Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

Type of the Paper (Article.)

High-Resolution, Multidimensional Phylogenetic Metrics Identify Class I Aminoacyl-tRNA Synthetase Evolutionary Mosaicity and Inter-modular Coupling

Charles W. Carter, Jr[1], Alex Popinga[2], Remco Bouckaert[2], and Peter R. Wills[3]

[1]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260
ORCID ID (Carter): /0000-0002-2653-4452
[2]Centre for Computational Evolution, University of Auckland, PB 92019, Auckland 1142, New Zealand
ORCID ID (Bouckaert): /0000-0001-6765-3813
[3]Department of Physics and Te Ao Marama Centre for Fundamental Inquiry, University of Auckland, PB 92019, Auckland 1142, New Zealand
ORCID ID (Wills): /0000-0002-2670-7624
* Correspondence: carter@med.unc.edu; Tel.: 1-919-966 3263

Abstract: The provenance of the aminoacyl-tRNA synthetases (aaRS) poses challenging questions because of their role in the emergence and evolution of genetic coding. We investigate evidence about their ancestry from curated structure-based multiple sequence alignments of a structurally invariant "scaffold" shared by all 10 canonical Class I aaRS. Three uncorrelated phylogenetic metrics—residue-by-residue conservation, its variance, and row-by-row cladistic congruence—imply that the Class I scaffold is a mosaic assembled from distinct, successive genetic sources. These data are especially significant in light of: (i) experimental fragmentations of the Class I scaffold into three partitions that retain catalytic activities in proportion to their length; and (ii) evidence that two of these partitions arose from an ancestral Class I aaRS gene encoding a Class II ancestor in frame on the opposite strand. Phylogenetic metrics of different modules vary in accordance with their presumed functionality. A 46-residue Class I "protozyme" roots the Class I molecular tree prior to the adaptive radiation of the Rossmann dinucleotide binding fold that refined substrate discrimination. Such rooting is consistent with near simultaneous emergence of genetic coding and the origin of the proteome, resolving a conundrum posed by previous inferences that Class I aaRS evolved long after the genetic code had been implemented in an RNA world. Further, pinpointing discontinuous enhancements of aaRS fidelity establishes a timeline for the growth of coding from a binary amino acid alphabet..

Keywords: BEAST2; DensiTree; protein mosaic structure; RNA World hypothesis

1. Introduction

The emergence of the aminoacyl-tRNA synthetases, aaRS, is a quintessential chicken and egg puzzle whose solution would demystify the origins of coded protein synthesis. How did aaRS enzymes gain the reflexive property of collectively being able to use relationships in the universal genetic code to convert the sequences of base triplets in their own genes into functional amino acid sequences that make the code work? The detailed trajectory by which genes for the two essential superfamilies appeared, acquired catalytic proficiencies and radiated to refine their dual amino acid and cognate tRNA specificities is thus a crucial chapter in the book of life.

Both aaRS Classes [1-3] have separate catalytic and anticodon-binding domains. Only the catalytic domains within each superfamily share the same

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

architectures [4]. Anticodon-binding domains are idiosyncratic and, by consensus probably have a distinct evolutionary origin [5]. We develop high-resolution phylogenetic metrics here as a quantitative framework to show that even the shared architectures of the large, variable catalytic domains are probably mosaics assembled from far smaller peptides.

Phylogenetic clades for each amino acid in the Class I aaRS superfamily tree [6-12] are monophyletic and divide into three subclasses [4, 13]. Subclass IA includes IleRS, ValRS, LeuRS, and MetRS; subclass IB GluRS and GlnRS; and Subclass IC TyrRS and TrpRS. CysRS and ArgRS, assigned originally to Subclass IA [13], are difficult to assign, with one or the other appearing instead with Subclass IB [for example, see 9].

Genetic deconstruction and experimental biochemistry have suggested significant mosaicity within the Class I aaRS catalytic domains (Figure 1). Two segments nested within these domains—protozymes (from πρωτο = first [14]) and urzymes (from Ur = original, authentic [15-17])—represent successive intermediate evolutionary states of increasing length and dating from well before the Last Universal Common Ancestor [18, 19].

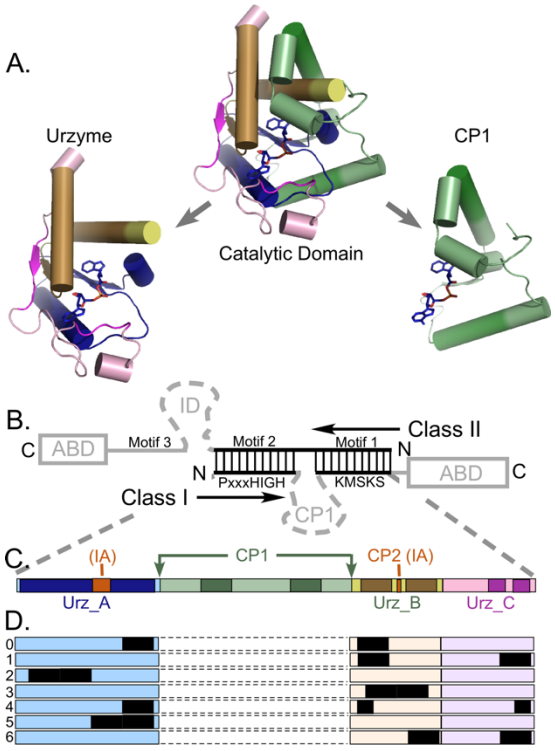


Figure 1. Structural modules underlying the hypothesis and data organization. **A**. Deconstruction of Class I urzyme and internal CP1 insertions that together make up the Class I catalytic domain. Cartoons were prepared with Pymol [20] from coordinates (PDB ID 1I6L) for the tryptophanyl-tRNA synthetase, the smallest Class I aaRS. Secondary structures displayed here are conserved in all 10 Class I aaRS. The activated aminoacyladenylate is drawn as sticks. Size variation in other Class I aaRS arises from further insertions within CP1. Anticodon-binding domains are idiosyncratic, and not shown. **B**. Overlapping portions of Class I and II aaRS as envisioned in an ancestral bidirectional gene [21] coincide with the respective urzymes [17, 22]. Vertical lines denote ancestral base-pairing between the respective genes. Grey segments were presumably more recent additions. Insertion of CP1 is incompatible with bidirectional coding. **C**. Schematic location of CP1 between two roughly equal fragments of the urzyme, colored to allow identification of structural fragments in **A**. Intensely colored segments are highly conserved secondary structures in all 10 Class I aaRS and compose the Class I scaffold. Subclass I A enzymes contain one or more additional insertions (red). **D**. Sampled alignments consisting of different segments totaling 20 amino acids, selected as described in Methods. **C** and **D** amplify the Class I base-paired portion of **B**. The protozyme (Urz_A) is the amino terminal β-α-β crossover connection (blue), the amino acid binding pocket (Urz_B) is formed by the protozyme and two intermediate α-helices (amber). The KMSKS loop (Urz_C) is rose. Previous work [23] established that the CP1 insertion (green) has little impact on the enzymatic properties of the TrpRS urzyme unless the anticodon-binding domain is also present. Colors match those in (**A**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

Urzymes [24] are cores whose ~130 residues form a nearly intact active site. Biochemical experiments show that urzymes retain ~60 % of full-length aaRS catalytic proficiency—estimated as the transition-state stabilization free energy—for both amino acid and tRNA substrates [16, 25, 26] (Figure 1). Class I and II urzymes also differentiate between the corresponding two sets of substrates, with Class I urzymes activating Class I, in preference to Class II amino acids by ~1.0 kcal/mole, and conversely [22, 27, 28].

Protozymes are 46 residue subsets from both Class I and II Urzymes that retain nearly half the full catalytic proficiency in the critical amino-acid activation reaction [14], whose uncatalyzed rate is rate-limiting for protein synthesis. The Class I protozyme coincides with the N-terminal crossover connection of the Rossmann fold, which forms the ATP binding site and contains a distinctive 3D packing motif that recurs in ~25% of the proteome [29] and imposes distinct, functionally relevant conformational states tightly coupled to catalysis in TrpRS [23, 30-35].

Fully active Class I and II protozymes have been expressed from a single gene designed to encode their structures on opposite strands. Experimental Michaelis-Menten parameters of all four protozymes—Class I and II; native and bidirectional—are, remarkably, the same within experimental error [14]. Tamura's laboratory [36] have replicated those results. Experimental [14, 17, 37] and bioinformatic evidence [38] therefore support the hypothesis of Rodin and Ohno [21, 39] that the two aaRS Classes descended from opposite strands of a single bidirectional gene.

Structural conservation, high catalytic proficiency in both essential reactions, and specificity for amino acids from the appropriate class all reinforce their role as experimental for different stages of aaRS molecular ancestry.

"Connecting peptide 1" or CP1 [40, 41] intersects the Class I aaRS Rossmann fold immediately after the protozyme between structurally homologous residues that are ~4.5 Å apart. CP1 can thus be replaced by a peptide bond without structural disruption to produce the Class I urzyme [15, 17] as detailed in Figure 1. CP1 insertions include the editing domains of the aaRS for aliphatic amino acids, and thus largely account for the variable size of Class I catalytic domains. Independent 3D superposition of aaRS crystal structures [6, 17, 42] revealed considerable structural conservation within the catalytic domains of all ten members of each Class—see Figure 1 of [42]. Surprisingly, structural homology across the Class I superfamily extends beyond the urzyme, into CP1. We call the secondary structures within these conserved cores "scaffolds".

Evidence for descent of Class and II aaRS from a bidirectional gene implies, *ipso facto*, that sequences inconsistent with bidirectional coding, like CP1, represent accretion of new genetic material. The CP1 insertions are incompatible with, and their introduction would necessarily have ended, bidirectional genetic coding of ancestral Class I and II aaRS (Figure 1B). Class I urzyme boundaries delimited decisively by potential bidirectional coding of Class II urzymes constitute only about 80% of the Class I scaffold; the remainder consists of 10-residue α-helical segments near the beginning and end of the CP1 insertion (Figure 1). Thus, if the Rodin-Ohno hypothesis is correct, then the CP1 insert must derive from a distinct genetic source.

To address the apparent contradiction between the Rodin-Ohno hypothesis and extended conservation into the CP1 segment, we introduce four phylogenetic metrics that, together, reinforce the conclusion that modular components within the structurally invariant segments of the ancestral aaRS have different genetic histories:

(i) We threaded sequences into closely-related crystal structures to assemble multiple sequence alignments (MSAs) based on three-dimensional structure superposition, to avoid depending on amino acid substitution matrices to define equivalent positions.

