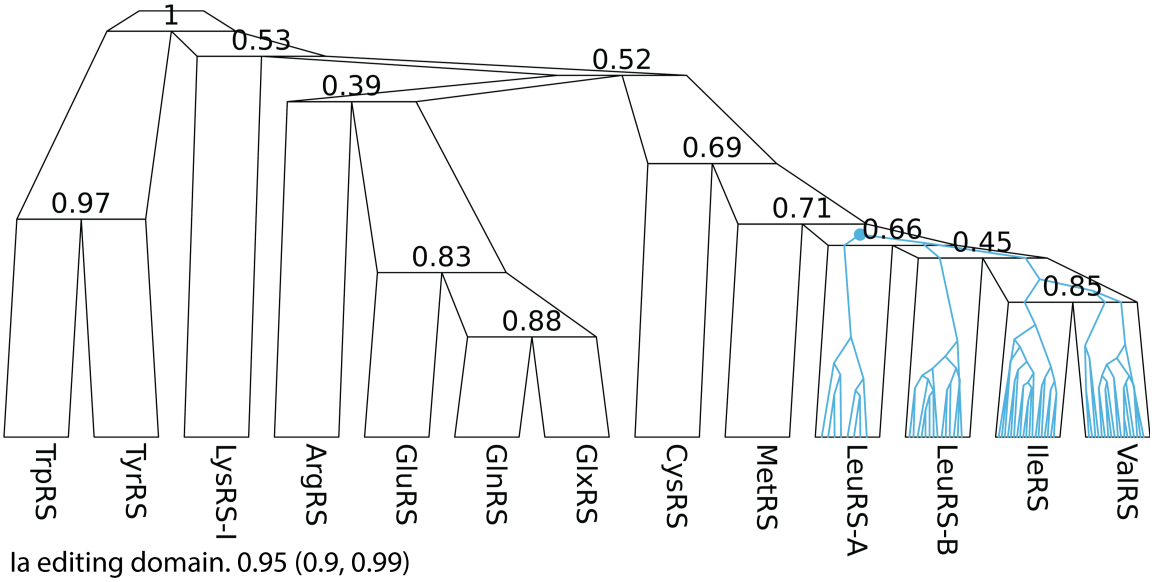

Fig. S8: Phylogeny of the Ia editing domain.

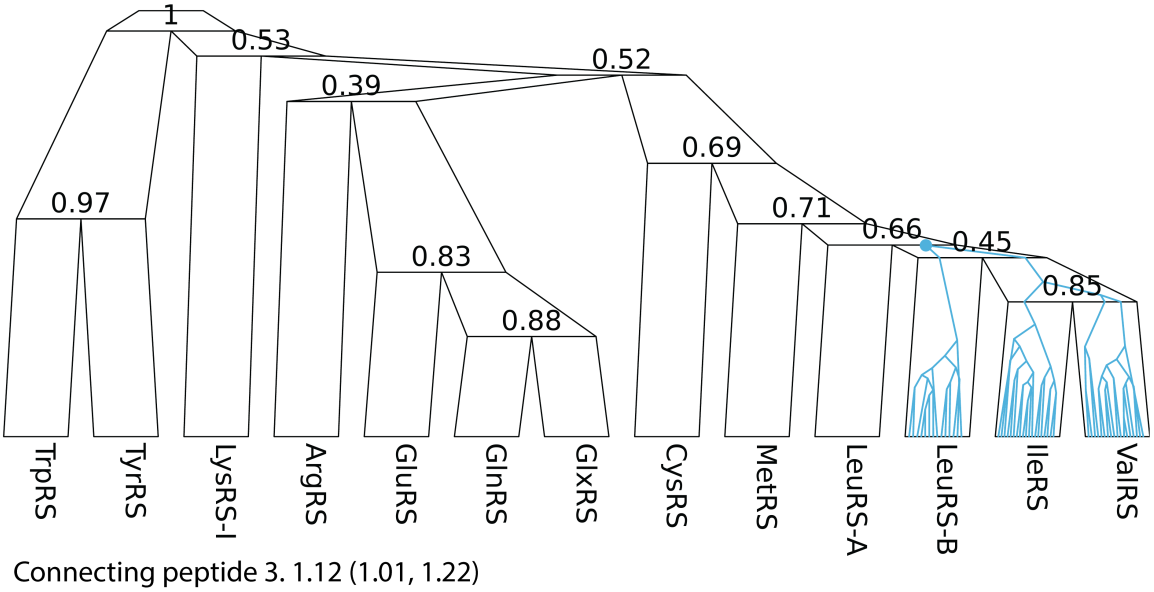

Fig. S9: Phylogeny of CP3.

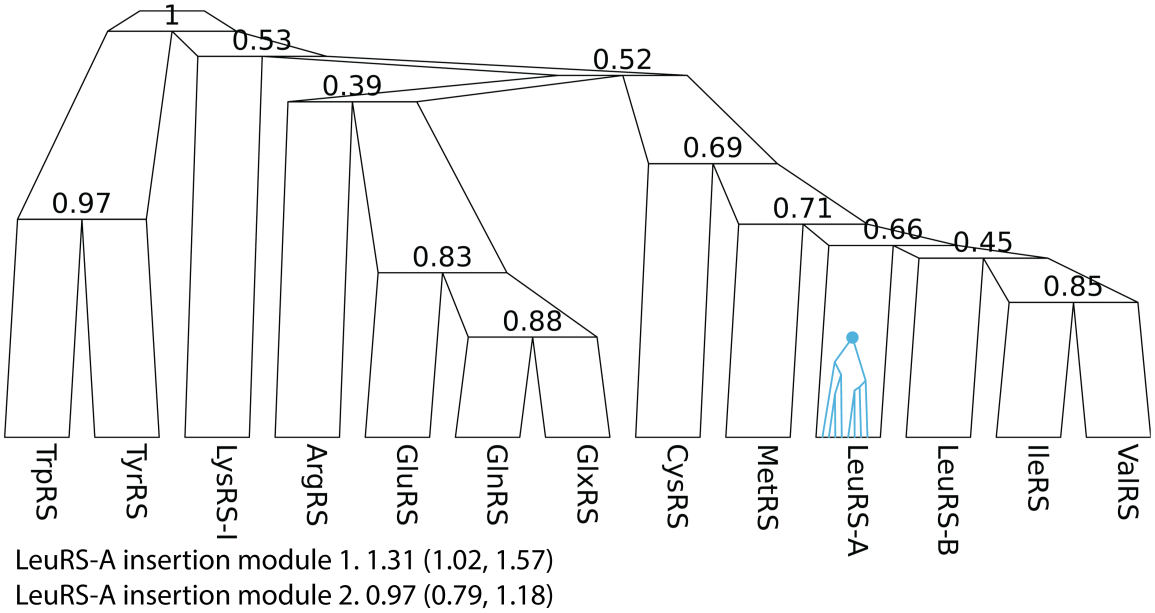

Fig. S10: Phylogeny of the two insertion modules unique to LeuRS-A.