(ii) We develop multi-dimensional phylogenetic metrics from the ensemble of phylogenetic trees obtained by Markov Chain Monte Carlo (MCMC) simulations to compare segments of the MSA. We show that these three metrics are uncorrelated, and demonstrate their statistical significance by multiple regression.

(iii) We increase the effective analytical resolution by applying the metrics to partitions of the MSA that have been extensively characterized experimentally, providing novel insight into the functional modularity of the superfamily and molecular phylogenetic support that they are successive evolutionary intermediates in the generation of the Class I aminoacyl-tRNA synthetase superfamily.

(iv) We identify key two-way interactions between mutation rates in different MSA segments. Both involve the amino acid binding site and one is central to the enhancement of amino acid specificity enabled by the CP1 insertion.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

These results support important modifications of conventional evolutionary scenarios for major parts of the proteome containing Rossmann dinucleotide fold domains, and strengthen the proposal that the genetic code development is coupled intimately to the structural evolution of aaRS.

2. Results

Phylogenetic metrics we will discuss are given in Table 1. They were derived from trees constructed for different subset MSAs—CP1; the urzyme; and its three distinct modules Urz_A (the protozyme), Urz_B (the amino acid binding site, and Urz_C (the PP_i binding site). They furnish unprecedented insight into the genetic modularity of the Class I aaRS. In summary, they suggest that CP1 is derived from a more recent and less cladistically coherent genetic entity, as previously posited [17]. Similarly, they suggest that the protozyme may be from an older genetic entity than other segments of Class 1 urzymes.

Table 1. Design matrix for regression analyses. Shaded columns are dependent variables; unshaded columns are independent variables that serve as predictors. Numerals in columns A, B, C, and CP1 are proportional to the total number of amino acids in the respective MSAs. Two additional independent variables, urzyme and protozyme, constructed in a related fashion, are not shown. NUMB is the number of amino acids in the alignment.

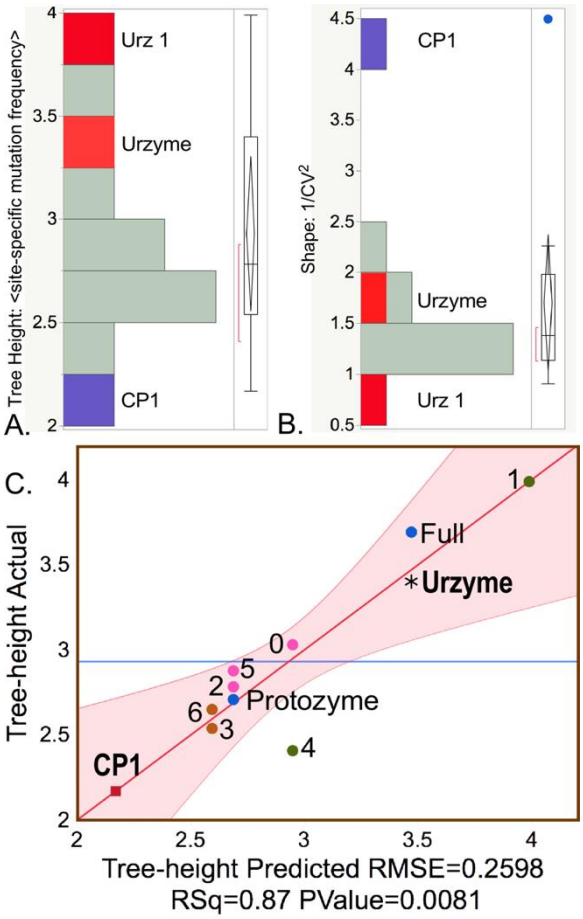
Subset	<S>$_{WAG}$	<S>$_{LG}$	<Q>	Tree Height	Shape	A	B	C	CP1	NUMB
Full	0.9	0.86	48.8	3.69	1.17	4	3	1	2	103
urzyme	0.87	0.82	52.4	3.40	1.60	4	3	1	0	83
Protoz	0.85	0.85	52.3	2.71	1.38	4	0	0	0	46
CP1	0.34	0.35	33.7	2.17	4.49	0	0	0	2	20
Urz_0	0.76	0.73	50.1	3.03	1.46	1	1	0	0	20
Urz_1	0.71	0.68	57.8	3.99	0.91	0	1	0.5	0	20
Urz_2	0.72	0.7	55	2.78	1.13	2	0	0	0	20
Urz_3	0.62	0.62	45.6	2.54	2.26	0	2	0	0	20
Urz_4	0.82	0.77	58.5	2.41	1.21	2	1	0.5	0	20
Urz_5	0.78	0.8	45.6	2.88	1.98	4	0	0	0	20
Urz_6	0.62		56.6	2.65	1.13	0	2	1	0	20

2.1 *CP1 has lower and more uniform apparent site-to-site mutation rates between aaRS for different amino acids.*

The BEAST 2 MCMC algorithm tracks the extent of site-to-site evolutionary of sequence variation in using the Tree Height and Shape metrics described in **Methods**. The Tree Height metric is summarized in (Figure 2A, B). Sequences responsible for the elevated urzyme Tree Height are identified by regression against the independent parameters of the design matrix (Table 1) in Fig 2C. Tree Height $= 2.17 - 1.4 * Segment\ B + 3.2 * Urzyme - 2.7 * protozyme + 1.7 * Segment\ B * protozyme$. All β coefficients are significant with P-values < 0.005. Foremost among the positive contributors is the urzyme. However, the interaction between the Protozyme (segment A) and segment B is also quite significant. We note that CP1 sequences exhibit substantially smaller Tree Height and variance (i.e. higher Shape), consistent with it being a more recent genetic entity.

The negative logarithm of the Shape parameter is highly correlated with the conservation quality, $\langle Q \rangle$, defined by Clustal [43, 44] (Figure 3A), lending intuitive insight into the physical meaning of the latter. The designation "Conservation Quality" is apparently misleading in suggesting that the significantly smaller $\langle Q \rangle$ value for the CP1 MSA (Figure 3B) implies that it is less well conserved, in apparent conflict with its reduced Tree Height. In fact, the colinearity of –log(Shape) and $\langle Q \rangle$ led us to pursue the equivalence between the two metrics. Fig 3C, D show that the two metrics depend in nearly identical fashion on predictor columns from the design matrix in Table 1. It appears then that the $\langle Q \rangle$ metric does not measure mutational variation itself, but rather the inverse of its variance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

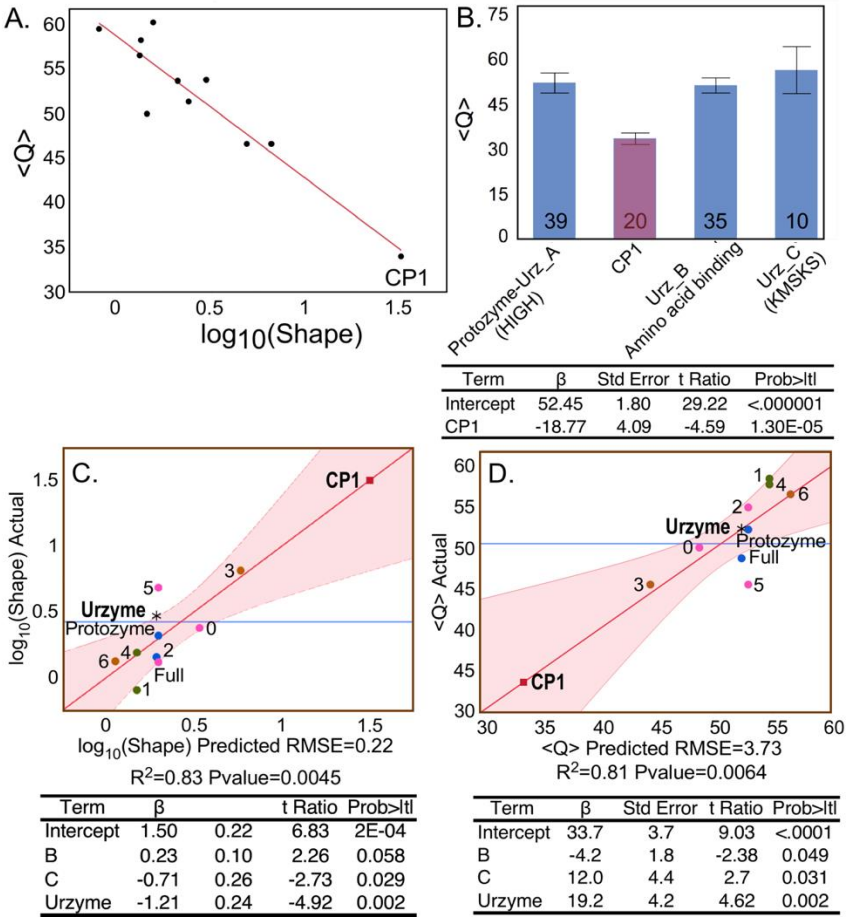


Term	β	Std Error	t Ratio	Prob>|t|
Intercept	2.17	0.26	8.35	0.0002
B	-1.39	0.32	-4.38	0.0047
Urzyme	3.21	0.61	5.27	0.0019
Protozyme	-2.70	0.57	-4.75	0.0031
B*Protozyme	1.66	0.33	5.05	0.0023

Figure 2. Histograms of the Tree Height (**A**) and Shape (**B**) metrics highlight differences between sequence variability within CP1 (blue) and urzyme (red) sequences. The Tree Height is the reciprocal of the estimated mutation rate per site. CP1 has the lowest site-specific mutation rate and the highest Shape parameter (blue dot implies statistical significance). The Urz_1 subset of amino acids includes a stretch of 10 amino acids along the specificity determining helix, where the highest site-specific mutation frequency occurs within the urzyme. **C**. Regression model showing the dependence of Tree Height on MSA subsets. The horizontal blue line in this and other regression curves is the average value of the dependent variable. Dots represent the various MSAs in the design matrix and are colored and distinguished by different symbols for identification. R^2 for this model is 0.87. Numerals refer to the 20 residue subsets defined in Fig 1. The F-ratio is 9.9 with a P-value of 0.008.

Moreover, the extended similarity between Figure 3C and Figure 3D suggest that the site-by-site metrics (Tree Height and Shape) can provide quite high resolution evidence on the evolution of modularity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1



Term	β	Std Error	t Ratio	Prob>|t|
Intercept	52.45	1.80	29.22	<.000001
CP1	-18.77	4.09	-4.59	1.30E-05

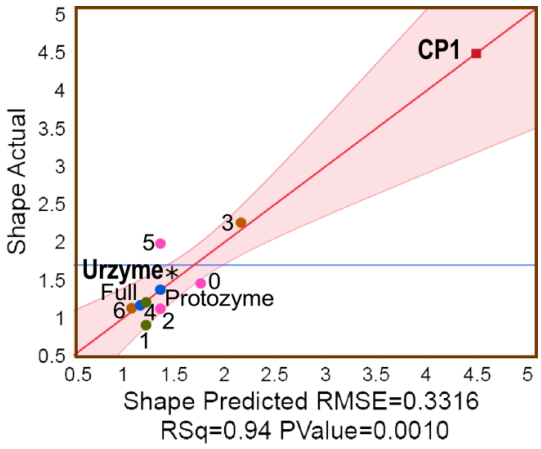
Term	β		t Ratio	Prob>|t|
Intercept	1.50	0.22	6.83	2E-04
B	0.23	0.10	2.26	0.058
C	-0.71	0.26	-2.73	0.029
Urzyme	-1.21	0.24	-4.92	0.002

Term	β	Std Error	t Ratio	Prob>|t|
Intercept	33.7	3.7	9.03	<.0001
B	-4.2	1.8	-2.38	0.049
C	12.0	4.4	2.7	0.031
Urzyme	19.2	4.2	4.62	0.002

Figure 3. Position-specific metrics. **A**. Colinearity of the logarithm of the Shape parameter of the gamma distribution of the Tree Heights estimated by BEAST2 for different MSAs derived from the Class I scaffold, and the mean conservation quality scores, $\langle Q \rangle$ [44] obtained by Clustal directly from respective MSAs. **B**. The conservation quality, <Q>, for the four segments of the Class I scaffold alignment. Class I signature peptides [4] are in parentheses. Error bars show the standard error of the mean over all positions within the segment. The numbers of amino acids in each segment are given for each histogram. **C**. and **D**. Regression models showing nearly identical dependence of log(Shape) and $\langle Q \rangle$, respectively, on the same predictors from the design matrix (Table 1). Dots represent different MSAs and are labeled and colored as in Figure 2 to emphasize the extraordinary similarity of site-by-site metrics derived two different ways.

Regression of Shape on the independent parameters of the design matrix in Table 1 (Figure 4) resulted in the unique model: $\text{Shape} = 1.37 + 0.4 * Segment\ B - 1.08 * Segment\ C + 1.56 * CP1 - 0.57 * Segment\ B * CP1$. All β coefficients are significant and two, shown in bold face, have P-values < 0.005.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1



Term	β	Std Error	t Ratio	Prob>|t|
Intercept	1.37	0.17	8.26	0.0002
B	0.40	0.16	2.45	0.0498
C	-1.08	0.40	-2.73	0.0342
CP1	**1.56**	**0.19**	**8.43**	**0.0002**
B*CP1	**-0.57**	**0.10**	**-5.96**	**0.001**

Figure 4. Shape parameter dependence. **A.** Plot of the multivariate regression model for Shape. $R^2 = 0.94$ for the regression and the P value for the F ratio is 0.0006. **B**. Regression coefficients and their Student t-test probabilities. Points for trees constructed using both WAG and LG substitution matrices are shown for comparison. Significant predictors of Shape are the amino acid positions in CP1, those in the B-fragment containing the amino acid specificity-determining helix (sand; Figure 1A), and their two-way interaction.

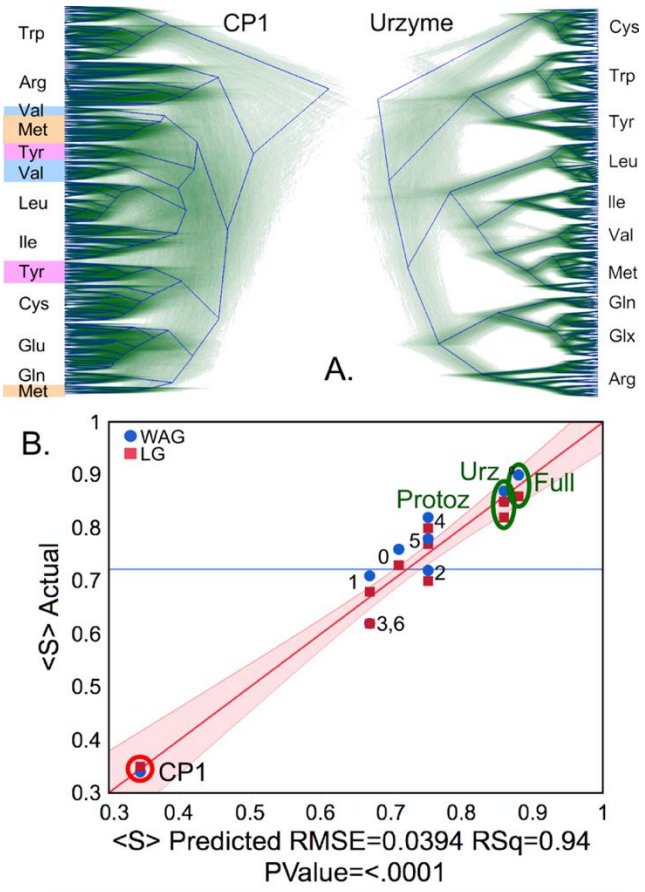
2.2 *Class I urzyme- and CP1-based trees form distinctly different clades.*

Urzyme and CP1 partitions of the MSAs produce substantially different trees (Figure 5A). In particular, although all ten Class I aaRS clades are monophyletic in the trees for the urzyme MSAs, three enzymatic clades in the CP1 trees—MetRS, ValRS, and TyrRS—are polyphyletic. Moreover, the urzyme clades are constrained by tight envelopes, whereas the CP1 envelopes are poorly defined.

Support, S_i, the fraction of all trees for which each aaRS clade, i, is monophyletic, was averaged over all Class I aaRS types to give the mean support, $\langle S \rangle = \sum S_i / 10$. That operation was repeated for trees built for the full scaffold MSA (Full), the urzyme, CP1, protozyme (segment A in Figure 1), and seven subset MSAs each consisting of twenty amino acids in blocks of five residues as described in Methods. For the eleven rows of the design matrix (Table 1), we computed $\langle S \rangle$ for populations of trees constructed using the conventional WAG [45] substitution matrix and repeated using the more recent LG [46] substitution matrix used in [12].

Contributors to the variance of $\langle S \rangle$ were assessed using stepwise multiple regression, which resulted in the unique model: $\langle S \rangle = 0.67 + 0.04 * protozyme - 0.33 * CP1 + 0.1 * protozyme * CP1$ summarized in Fig 5. All three coefficients are significant at better than the 99% confidence level, with P-values < 0.005. As suggested by its position in the regression curve in Figure 5, trees built from CP1 residues have significantly less support than those from the urzyme or any of its 20-residue subsets. The A segment coincides with the protozyme MSA; its dominant impact on the variance of $\langle S \rangle$, contributing positively both via its intrinsic effect and by its two-way interaction with CP1, is consistent with its being close to the root of the Class I aaRS superfamily tree.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1



Term	β	Std Error	t Ratio	Prob>|t|
Intercept	0.67	0.03	26.32	<.0001
Protozyme	0.04	0.01	4.34	0.0034
CP1	**-0.33**	**0.06**	**-5.87**	**0.0006**
Protozyme*CP1	**0.10**	**0.02**	**4.73**	**0.0021**

Figure 5. Phylogenetic support. **A.** DensiTree representations of the urzyme and CP1 segments of the Class I structural scaffold. Each aaRS is monophyletic in the urzyme alignment, whereas several of the clades in the CP1 alignment, highlighted in color, have multiple ancestries. **B.** Regression model for $\langle S \rangle_i$. Dots are colored and labeled as in Fig 3. Values in the Regression table are only for the WAG matrix, as use of both WAG and LG matrices unduly enhances the P-values of the β-coefficients. The R^2 and P value of the model's F-ratio are shown under the X-axis. Coefficients and their statistics for the model are given in Table 2. Individual $\langle S \rangle$ values are labeled to enable comparison with Fig 1C. R^2 was 0.94 for 20 observations, and the F-ratio for the regression table was 92.4 with a P-value <0.0001.

Thus, although the 3D structures from the urzyme and CP1 partitions have comparable structural homology, they have markedly different phylogenetic signatures.

2.3 *Phylogenetic metrics—Tree Height, Shape, and Support—reveal significant, high-resolution genetic mosaicity.*

Our quantitative evidence probes far deeper into evolutionary time than previous phylogenetic analyses of Class I aaRS [7, 11, 18, 19, 47]. That depth both calls for caution and is a source of great interest. Constructing aaRS trees is fundamentally ambiguous because at each node in any conceivable tree, the coding alphabet and dimension of the operational substitution matrix necessarily both change by integer steps as Class I and II trees branch into multiple families. Trees for the two superfamilies are thus necessarily interdependent, so that the dimensions of all possible substitution matrices start from 2 and end at 20. Thus, it is uncertain what should be inferred from phylogenetic metrics for a single superfamily.

These difficulties are substantially offset by the consistency of site-by-site and row-by-row metrics with the construction and experimental characterization of aaRS protozymes [14] and urzymes [22-25, 27, 31, 37] (Figure 1). Such deep evolutionary intermediates are, at present, manifestly unique to the aaRS. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

extensive experimental and structural context of that consistency strengthens our conclusions even without comparable analysis of Class II aaRS, now in progress, especially in light of the following observations.

2.3.1 The three metrics are uncorrelated.

The CP1 MSA is a substantial outlier for all three metrics, suggesting that the metrics may be correlated. Removing the CP1 entries from the design matrix eliminates any correlation between Tree Height, Shape, or $\langle S \rangle$, Table 2. The three types of metrics are therefore essentially uncorrelated and provide independent insights.

Table 2. Cross-correlation R^2 values between phylogenetic metrics

	Tree Height	Shape
Shape	0.15	
<S>	0.10	0.03

2.3.2 Differences between urzyme and CP1 sequences are statistically meaningful.

If the Class I aaRS sequence partitions compared in Table 1 were all drawn from continuously replicated ancient genetic sources, subject to comparable selection history since their emergence, the null hypotheses would produce similarly conserved sequences and comparably congruent trees for the urzyme and CP1 partitions. The log-worth values (i.e., –log(P)) for CP1's higher mutational frequency (Tree Height; 2.7), its variance (Shape; 3.3) and lack of congruence for phylogenetic trees, ($\langle S \rangle$; 3.4), imply with high statistical significance that all metrics for CP1 arise from a different genetic population than the urzyme sequences, corroborating the argument—based on bidirectional coding ancestry—that the CP1 sequences represent genetic information acquired more recently by the urzymes.