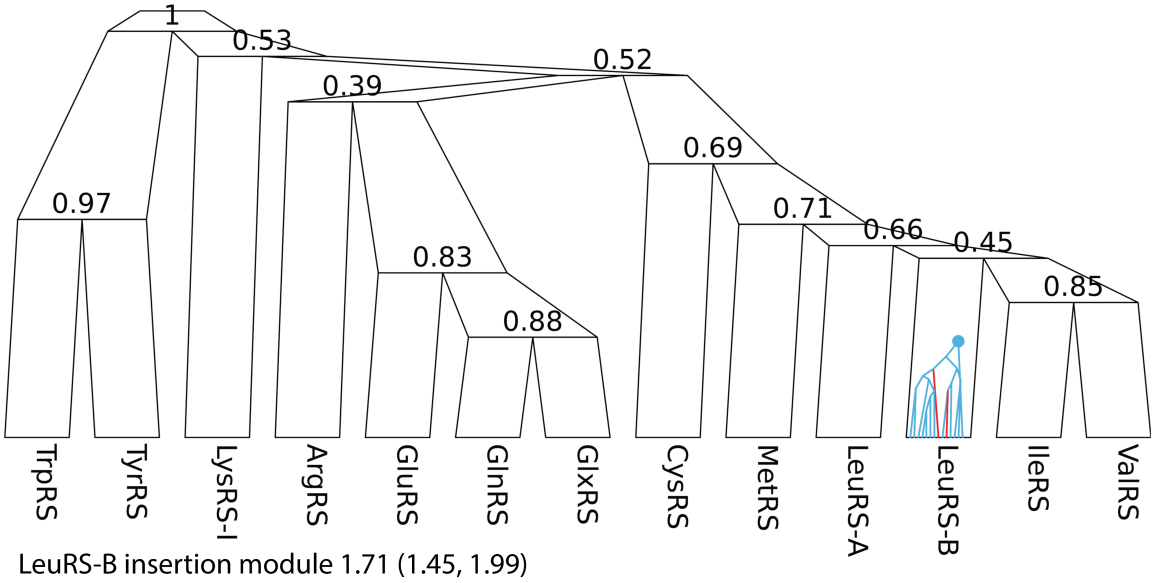

Fig. S11: Phylogeny of the LeuRS-B IM. The two lineages in red are lacking the insertion module, due to deletion events. These two bacterial taxa are *Clostridium botulinum* and *Mycoplasma hyopneumonia*.

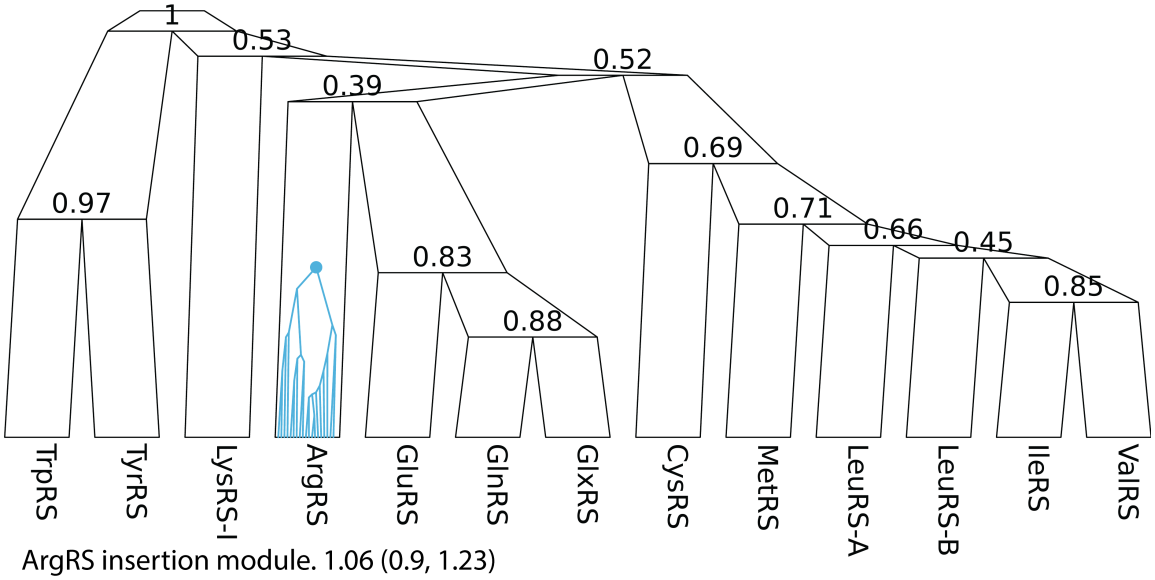

Fig. S12: Phylogeny of the ArgRS IM.

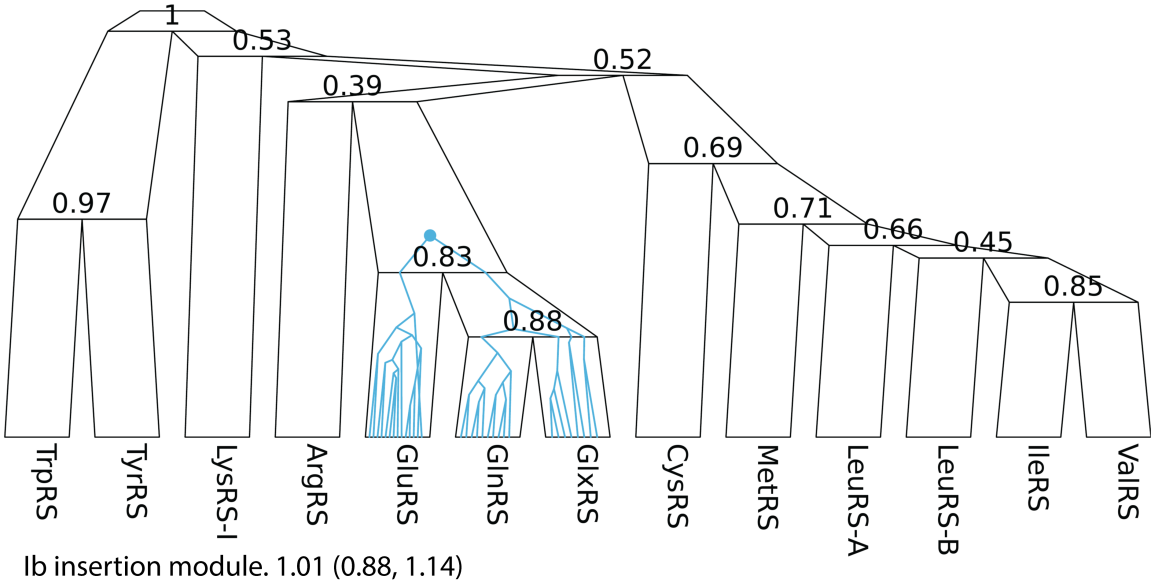

Fig. S13: Phylogeny of the *Ib* IM.

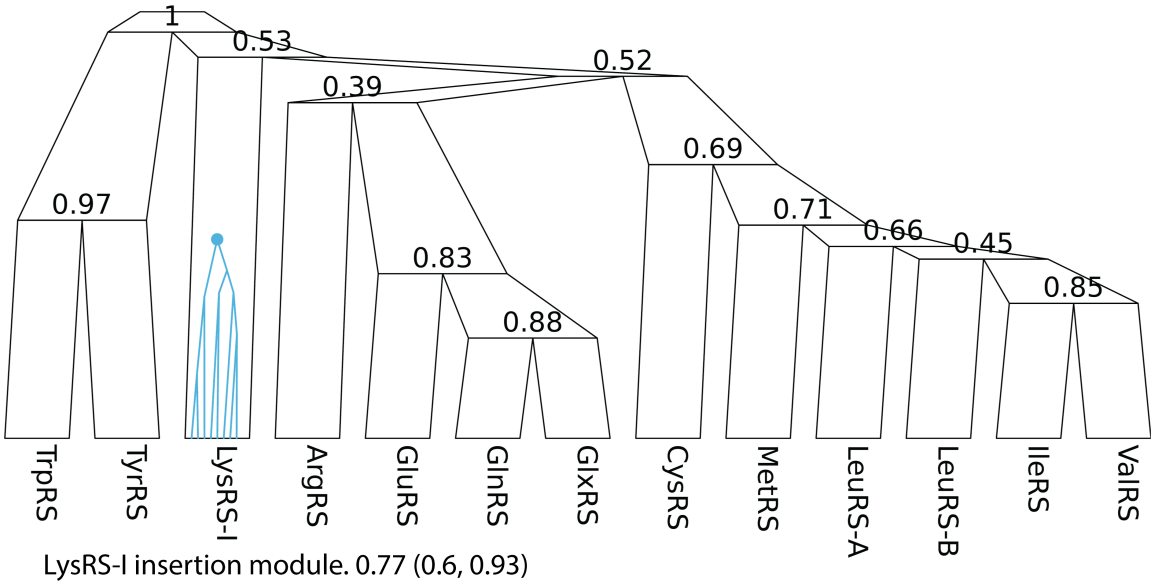

Fig. S14: Phylogeny of the LysRS-I IM.

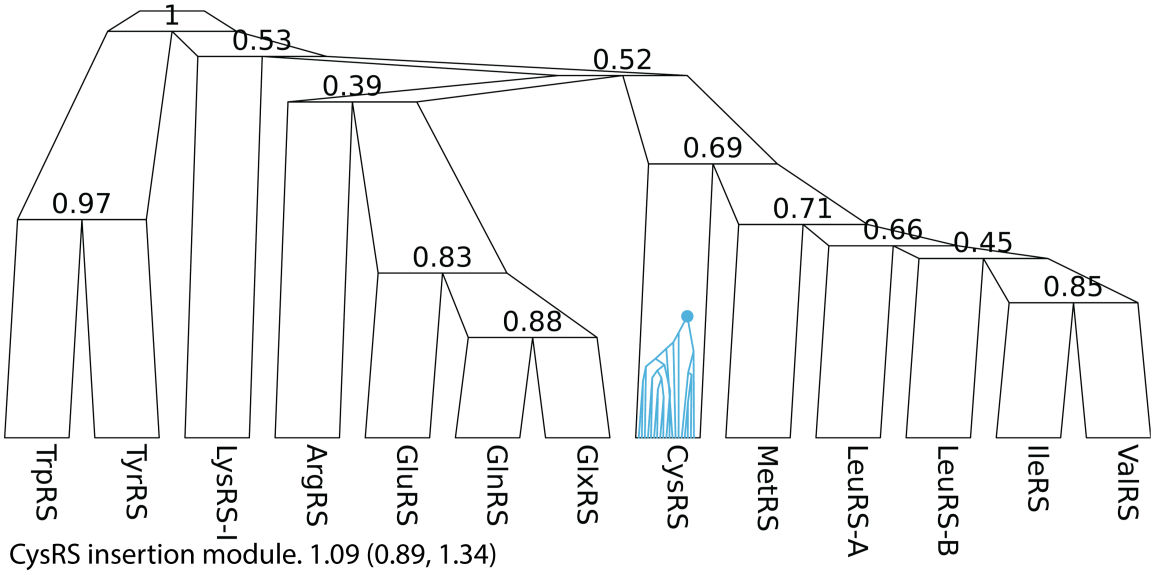

Fig. S15: Phylogeny of the CysRS IM.