2.3.3 The lengths of different segments drawn from the MSA have no detectable impact on any phylogenetic metric.

One might suppose that degraded congruence of the CP1 trees results from the fact its MSA has only ~0.25 as many amino acids as the urzyme MSA. However, including the NUMB parameter in Table 1 fails to reduce the variance of regression models for any score (Figures 2-5). The insignificant impact of NUMB and the clustering of the 20-residue $\langle S \rangle$ values with the intact urzyme MSA (Figure 5) confute that expectation.

2.3.4 Threading does not force any particular comparison between different aaRS types.

Threading increased the reliability of structure-based alignments within any aaRS type by adding sequences. Scaffold positions were rigorously defined as structural homologs from the close proximity of their alpha carbon coordinates in multiple PDB structure alignments (i) among aaRS for any single amino acid, drawn from very diverse bacterial species, that only then (ii) produced a grand scaffold MSA across all Class I aaRS types. Additions produced by threading thus have only second-order effects on structural superpositions of aaRS for different amino acids, adding precision to our analysis without overtly influencing choices on which our conclusions depend.

2.3.5 Distinguishing features of CP1 sequences are evident without considering indels.

Figures 2-5 together illustrate a comprehensive, near-optimally quantitative three-dimensional comparison of structural partitions in the Full Class I aaRS scaffold MSA. The crux of what the data suggest is that CP1 is more recent than the urzyme, its evolutionary divergence and variance—evidenced by its Tree Height and Shape—is much reduced, and its phylogenetic consistency—evidenced by $\langle S \rangle$—is also much reduced, relative to the corresponding metrics for urzyme segments. This counterintuitive conclusion is evident, from sequences with strict 1-1 correspondences between 3D crystal structures, excluding the large and variable-length indels that dominate CP1 insertions in most aaRS types. We consider this key observation in greater detail in the Discussion.

2.3.6 Neither convergent evolution nor horizontal gene transfer is a likely explanation for the urzyme/CP1 distinction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

The congruent clade structure of urzyme-derived sequences from the scaffold separates into consensus groupings characterizing Class I aaRS as a coherent superfamily the explanation of which does not require reference to mechanisms beyond mutation and selection from a single common ancestor. The aberrant behavior of sequence variations in the CP1 insertion (Figs. 2-5) might suggest appeal to such processes. Treatments of Class I aaRS evolution based on full MSAs [10, 48, 49] show evidence of horizontal gene transfer (HGT), genetic transpositions and large scale insertion/deletion events, of which CP1 is the foremost example.

Wherever full-length bacterial Class I enzymes with a particular amino acid substrate specificity are represented by more than one canonical structure, as has been described for IleRS and MetRS [see Fig. 3 in 48], that bipartite distribution in genome space is adequately explained in terms of early HGT into bacteria from an archaean/eukaryotic ancestor, but not in terms of convergent evolution. CP1 sequence disparities at homologous sequence positions in TyrRS, MetRS, and ValRS behave in the opposite manner: the higher variance of their $\langle Q \rangle$ values producing lower $\langle S \rangle$ values by allowing their evolutionary paths to wander widely, often crossing in sequence space, instead of forming multiple well-defined clades reasonably distant from one another in sequence space that could arise as a result of HGT.

2.4 Phylogenetic metrics have functional interpretations.

Regressions in Figure 3 were performed to demonstrate equivalence of the BEAST2 –log(Shape) and Clustal $\langle Q \rangle$ metrics. The dominant effect of the urzyme accounts for >60% of the variance in models for both dependent variables, with the balance coming from the opposing effects of the B and C subsets (whose β coefficients are of opposite sign). Regression models of the three independent metrics derived from BEAST2 tree constructions all depend heavily on significant two-way interactions between segments of the different MSAs. The Class I aaRS modules are experimentally sufficiently well-characterized to sustain functional interpretations of the three two-way interaction terms for site-to-site (Figures 2, 4) and row-by-row Support (Figure 5) metrics. These interpretations shed light on how evolutionary changes enhanced genetic coding.

2.4.1. CP1 forms a structural annulus constraining the urzyme's two halves (Figure 6A).

The simplest of the 10 CP1 insertions (in TrpRS) is only 74 residues long. Its structure is a ring that wraps around the protozyme and specificity-determining helix on one side of the active-site opening. Molecular dynamic simulations of the TrpRS urzyme [26] show that in these two segments, which together form the amino acid binding site in Class I aaRS, exhibit extensive relative motion. Further, the distance between the two parts of the amino acid binding site decreases in the TrpRS catalytic conformational transition and that motion is coupled to the relative motion of CP1 and the anticodon-binding domain. Steady state kinetic measurements of amino acid specificity confirm that this relative domain motion is coupled to the relative specificity for Tryptophan vs Tyrosine [31]. This structural feature provides a functional interpretation for how the three two-way interactions (Figures 3-5) help determine the phylogenetic metrics.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

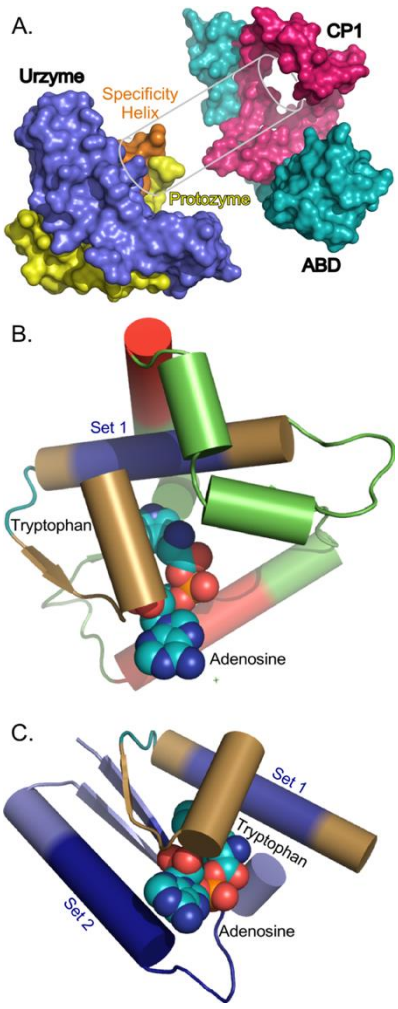


Figure 6. A. Structural relationships between the TrpRS urzyme, CP1, and the anticodon-binding domain (ABD). Coloring differs from that in Figures 1. The CP1 motif forms an annulus that constrains motions of the specificity determining helix (sand) and the Protozyme (yellow), constraining in turn the effective size of the amino acid binding pocket when the ABD (teal) changes its orientation. Experimental evidence [31] described in the text confirms that these constraints enable full-length TrpRS to reject tyrosine in the transition state complex for tryptophan activation. **B**. Structural cartoon with details of the interaction illustrated in **A**. Coloring is the same as that in Figure 1. The specificity-determining helix is the site of the Set 1 segment, which is colored in Blue. The TrpRS CP1 module is green and the scaffold segments are red. Note the close contact between CP1—especially the red scaffold segment—and Set 1. **C.** Interactions between the TrpRS protozyme—locus of the ATP binding site, and segment B, locus of the amino acid binding site. Subsets 2 and 1, which are contained entirely within the respective segments, are highlighted in dark blue.

2.4.2. The Tree Height dependence on segment B (Figure 2C) changes sign, depending on whether the protozyme is present.

As noted in regard to Figures 2A&B, the urzyme itself has a dominant impact increasing the value of the Tree Height parameter, and this effect is reduced substantially by the protozyme but increased by the protozyme*Segment B interaction. Yet, sequences in Set 1 from within Segment B have the highest Tree Height. Another reflection of the same phenomenon is that the protozyme and urzyme appear in very much the same place on the regression plots in Figures 2 and 5, yet are quite well separated in Figure 4, where the protozyme is midway between the urzyme and CP1.

The interaction in the Tree Height regression model is between the protozyme—locus of ATP bnding—and segment B—locus of amino acid binding (Figure 6B&C). The interaction term arises because of the contravariant effects of the two different binding sites. The protozyme ATP binding site is common to all Class I aaRS. Segment B is part of the amino acid binding site, and thus would be expected to exhibit the most significant evolutionary sequence variation between families specific for different amino acids.

2.4.3. The Shape dependence on segment B (Figure 4) changes sign, depending on whether CP1 is present.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

Thus, the interaction between CP1 and the amino acid binding site of the urzyme constrains the amino acid binding site dynamically. Structural relationships between CP1 and the amino acid binding site in TrpRS are highlighted in Figure 6B. The experimental demonstration of the interaction between CP1 and the amino acid binding site validates the sign and strength of the B*CP1 contribution to Shape much as a pre-formed space in a partially assembled puzzle validates the outline of the missing piece.

2.4.4. The dependence of Support on CP1 (Figure 5B) changes sign, depending on whether the protozyme is present.

Residues within the protozyme contribute even more decisively to the average value of the Support parameter than do residues located elsewhere in the urzyme. CP1 has the most significant impact on the regression model for the Support metric. Its coefficient is -0.33, more than three times that of the next most important predictor. This effect can be seen in the regression plot in Figure 5, in which the MSA for CP1 is more widely separated from the other MSAs than in the models for any other metric. However, the presence of the protozyme sequences in the Full MSA is sufficient to change the impact of CP1 from negative to positive, giving the β-coefficient of +0.1 for the protozyme*CP1 interaction term. This functional interpretation and the widespread occurrence of the protozyme packing motif [29] and its functional activation of ATP [14], reinforce the conclusion that the protozyme was the original root of the entire superfamily and is older than the urzyme itself, as previously proposed [50].

3. Discussion

The progressive biochemical functionality of aaRS protozymes and urzymes; the fact that their structures are universally conserved within both Classes; and the evidence that they descended from one bidirectional ancestral gene all imply that they are legitimate experimental models for ancestral evolutionary aaRS forms that participated throughout the emergence of genetic coding and well before LUCA [24]. Moreover, to date, no such sequential intermediates have been derived for other superfamilies. Thus, more appears to be known about the modular evolution of the two aaRS superfamilies than for any other ancient superfamily, making the Class I aaRS an appropriate subject for this work.