Class II Insertion Module Phylogenies

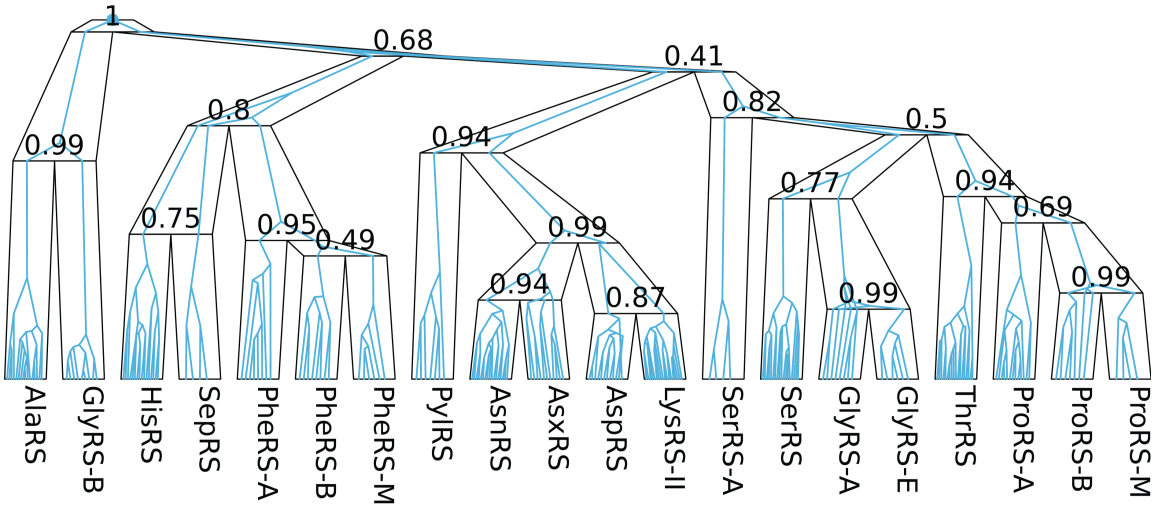

Fig. S16: Phylogeny of the three structural modules common to Class II and their estimated relative amino acid substitution rates (and 95% credible interval). Family tree internal nodes are labelled by posterior support. All of the consequent Class II modules are assumed to have a phylogeny which is a clade of the coloured phylogeny above.

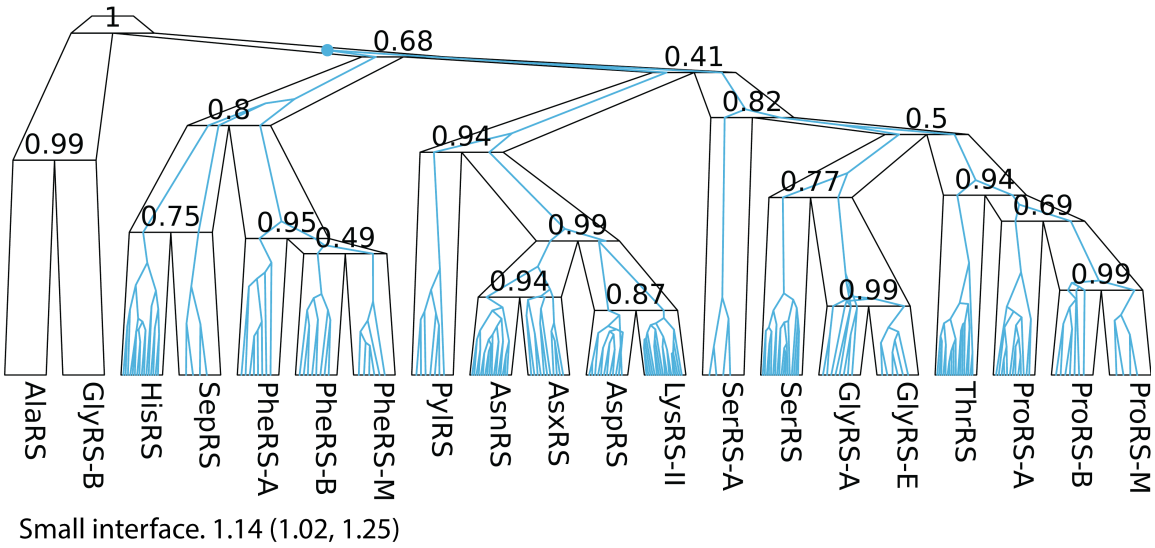

Fig. S17: Phylogeny of the small interface.

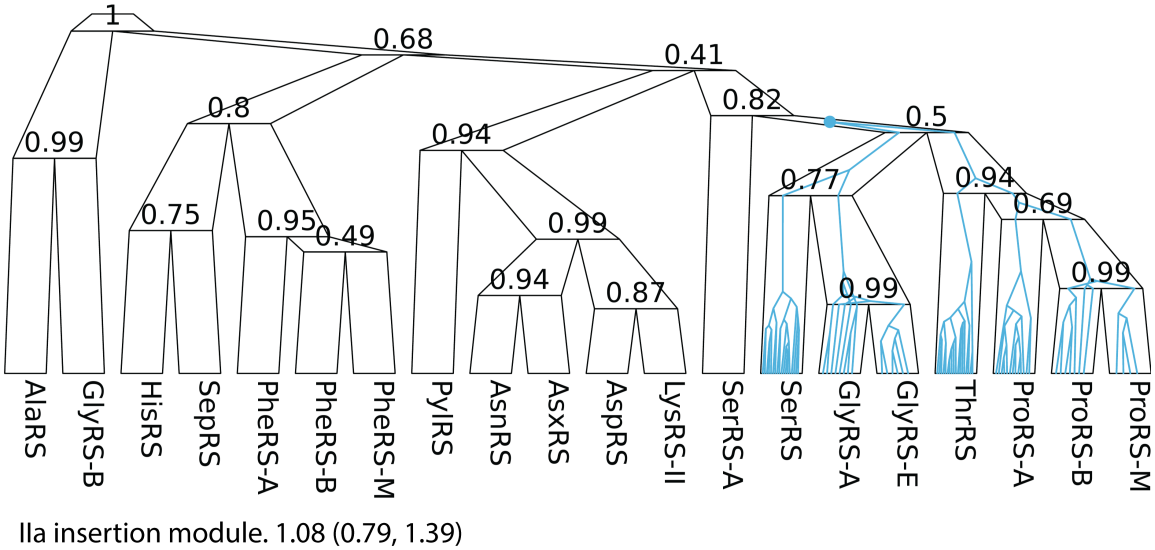

Fig. S18: Phylogeny of the *IIa* IM.

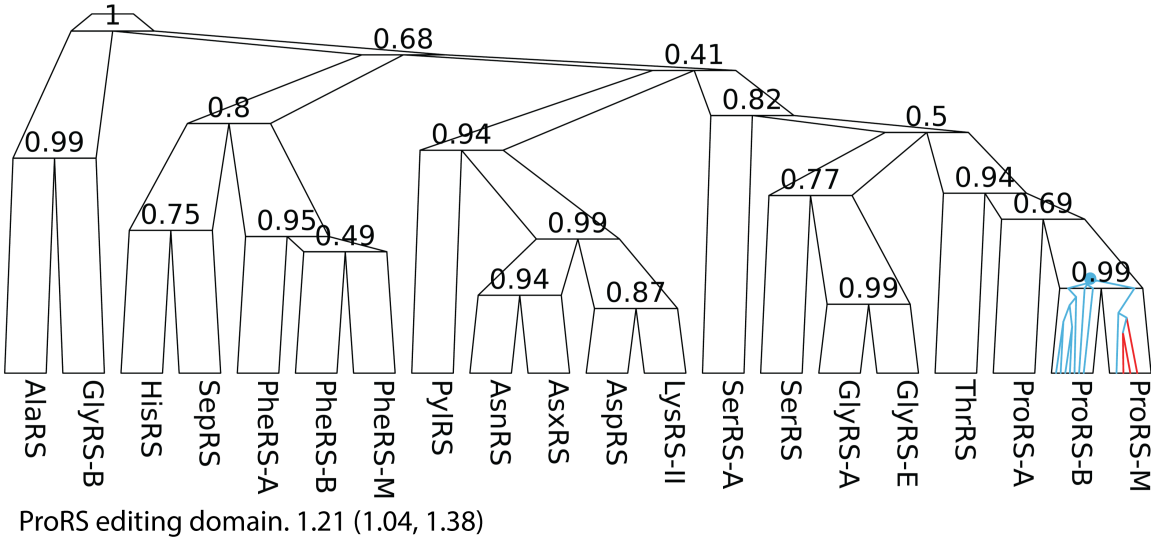

Fig. S19: Phylogeny of the ProRS editing domain. The lineages in red are lacking the insertion module, due to a deletion event. These mitochondrial taxa are *Drosophila melanogaster*, *Mauremys reevesii*, and *Schizosaccharomyces pombe*.

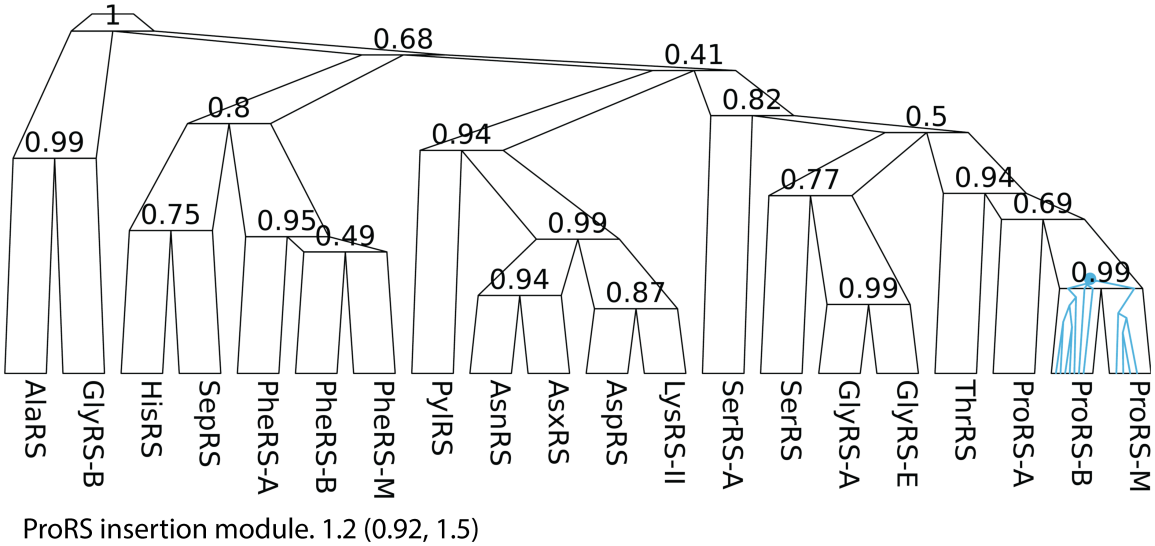

Fig. S20: Phylogeny of the ProRS IM, N-terminal of the ProRS editing domain.

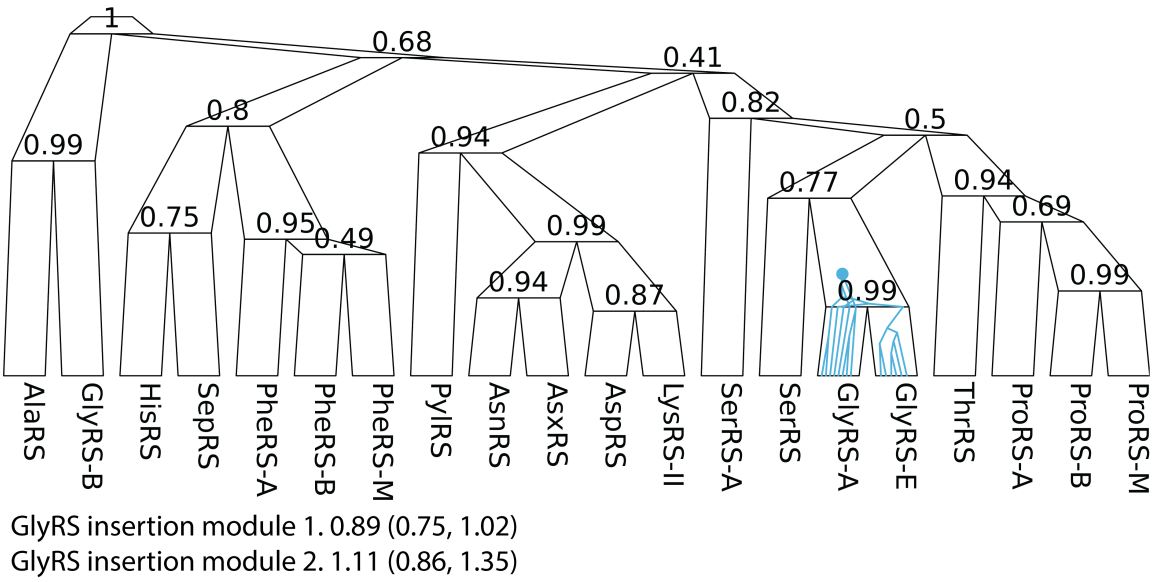

Fig. S21: Phylogeny of GlyRS IM 1 and 2.