The biological importance of our results is that they furnish new phylogenetic support for an evolutionary trajectory assembling different polypeptide sequences with successive capabilities necessary for the emergence and refinement of genetic coding (Figure 7):

- Mobilization of ATP as an energy source for amino acid activation (protozyme)
- Simultaneous recognition of amino acid and tRNA substrates and a rudimentary binary code (second half of the urzyme)
- Insertion of an ancestral CP1, perhaps from an RNA transposable element, to produce a rudimentary catalytic domain and terminate bidirectional coding (CP1)
- Assimilation of idiosyncratic anticodon binding domains (ABD)
- Expansion of the coding alphabet via mutational generation of allosteric coupling between CP1 and ABD modules and multistage error correction (editing domains).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

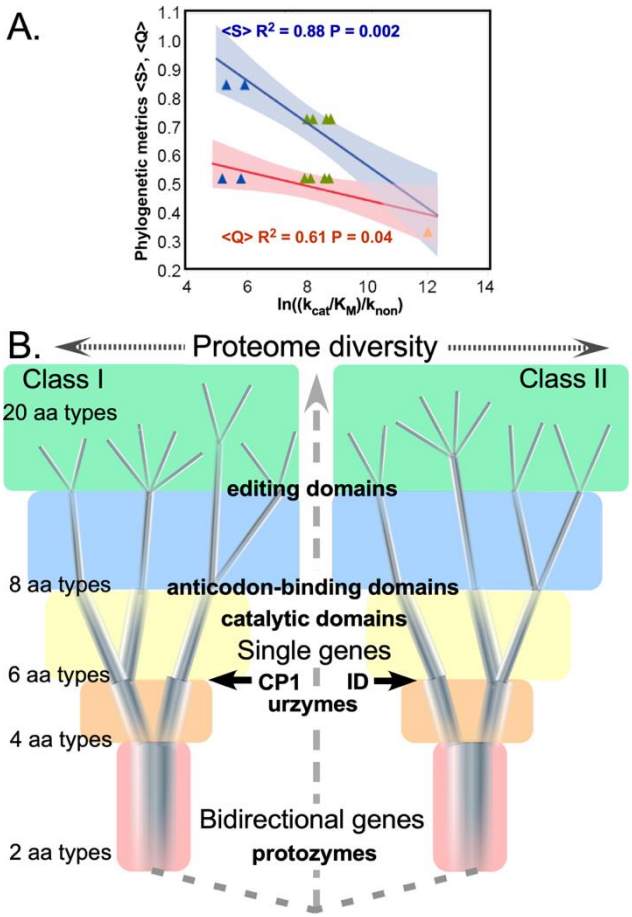


Figure 7. Assignment catalysis and code evolution. **A.** Correlations between rate accelerations and phylogenetic metrics, $\langle Q \rangle$ and $\langle S \rangle$. Parameters for regression against experimental catalytic proficiencies for corresponding putative evolutionary intermediate constructs are both significant. Blue, green, and amber triangles represent protozymes, urzymes, and catalytic domains. **B.** Timeline for growth of the genetic coding alphabet from a two-letter code. Introducing new aaRS into the context of the ancestral bidirectional gene (red background and dashed connecting lines) simultaneously enhanced specificity and created fundamental changes in selection pressure. Different colored backgrounds signify altered selection pressures that apply to all extant aaRS at a given stage of the coding alphabet as well as the scale of the extant proteome possible with successive alphabet sizes. Increasing cardinality of the alphabet induces (i) sequence space inflation as a greater number of distinguishable sequences can be specified; and (ii) restriction in the diameters of the quasispecies as they approach fully coded sequences in the final 20-letter alphabet. Bold face landmarks (CP1 and ID) denote qualitative changes in aaRS architecture shown experimentally to enrich specificity as discussed in the text. Note that this timeline represents evolutionary events before the LUCA.

3.1. *Primordial aaRS quasispecies covered progressively smaller regions of sequence space, closer to the "root".*

Early aaRS evolution cannot have been an ordinary mutation/selection process. All solutions to the problem normally framed as a "chicken and egg problem" [51] imply historical context and a continuity principle must be defined by the genotype-phenotype mapping [52]: at any stage during the emergence of coding some prior system must have been interpreted by extant aaRS protogenes. A key aspect of such prior systems is that as the alphabet size, diversity, and consequent fidelity increased, they would, *ipso facto*, have created more narrowly targeted selection pressures, strongly coupling mutation and selection in early stages of genetic coding (Figure 7B). The subsequent coevolution of CP1 and urzyme sequences would necessarily have preserved urzyme functionality, while the new CP1 could adapt flexibly to its developing role of enhancing specificity as described in the next section.

Different background colors in Figure 7B denote how branching of the tree to allow introduction of the n^{th} amino acid into the alphabet enforces a highly cooperative re-optimizing of the n–1 aaRS types already present. As each new, refined amino acid type emerged, all extant gene sequences adapted to opportunities introduced by progressively finer discrimination between amino acid side chain physical chemistry. In turn, adaptation to a more diverse alphabet sharpens the new aaRS fidelities (i.e., the contraction of branch

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

thicknesses in Figure 7B), implicating a bootstrapping feedback and enhancing the cooperativity of the transition to higher-dimension coding alphabets [50]. That cooperativity creates a Lamarckian-like correlation between selection pressure and its outcome—the result of any mutation being nearly synonymous with the selection pressure it faced, especially as the code differentiated. Survival would have depended on the relationship between the shapes of fitness landscapes and error rates of catalysis by the extant quasispecies [53]. The earliest alphabets, including at least the first two amino acid types, were also tightly constrained by bidirectional coding (rose-shaded background, Figure 7B).

We have argued [50, 54] that a single functional island in sequence space (i.e. quasispecies) would invariably have been a strong attractor, irrespective of detailed features of the fitness landscape that stabilized it, because all mutations that moved the system slightly off its optimum phenotype would be subject to strongly restorative selection pressures. However, if ancestral protozymes from a bidirectional gene had broad, relatively flat (and necessarily co-dependent) fitness landscapes, matched to correspondingly high error rates, that could have favored bifurcated quasispecies that enhanced genetic coding by recruiting new amino acid types and simultaneously increasing the precision with which child specificities could be encoded.

3.2. *Evolutionary refinements of protein catalysis and specificity were predicated on expanding the genetic code.*

The Tree Height, Shape, and Support metrics identify differences in the genetic origins of successively acquired contributors to aaRS function: protozyme=>urzyme=>catalytic domain with CP1. They correlate inversely with experimentally measured catalytic proficiencies for the catalysts from which the corresponding sequences are derived (Figure 7A). Those proficiencies themselves correlate quantitatively with parallel increases in catalytic proficiency with the concomitant additions of mass in the Class I and II aaRS evolutionary intermediates they entail [See Figure 6 in 24]. Both phylogenetic and functional properties of these intermediates thus appear to probe far deeper into evolutionary time than previous phylogenetic analyses of Class I aaRS [7, 11, 18, 19, 47].

The precision with which protein active sites distinguish substrates from one another, and transition states from substrates, was the result of the evolutionary process we wish to characterize. Expansion of genetic coding itself depended critically on developing a system for the placement of a precisely defined amino acid sidechain at a particular point on an aaRS peptide backbone. Phylogenetic analysis of segmental Class I aaRS MSAs represents a uniquely promising opportunity to test the hypothesis that contemporary enzymes are mosaic structures rooted in simpler catalytic polypeptides and assembled from detectably different genetic ancestors.

3.3. *Phylogenetic metrics identify meaningful fine structure and covariation within Multiple Sequences Alignments.*

Sections 2.2-2.5 exploit quantitative metrics compiled during the Markov Chain Monte Carlo exploration of the phylogenetic landscape to expose differences in how different segments of the overall MSA behave. The statistical coherence of these phylogenetic metrics alone justifies their novel application here. Their functional significance emerges only in the context of partitioning the overall MSA according to structural and functional criteria established within other disciplines and in the presence of suitable controls for the effect of sequence length (subsets 0-6 from the urzyme MSA; Figure 1C, Table 1).

In turn, the phylogenetic analysis provides novel validation of decisions that guided experimental work on partitioning the MSA. Identification of how the Tree Height (Figure 2B), Shape (Figure 4B), and Support (Figure 5B) depend on two-way intermodular interactions validates experimental work demonstrating how intermodular coupling contributes to catalysis and specificity. Further, because these interactions arise from the coherence across the Class I superfamily, they imply that similar interactions occur between ATP and amino acid binding sites and between amino acid binding sites and CP1 in all or most Class I aaRS.

The resulting Bayesian phylogenetic evidence of fine structure within protein domains has no precedent. These observations suggest that this work points toward more substantial applications of software like BEAST2 [55] to a broader range of evolutionary questions.

3.4. *High-resolution structural modularity implies discontinuities in the evolution of genetic coding.*

If CP1 insertions were indeed assimilated from one or more similar genetic sources after Class I aaRS urzymes had evolved significantly from ancestral protozymes, it could have at least three noteworthy biological implications:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

(i) Class I protozymes, whose high catalytic activity mobilizes ATP for biosynthesis, an activity found in many proteins—may be the root of that major portion of the proteome built from β-α-β crossover connections—the Rossmannoid protein superfamily [29] and potentially β-barrel proteins [56, 57], which are central to intermediary metabolism and nucleotide biosynthesis [58].

(ii) AARS protozyme and urzyme populations would have functioned first as quasispecies in translation, limiting the sophistication of the early proteome.

(iii) CP1 assimilations would have transformed selection pressures for subsequent aaRS evolution by facilitating enhanced fidelity.

3.5. Insertion of CP1 likely enabled saltatory improvements in fidelity.

Structural and biochemical data suggest that the CP1 insertions created stepwise enhancements in the evolution of genetic coding by enabling conformationally-driven mechanisms to increase specificity. The shortest CP1 insertions have ~75 residues in TrpRS and TyrRS that recur essentially intact in the longer CP1 insertions of the remaining eight Class I aaRS [17, 42]. It seems likely that the initial insertion needed to be that long. CP1 must wrap around the urzyme (Figure 6) to constrain relative movements of the protozyme and specificity helix that form the amino acid binding site [26]. For that reason, an earlier hypothesis as to its origin referenced near-simultaneous insertion of a mobile genetic element into all extant Class I urzymes [17], in which case its root sequence would have been more recent than that of the urzymes, yet earlier than the remaining sequences in contemporary full-length aaRS enzymes.

Amino acid specificities [22, 27] suggest that although capable of 10^9-fold rate accelerations, Class I and II urzymes select an amino acid from the correct Class only 80% of the time. However, Wills & Carter [28] note that within-Class aaRS urzyme specificity is consistent with each Class distinguishing two kinds of amino acids, to operate a four-letter coding alphabet. These modest fidelities suggest a fundamental limit to the precision of which bidirectional coding was ultimately capable.