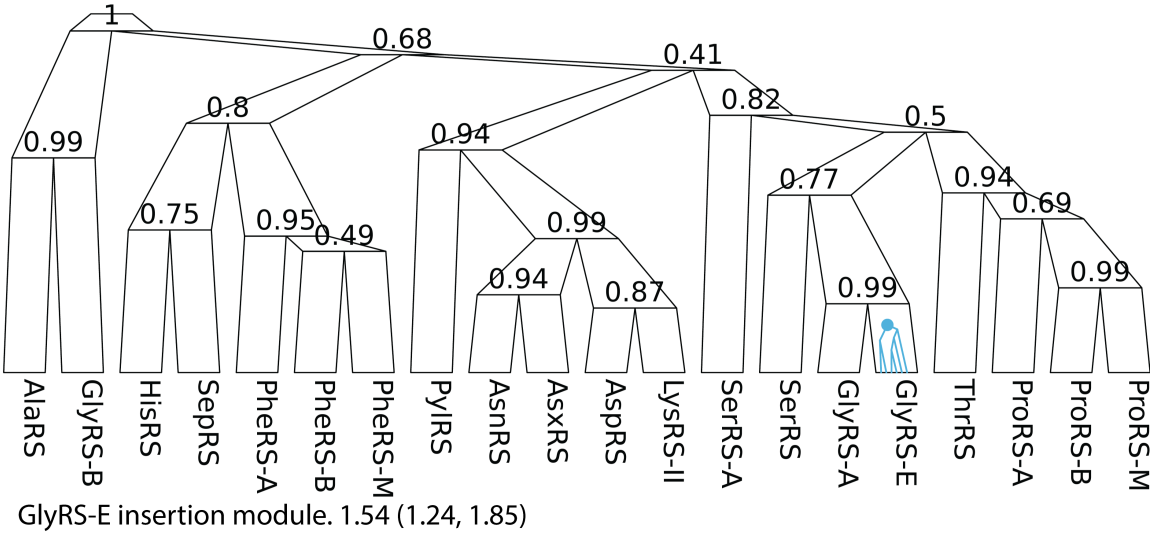

Fig. S22: Phylogeny of the GlyRS-E IM.

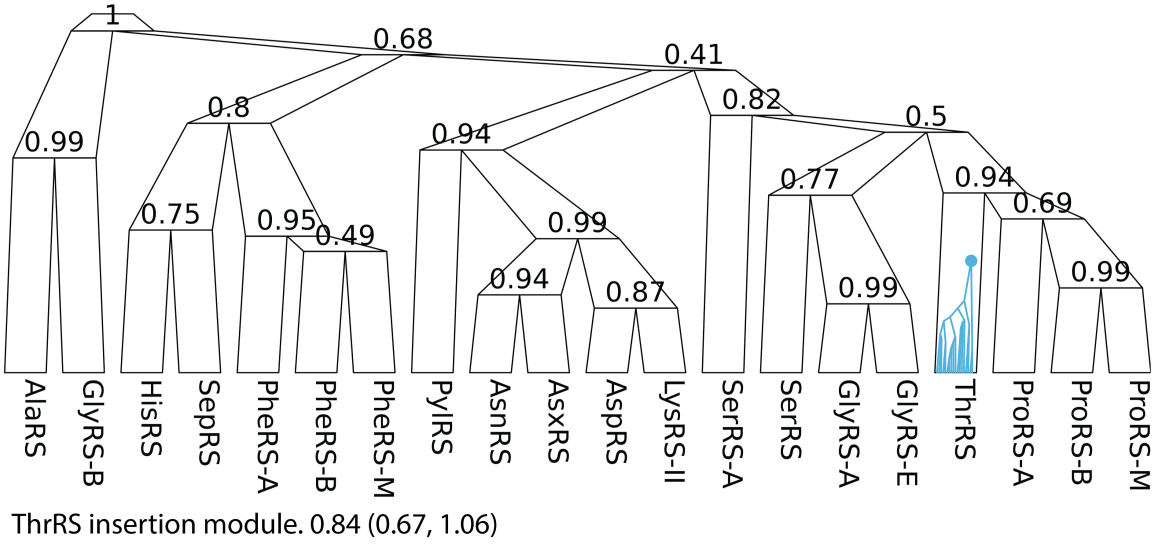

Fig. S23: Phylogeny of the ThrRS IM.

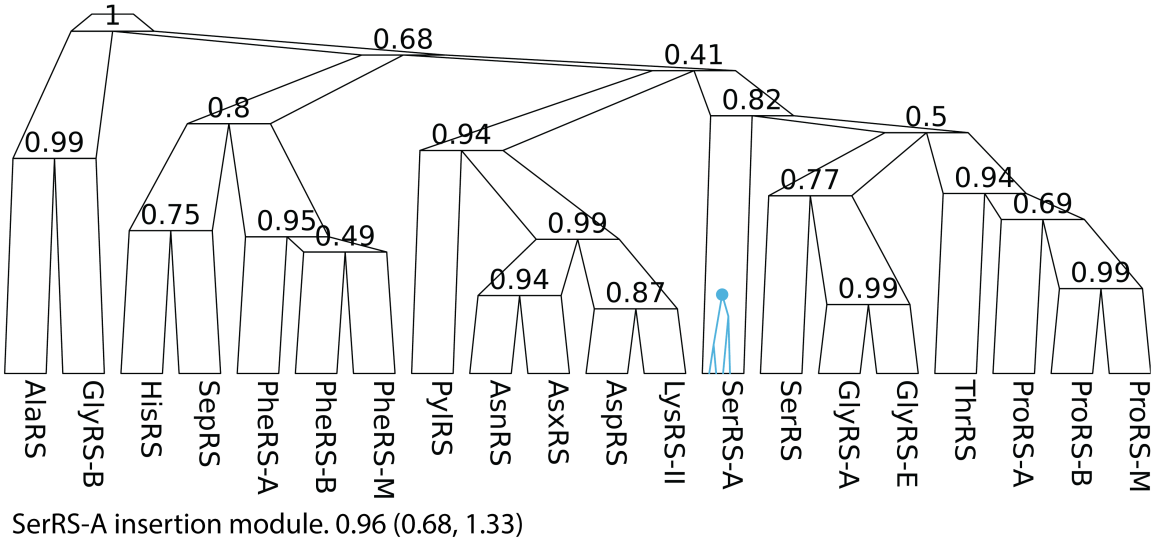

Fig. S24: Phylogeny of the SerRS-A IM.

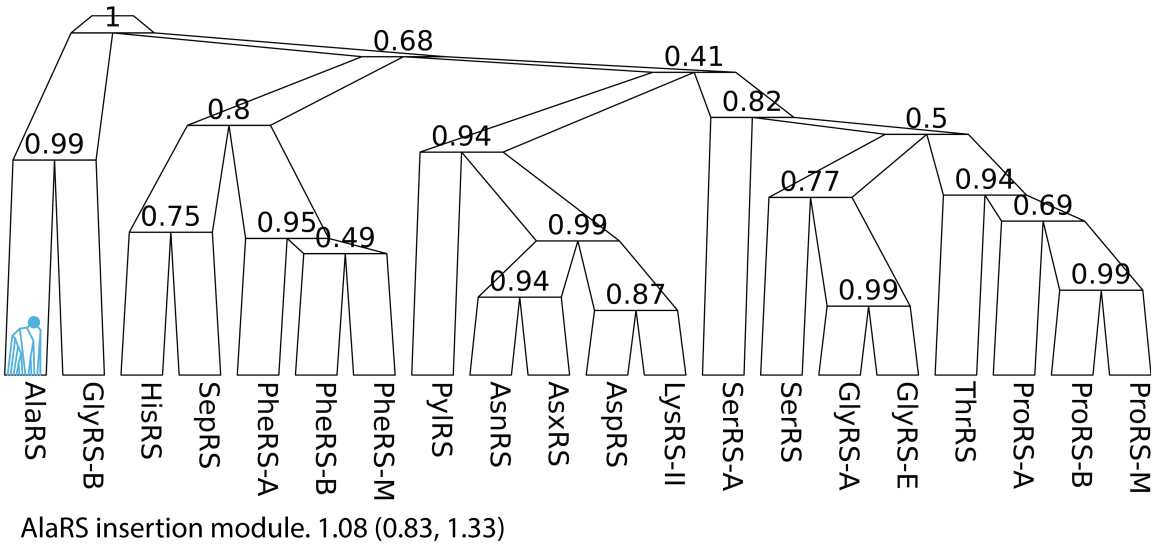

Fig. S25: Phylogeny of the AlaRS IM.

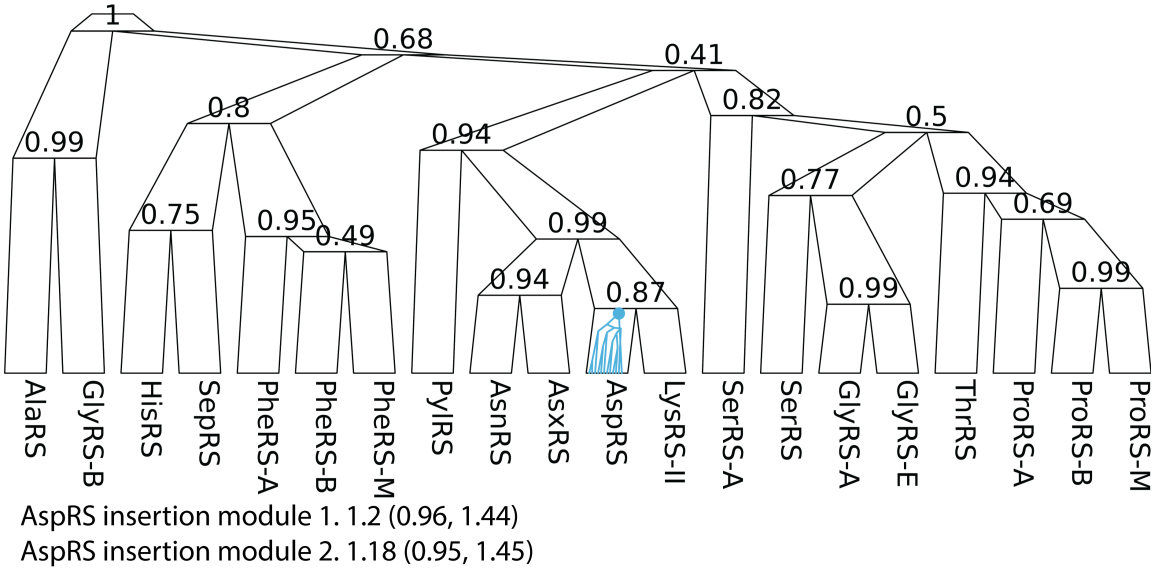

Fig. S26: Phylogeny of AspRS IM 1 and 2.

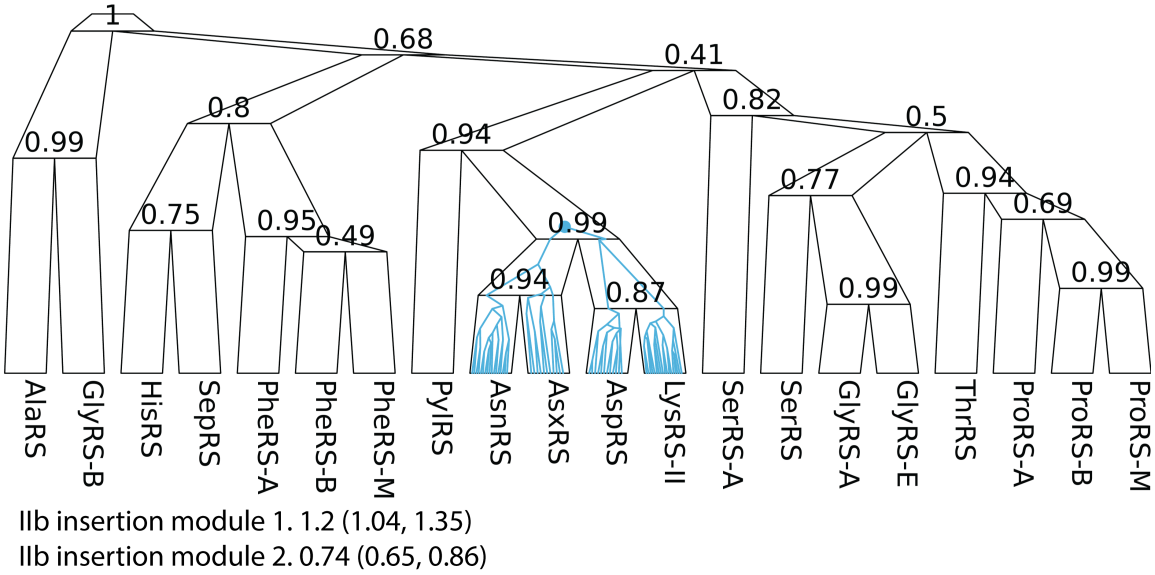

Fig. S27: Phylogeny of *IIb* IM 1 and 2.