The most evident contribution of CP1 to fidelity is that the editing domains present in the larger subclass IA aaRS for aliphatic amino acids Ile, Val, and Leu are elaborations of the CP1 motif present in the simplest subclass IC aaRS for Tyr and Trp. It seems likely, however, that CP1 functioned even earlier to enhance fidelity by dynamically constraining the volume and configuration of the amino acid binding pocket. Several groups found that comprehensive mutation of side chains in the immediate vicinity of the amino acid substrate, all within the urzyme architecture, would not change specificities of subclass IB GlnRS to Glu [59, 60] or subclass IC TrpRS to Tyr [61]. Changing GlnRS specificity to Glutamate (Bullock, et al. [60]) required wholesale mutations in the second layer surrounding the amino acid binding site outside the urzyme, but within in the GlnRS CP1 domain.

Similarly, a modular thermodynamic cycle comparing specificities of full-length TrpRS, urzyme, and urzyme plus either CP1 or the anticodon binding domain (ABD) showed rejection of Tyrosine by *G. stearothermophilus* TrpRS requires cooperation between CP1 and the ABD [31]. CP1 must coordinate its movements with those of the ABD to perform the allosteric communication necessary to enhance side chain selectivity beyond the modest capabilities of the urzyme [23, 33]. In both cases, fine tuning specific recognition of amino acid substrates required insertion of CP1 and the ABD and, subsequently, coupling between them.

Inserting the CP1 motif would necessarily have disrupted bidirectional coding (Fig 1B), thus dividing aaRS evolutionary histories decisively into distinct stages (Figure 7, between orange and yellow backgrounds). Selective advantages of CP1 insertion thus appear to have been to (i) end constraints imposed by bidirectional coding and (ii) transcend the fundamental limitation on specificity posed by the urzyme architecture. Either or both would have allowed substantial, discontinuous, increases in fidelity to develop. CP1 therefore likely dramatically transformed the Class I aaRS fitness landscape, and was likely necessary to expand the coding alphabet.

3.6. A revised branching order suggests that Class I aaRS protozymes and urzymes root the Rossmannoid superfamily.

Putting Class I aaRS protozymes at the root of that superfamily reconfigures many branching orders within the proteome to reflect that aaRS urzymes were not a late-developing branch in the Rossmannoid superfamily radiation, but instead were ancestral to it (Figure 8). Descent of Class I and II aaRS from a single bidirectional ancestral gene [14, 21, 22, 24, 27, 38, 39] would underscore the likelihood that the aaRS of both families diverged, rather than converging to similar functions from different sources. Thus, the genetic coding table itself likely evolved by bifurcating pre-existing aaRS genes into specialized enzymes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

whose more discriminating specificities for tRNA and amino acid substrates enabled daughter synthetases to differentiate groups amino acids that previously had functioned as synonymous [50, 54, 62, 63]. It would be surprising if other branches of the proteome had not diverged from the ancestral aaRS, as suggested in Figure 6.

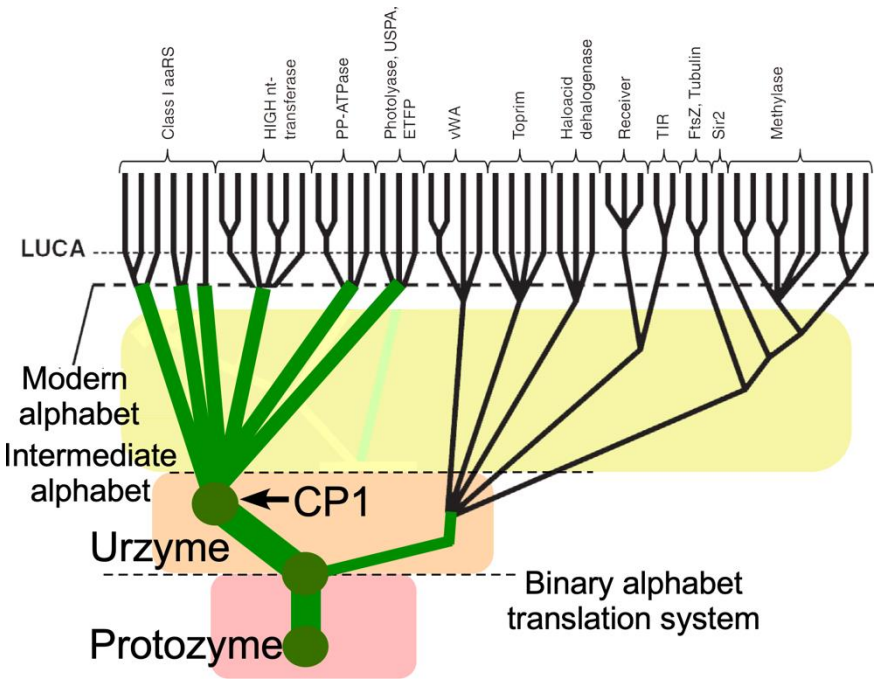


Figure 8. Modified Class I aaRS phylogeny consistent with bidirectional coding ancestry in a peptide/RNA world. Adapted with permission from Fig 12-2A of [64]. Green circles represent key stages in the emergence of genetic coding, beginning with ATP-dependent amino acid activation by protozymes coded by a bidirectional gene [14]. Backgrounds are colored as in Figure 7B to approximate the expansion of the proteome possible with suggestive increases in the dimension of the alphabet.

4. Materials and Methods

Amino acid chemistry underlies the metric form of amino acid substitution matrices (aaSM) generally required for the logically circular process of creating an amino acid sequence alignment by optimizing the constraints provided by some aaSM and then building a phylogenetic tree from the resulting alignment according to those same constraints [12].

We superposed the available bacterial Class I aaRS crystal structures to base the multiple sequence alignment (MSA) exclusively on precise three-dimensional structural homology, freeing the MSA from dependence on empirical substitution matrices and estimates of relative rates of mutation [65]. In the first place, this was done within each aaRS type. The identification of sequence positions displaying very high conservation of both structure and amino acid occupancy provided unambiguous anchors for the production of much more extensive sequence alignments. Popinga [66] performed this task through extensive use of HHpred [67] and MODELLER [68] to thread amino acid sequences for each aaRS type using experimentally determined 3D structures as templates. This procedure produced, for each aaRS type, an expanded MSA in which various well-conserved secondary structural regions were near-perfectly structure- and sequence-aligned between species known to have diverged not long after the LUCA.

In the final stage of alignment, the structures of the individual aaRS types were brought together to identify regions of universal structural homology, along lines similar to those described by Chaliotis [69], to roduce what we refer to as the Class I aaRS “scaffold”. The product was an MSA across all Class I aaRS types in which each sequence position could be assumed to have arisen, as far as is reasonably possible, from the same LUCA-ancestral codon. While the validity of this assumption is by its very nature untestable, the close structural homology of Class I aaRS of all types and the rigor of our procedure gives us confidence and justification as good as that underlying any cross-species phylogenetic enterprise. However, we took further steps to restrict and constrain the data upon which we later built putative phylogenies. First, individual scaffold elements are more extensive in some extant aaRS types than others and it is not possible to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

identify proper residue-by-residue homology among the loops, turns and structurally disordered regions that join them. All these were excluded from the Class I "scaffold" MSA. Second, the earliest domain of life is unknown and highly controversial, without consensus [70-72]. However, we included only eubacterial aaRS for the following reasons.

The substantially stronger codon middle-base pairing frequency and the steeper slope between independently reconstructed ancestral eubacterial Class I and II sequences ([38]; Chandrasekaran, unpublished data; Carter & Wills unpublished data) provide evidence that bacterial aaRS sequences are both closer to the ancestral root and less convoluted by horizontal gene transfer than those from other domains. Analysis of loop structures is a contested issue; however there is certainly no consensus that they contain reliable information about the evolutionary origin of biology's three main evolutionary domains [73-77]

It is generally accepted that cytoplasmic proteins of bacteria have been subject to many fewer complex selection pressures than their archeal and eukaryotic counterparts. The complex functional and regulatory roles played by some aaRS proteins and their involvement in numerous genetic syndromes attests to this [78].

Thus, the final MSA contained roughly bacterial 20 sequences for each of 10 Class I aaRS, circumventing as many problems as possible in aligning sequences that diverged to produce different substrate specificities in the pre-LUCA æon. Use of scaffold sequences for phylogenetic analysis gave the best guarantee that results would reflect relatively neutral evolution within the context of asymmetric, specialized selection pressures producing variant substrate specificities by fine-tuning the size, shape and chemistry of more intricately constructed amino acid side-chain structures and pockets.

The MSA for the Class I scaffold was output using VMD [79], and is provided in the supplement (Supplement_MSA_files.txt) together with all subsets used in this work. The scaffold fasta file was then partitioned along lines of the experimental deconstruction of the Class I aaRS superfamily [24] into the disjointed segments of the urzyme, separated by structurally conserved segments from CP1. Finally, because the urzyme (83) and CP1 (20) partitions of the scaffold MSAs have different sequence lengths, seven subsets comprising 20 sequence positions distributed throughout the urzyme were selected arbitrarily by a balanced, randomized procedure [80] to test the effect of sequence length on our phylogenetic signatures. We invoked the usual "zeroth order"assumption that the evolution of any sequence position was statistically independent of all other positions.

We compiled three complementary metrics to compare the overall scaffold MSA to its Class I urzyme, protozyme, and CP1 subsets. BEAST2 accumulates two column-by-column metrics from the ensemble of trees generated by Markov Chain Monte Carlo simulation. The Tree Height reflects the mean overall mutation rates. Site-specific mutation rates are fitted to a gamma distribution [81] by adjusting a Shape parameter, $\alpha = 1/CV^2$, where CV is the coefficient of variation or the ratio of the standard deviation to the mean. Shape is thus a measure of Tree Height variance.

The conservation quality, Q_j, defined by Clustal [43, 44] was computed down each column, j, of the grand MSA, which included all aaRS types. While the definition of Q seems convoluted, it has been constructed as a metric whose value reflects the degree of amino acid diversity generated by typical evolutionary amino acid substitution processes, reflected in the substitution matrix that it uses as a reference. Different matrices do not give widely divergent Q parameters. We used the ClustalW default matrix (PAM 250; [82]) and calculated the average, $\langle Q \rangle$, over all positions within a partition of the MSA a parametric representation of the partition-wide variation in amino acid occupancy calculated column-by-column over individual sequence positions. In the course of the analysis it emerged that $\langle Q \rangle$ is actually nearly co-linear with the log(Shape) parameter (Figure 3).

A third, row-by-row metric was derived from molecular kinships between rows of the MSA within each partition by clustering the sequences into clades according to their evolutionary origin. Phylogenetic trees were computed using BEAST2 [55], allowing the use of multiple amino acid substitution matrices (primarily WAG and LG). Trees were visualized with DensiTree [83] and FigTree [84]. For the full Class I scaffold and each partition of interest, DensiTree was used to extract ~10,000 trees generated by BEAST2 in building trees. From these we calculated values of a parameter representing the support for clades that each corresponded exclusively to an individual aaRS type, i. This support, S_i, is the percentage of all trees for which the aaRS type in question appears to be monophyletic, meaning that the leaves of that aaRS type descend from the same most recent branch point (common ancestor) with no descendants arising from a different

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

aaRS type. The mean phylogenetic support, $\langle S \rangle$, is a metric derived from a row-by-row calculation over all 10,000 phylogenetic trees generated using the sequence data for a chosen partition of the scaffold MSA. Cross correlation between the three metrics is negligible (Table 2), so they reflect independent aspects of the MSAs.

Columns for various predictors were appended to the table of $\langle Q \rangle$ and $\langle S \rangle$ values to form the design matrix (Table 2) used for multiple regression analysis with the JMP software [85]. Multiple Regression was used to assess the statistical strength of contributions of the predictors (i.e., the independent variables, A, B, C, CP1 defined in Fig 1 and the number of amino acids) to $\langle Q \rangle$ and $\langle S \rangle$ values, using stepwise searches to identify the best set of predictors followed by least squares estimation of the corresponding coefficients and their Student t–test P values.

5. Conclusions

Understanding how evolution of aaRS•tRNA cognate pairs effected the stepwise increases in dimensionality of entries into the universal genetic coding table is pivotal to the origin of biology. To that end we have sought relevant data from carefully curated, structure-based amino acid sequence alignments (MSAs) of secondary structures shared by all members of the Class I aaRS superfamily. Three uncorrelated phylogenetic metrics identify significant, high-resolution mosaicity, consistent with assembly from distinct genetic sources. All metrics reinforce prior arguments that Class I aaRS evolved by a succession of intermediate states—protozyme=>urzyme=>Catalytic domain—with increasingly sophisticated catalytic [14, 22, 24, 27, 37] and discriminatory [50, 54, 86] capabilities. Regression analyses identify covariation of sequences as statistically significant two-way intermodular interactions that facilitate functional interpretations and help validate our approach.

Genetic coding and the proteome probably emerged from an RNA•polypeptide partnership. Previously accepted phylogenies of the Class I aaRS root well before the Last Universal Common Ancestor (LUCA) [7, 11, 18, 19, 47], but appeared also to have radiated after earlier bifurcations in the immense metafamily [7] containing folds based on the Rossmann dinucleotide-binding fold [87]. Genetic coding by protein aaRS was therefore thought, necessarily to have replaced a prior implementation based on ribozymal assignment catalysts [64, 88]. This paradoxically late radiation of protein assignment catalysts has been the foremost phylogenetic evidence favoring the RNA World hypothesis.

Evidence described here that CP1 is a more recent acquisition by ancestral aaRS urzymes supports a plausible alternative branching order (Figure 8). Attributing the apparently late adaptive radiation of Class I aaRS to CP1 achieves consistency with an origin of genetic coding from a bidirectional gene administering a binary coding alphabet in a peptide/RNA world. Resolving the paradox in this way complements—without necessarily contradicting—the phylogenetic analyses of Koonin [64, 88].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

Supplementary Materials: The following are available online at www.mdpi.com/Distinct_Origins_Supplement_MSA_files.txt.

Author Contributions: Conceptualization, P.R.W, R.B. and C.W.C., Jr.; methodology, P.R.W, A.P., and R.B.; software, R.B.; validation, P.R.W., R.B., and C.W.C., Jr.; formal analysis, C.W.C., Jr.; curation, A.P and P.R.W.; writing—original draft preparation, P.R.W. and C.W.C., Jr; writing—review and editing, P.R.W., R.B., and C.W.C., Jr; funding acquisition, C.W.C., Jr. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institute of General Medical Sciences (Grant number R01-78227 to C.W.C., Jr). Open Access charges were supported by The Alfred P. Sloan Foundation (Grant number G-2021-16944 to C.W.C., Jr.)

Acknowledgments: We are grateful to Corbin Jones, Greg Fournier, Günter Wachtershäuser, and Bill Martin for helpful feedback in preparing the manuscript and Loren Williams for suggesting that ribosomal protein sequences may exhibit a similar genetic mosaicity.

Conflicts of Interest: The authors know of no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

6. References

1. Eriani, G.; Delarue, M.; Poch, O.; Gangloff, J.; Moras, D., Partition of tRNA Synthetases into Two Classes Based on Mutually Exclusive Sets of Sequence Motifs. *Nature* **1990,** 347, 203-206.
2. Cusack, S.; Berthet-Colominas, C.; Härtlein, M.; Nassar, N.; Leberman, R., A second class of synthetase structure revealed by X-ray analysis of *Escherichia coli* seryl-tRNA synthetase at 2.5 Å. *Nature* **1990,** 347, (6290), 249-255.
3. Ruff, M.; Krishnaswamy, S.; Boeglin, M.; Poterszman, A.; Mitschler, A.; Podjarny, A.; Rees, B.; Thierry, J. C.; Moras, D., Class II Aminoacyl Transfer RNA Synthetases: Crystal Structure of Yeast Aspartyl-tRNA Synthetase Complexed with $tRNA^{Asp}$. *Science* **1991,** 252, (6), 1682-1689.
4. Carter, C. W., Jr., Cognition, Mechanism, and Evolutionary Relationships in Aminoacyl-tRNA Synthetases. *Ann. Rev. Biochem.* **1993,** 62, 715-748.
5. Schimmel, P.; Giegé, R.; Moras, D.; Yokoyama, S., An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Nat. Acad. Sci. USA* **1993,** 90, 8763-8768.
6. O'Donoghue, P.; Luthey-Schulten, Z., On the Evolution of Structure in Aminoacyl-tRNA Synthetases. *Microbiol. Mol. Biol. Rev.* **2003,** 67, (4), 550–573.
7. Aravind, L.; Anantharaman, V.; Koonin, E. V., Monophyly of Class I Aminoacyl tRNASynthetase, USPA, ETFP, Photolyase, and PP-ATPase Nucleotide-Binding Domains: Implication for Protein Evolution in the RNAWorld. *PROTEINS: Struct. Funct. Gen.* **2002,** 48, 1–14.
8. Leipe, D. D.; Wolf, Y. I.; Koonin, E. V.; Aravind, L., Classification and Evolution of P-loop GTPases and Related ATPases. *J. Mol. Biol.* **2002,** 317, 41-72.
9. Roach, J. M.; Sharma, S.; Kapustina, M.; Carter, C. W., Jr., Structure alignment via Delaunay tetrahedralization. *PROTEINS: Struct. Funct. Bioinf.* **2005,** 60, (1), 66-81.
10. Wolf, Y. I.; Aravind, L.; Grishin, N. V.; Koonin, E. V., Evolution of Aminoacyl-tRNA Synthetases—Analysis of Unique Domain Architectures and Phylogenetic Trees Reveals a Complex History of Horizontal Gene Transfer Events. *Genome Research* **1999,** 9, 689–710.
11. Caetano-Anollés, G.; Wang, M.; Caetano-Anollés, D., Structural Phylogenomics Retrodicts the Origin of the Genetic Code and Uncovers the Evolutionary Impact of Protein Flexibility. *Plos One* **2013,** 8, (8), e72225.
12. Shore, J.; Holland, B. R.; Sumner, J. G.; Nieselt, K.; Wills, P. R., The Ancient Operational Code is Embedded in the Amino Acid Substitution Matrix and aaRS Phylogenies. *J. Mol. Evol.* **2019,** 88, 136–150.
13. Cusack, S., Eleven down and nine to go. *Nat. Str. Biol.* **1995,** 2, 824-831.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

14. Martinez, L.; Jimenez-Rodriguez, M.; Gonzalez-Rivera, K.; Williams, T.; Li, L.; Weinreb, V.; Chandrasekaran, S. N.; Collier, M.; Ambroggio, X.; Kuhlman, B.; Erdogan, O.; Carter, C. W. J., Functional Class I and II Amino Acid Activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. *J. Biol. Chem.* **2015,** 290, (32), 19710–19725.
15. Hobson, J. J.; Li, Z.; Carter, C. W., Jr., A leucyl-tRNA synthetase urzyme: authenticity of tRNA Synthetase urzyme catalytic activities and production of a non-canonical product. *Nucl. Acids Res.* **2021,** submitted.
16. Li, L.; Weinreb, V.; Francklyn, C.; Carter, C. W., Jr, Histidyl-tRNA Synthetase Urzymes: Class I and II Aminoacyl-tRNA Synthetase Urzymes have Comparable Catalytic Activities for Cognate Amino Acid Activation. *J. Biol. Chem.* **2011,** 286, 10387-10395.
17. Pham, Y.; Li, L.; Kim, A.; Erdogan, O.; Weinreb, V.; Butterfoss, G.; Kuhlman, B.; Carter, C. W., Jr, A Minimal TrpRS Catalytic Domain Supports Sense/Antisense Ancestry of Class I and II Aminoacyl-tRNA Synthetases. *Mol Cell* **2007,** 25, 851-862.
18. Fournier, G. P.; Alm, E. J., Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. *J. Mol. Evol.* **2015,** 80, 171-185.
19. Fournier, G. P.; Andam, C. P.; Alm, E. J.; Gogarten, J. P., Molecular Evolution of Aminoacyl tRNA Synthetase Proteins in the Early History of Life. *Orig Life Evol Biosph* **2011,** 41 621–632
20. Pymol *The PyMOL Molecular Graphics System*, Schrödinger, LLC.
21. Rodin, S. N.; Ohno, S., Two Types of Aminoacyl-tRNA Synthetases Could be Originally Encoded by Complementary Strands of the Same Nucleic Acid. *Orig. Life Evol. Biosph.* **1995,** 25, 565-589.
22. Carter, C. W., Jr.; Li, L.; Weinreb, V.; Collier, M.; Gonzales-Rivera, K.; Jimenez-Rodriguez, M.; Erdogan, O.; Chandrasekharan, S. N., The Rodin-Ohno Hypothesis That Two Enzyme Superfamilies Descended from One Ancestral Gene: An Unlikely Scenario for the Origins of Translation That Will Not Be Dismissed. *Biology Direct* **2014,** 9, 11.
23. Li, L.; Carter, C. W., Jr, Full Implementation of the Genetic Code by Tryptophanyl-tRNA Synthetase Requires Intermodular Coupling. *J. Biol. Chem.* **2013,** 288, (29 November), 34736–34745.
24. Carter, C. W., Jr., Coding of Class I and II aminoacyl-tRNA synthetases. *Advances in Experimental Medicine and Biology: Protein Reviews* **2017,** 18, 103-148.
25. Li, L.; Francklyn, C.; Carter, C. W., Jr, Aminoacylating Urzymes Challenge the RNA World Hypothesis. *J. Biol. Chem.* **2013,** 288, 26856-26863.
26. Pham, Y.; Kuhlman, B.; Butterfoss, G. L.; Hu, H.; Weinreb, V.; Carter, C. W., Jr, Tryptophanyl-tRNA synthetase Urzyme: a model to recapitulate molecular evolution and investigate intramolecular complementation. *J. Biol. Chem.* **2010,** 285, 38590-38601.
27. Carter, C. W., Jr, What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life* **2015,** 5, 294-320.
28. Wills, P. R.; Carter, C. W., Jr., Impedance matching and the choice between alternative pathways for the origin of genetic coding. *International Journal of Molecular Sciences* **2020,** 21, 7392.
29. Cammer, S.; Carter, C. W., Jr., Six Rossmannoid Folds, Including the Class I Aminoacyl-tRNA Synthetases, Share a Partial Core with the Anticodon-Binding Domain of a Class II Aminoacyl-tRNA Synthetase. *Bioinformatics* **2010,** 26, (6), 709-714.
30. Carter, C. W., Jr., Escapement mechanisms: efficient free energy transduction by reciprocally-coupled gating. *Proteins: Structure, Function, and Bioinformatics* **2019,** 88, 710–717.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