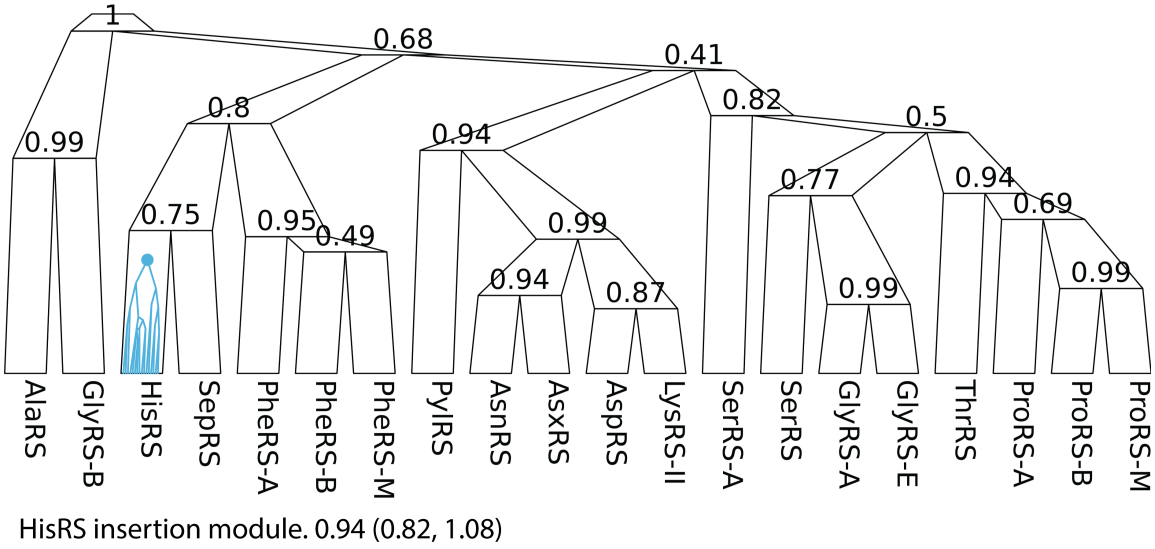

Fig. S28: Phylogeny of the HisRS IM.

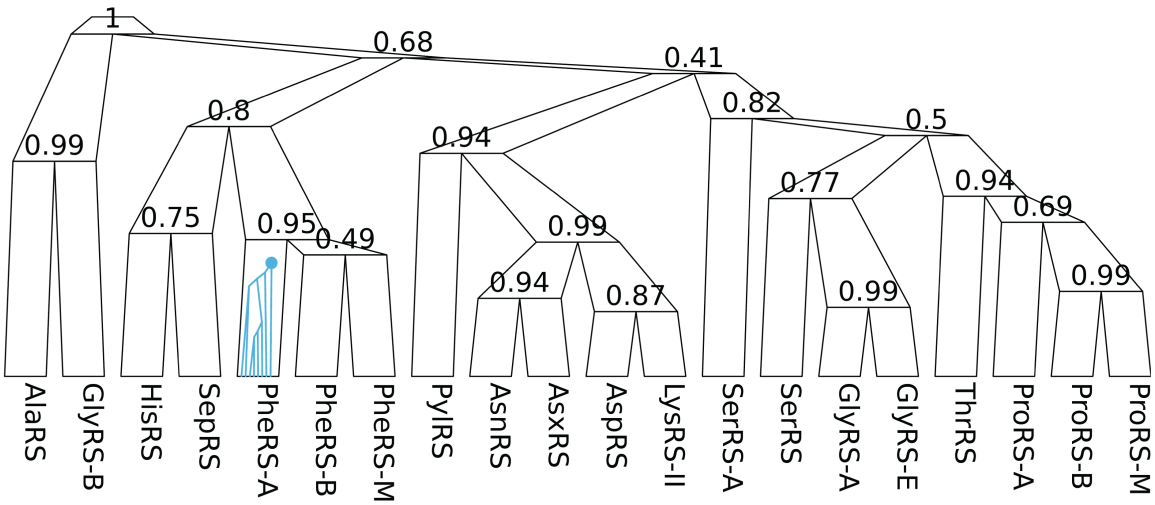

Fig. S29: Phylogeny of the PheRS-A IM.

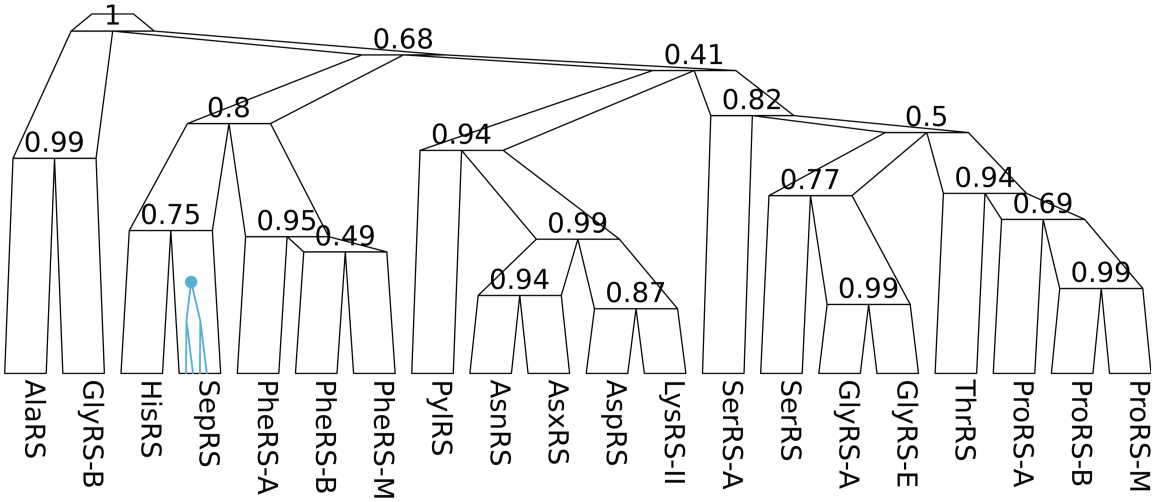

Fig. S30: Phylogeny of the SepRS IM.

References

[1] Bouckaert R, et al. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLOS Comput Biol 10(4):e1003537.

[2] Heled J, Bouckaert RR (2013) Looking for trees in the forest: summary tree from posterior samples. BMC Evolutionary Biology 13(1):221.

[3] Douglas J, Zhang R, Bouckaert R (2021) Adaptive dating and fast proposals: Revisiting the phylogenetic relaxed clock model. PLoS computational biology 17(2):e1008322.

[4] Wang S, Ma J, Peng J, Xu J (2013) Protein structure alignment beyond spatial proximity. Scientific reports 3(1):1–7.

[5] Müller T, Vingron M (2000) Modeling amino acid replacement. Journal of Computational Biology 7(6):761–776.