31. Weinreb, V.; Li, L.; Chandrasekaran, S. N.; Koehl, P.; Delarue, M.; Carter, C. W., Jr Enhanced Amino Acid Selection in Fully-Evolved Tryptophanyl-tRNA Synthetase, Relative to its Urzyme, Requires Domain Movement Sensed by the D1 Switch, a Remote, Dynamic Packing Motif *J Biol Chem* **2014,** 289, 4367-4376.
32. Weinreb, V.; Li, L.; Carter, C. W., Jr., A Master Switch Couples Mg^{2+}-Assisted Catalysis to Domain Motion in B. stearothermophilus Tryptophanyl-tRNA Synthetase. *Structure* **2012,** 20, 128-138.
33. Chandrasekaran, S. N.; Carter, C. W., Jr., Adding torsional interaction terms to the Anisotropic Network Model improves the PATH performance, enabling detailed comparison with experimental rate data *Structural Dynamics* **2017,** 4, 032103.
34. Chandrasekaran, S. N.; Das, J.; Dokholyan, N. V.; Carter, C. W., Jr., A modified PATH algorithm rapidly generates transition states comparable to those found by other well established algorithms. *Structural Dynamics* **2016,** 3, 012101.
35. Carter, J., Charles W.; Chandrasekaran, S. N.; Weinreb, V.; Li, L.; Williams, T. In *Combining multi-mutant and modular thermodynamic cycles to measure energetic coupling networks in enzyme catalysis* Structural Dynamics, American Crystallographic Association Annual Meeting, 2016; Pearson, A.; Benedict, J., Eds. American Crystallographic Association: American Crystallographic Association Annual Meeting, 2016.
36. Onodera, K.; Suganuma, N.; Takano, H.; Sugita, Y.; Shoji, T.; Minobe, A.; Yamaki, N.; Otsuka, R.; Mutsuro-Aoki, H.; Umehara, T.; Tamura, K., Amino acid activation analysis of primitive aminoacyl-tRNA synthetases encoded by both strands of a single gene using the malachite green assay. *BioSystems* **2021,** 208, (October), 104481.
37. Carter, C. W., Jr., Urzymology: Experimental Access to a Key Transition in the Appearance of Enzymes. *J. Biol. Chem.* **2014,** 289, (44), 30213–30220.
38. Chandrasekaran, S. N.; Yardimci, G.; Erdogan, O.; Roach, J. M.; Carter, C. W., Jr, Statistical Evaluation of the Rodin-Ohno Hypothesis: Sense/Antisense Coding of Ancestral Class I and II Aminoacyl-tRNA Synthetases. *Molecular Biology and Evolution* **2013,** 30, (7), 1588-1604.
39. Rodin, A.; Rodin, S. N.; Carter, C. W., Jr., On Primordial Sense-Antisense Coding. *Journal of Molecular Evolution* **2009,** 69, 555-567.
40. Burbaum, J. J.; Schimmel, P., Assembly of a Class I tRNA Synthetase from Products of an Artificially Split Gene. *Biochem.* **1991,** 30, 319-324.
41. Burbaum, J. J.; Starzyk, R. M.; Schimmel, P., Understanding Structural Relationships in Proteins of Unsolved Three-Dimensional Structure. *PROTEINS: Struct. Funct. Gen.* **1990,** 7, 99-111.
42. Wills, P. R., Genetic Information, Physical Interpreters and Thermodynamics; The Material-Informatic Basis of Biosemiosis *Biosemiotics* **2014,** 7, 141–165.
43. Thompson, J. D.; Higgins, D. G.; Gibson, T. J., Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994,** 22, 4673-4680.
44. Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G., The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **1997,** 24, 4876-4882.
45. Whelan, S.; Goldman, N., A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution* **2001,** 18, (5), 691-699.
46. Le, S. Q.; Gascuel, O., An Improved, General Amino-Acid Replacement Matrix. *Molecular Biology and Evolution. 2008* **2008,** 25, ((7)), 1307-20.
47. Katsnelson, M. I.; Wolf, Y. I.; Koonin, E. V., Towards physical principles of biological evolution. *arXiv* **2018,** orgabs1709.00284.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

48. Shiba, K., Glycyl-tRNA Synthetases. In *The Aminoacyl-tRNA Synthetases* Ibba, M.; Francklyn, C.; Cusack, S., Eds. Landes Bioscience: Austin, 2005.
49. Soucy, S. M.; Huang, J.; Gogarten, J. P., Horizontal gene transfer: building the web of life. *Nat. Rev. Gen.* **2015,** 16, (AUGUST), 472
50. Carter, C. W., Jr; Wills, P. R., Interdependence, Reflexivity, Fidelity, and Impedance Matching, and the Evolution of Genetic Coding. *Molecular Biology and Evolution* **2018,** 35, (2), 269-286.
51. Tyson, N. d. G., "Just to settle it once and for all: Which came first the Chicken or the Egg? The Egg – laid by a bird that was not a Chicken. *2013-01-28* **2013**, Tweet.
52. Fontana, W.; Schuster, P., Continuity in Evolution: On the Nature of Transitions. *SCIENCE* **1998,** 280, (29 MAY), 1451-1455.
53. Lauring, A. S.; Andino, R., Quasispecies Theory and the Behavior of RNA Viruses. *PLoS Pathogens* **2010,** 6 (7), e1001005.
54. Wills, P. R.; Carter, C. W., Jr, Insuperable problems of an initial genetic code emerging from an RNA World. *BioSystems* **2018,** 164, 155-166.
55. Bouckaert, R.; Vaughan, T. G.; Sottani, J. B.; Duchene, S.; Fourment, M.; Gavryushkina, A.; Heled, J.; Jones, G.; Kühnert, D.; De Maio, N.; Matschiner, M.; Mendes, F. K.; Müller, N. F.; Ogilvie, H.; du Plessis, L.; Popinga, A.; Rambaut, A.; Rasmussen, D.; Igor Siveroni; Suchard, M. A.; Wu, C.-H.; Xie, D.; Zhang, C.; Stadler, T.; Drummond, A. J., BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* **2019,** 15, (4), e1006650.
56. Goldman, A. D.; Beatty, J. T.; Landweber, L. F., The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J Mol Evol.* **2016,** 82, (Jan (1)), 17-26.
57. Hemmingsen, J. M.; Gernert, K. M.; Richardson, J. S.; Richardson, D. C., The tyrosine corner: a feature of most Greek key beta-barrel proteins. *Protein Science* **1994,** 3, (11), 1927-37.
58. Smith, J. L., Enzymes of nucleotide synthesis. *Current Opinion in Structural Biology* **1995,** 5, 752-757.
59. Perona, J. J.; Gruic-Sovulj, I., Synthetic and Editing Mechanisms of Aminoacyl-tRNA Synthetases. *Topics in Current Chemistry* **2014,** 344, 1-41.
60. Bullock, T.; Uter, N.; Nissan, T. A.; Perona, J. J., Amino Acid Discrimination by a class I aminoacyl-tRNA synthetase specified by negative determinants. *J. Mol. Biol.* **2003,** 328, 395-408.
61. Sever, S.; Rogers, K.; Rogers, M. J.; Carter, C. W., Jr.; Söll, D., *Escherichia coli* tryptophanyl-tRNA synthetase mutants selected for tryptophan auxotrophy implicate the dimer interface in optimizing amino acid binding. *Biochemistry* **1996,** 35, 32-40.
62. Carter, C. W., Jr.; Wills, P. R., The Roots of Genetic Coding in Aminoacyl-tRNA Synthetase Duality *Annual Review of Biochemistry* **2021,** 90, 349-373.
63. Shore, J.; Holland, B. R.; Sumner, J. G.; Nieselt, K.; Wills , P. R., The Ancient Operational Code is Embedded in the Amino Acid Substitution Matrix and aaRS Phylogenies. *Journal of Molecular Evolution* **2019,** 88, 136–150.
64. Koonin, E. V., *The Logic of Chance: The Nature and Origin of Biological Evolution*. Pearson Education; FT Press Science: Upper Saddle River, NJ, 2011.
65. Morrison, D. A., Multiple Sequence Alignment is not a Solved Problem. *arXiv* **2018**, 1808.07717.
66. Popinga, A. From the Origins of Life to Epidemics: Bayesian Inference, Stochastic Simulation, and Dynamics of Bioinformatic Systems. Doctoral, University of Auckland: Supplementary Data. http://github.com/alexpopinga/aaRS-Pipeline, accessed 11 April 2019, Auckland, NZ, 2019.
67. Söding, J.; Biegert, A.; Lupas, A. N., The HHpred interactive server for protein homology detection and structure prediction. *Nucl. Acids Res.* **2005,** 33, W244–W248

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 December 2021 doi:10.20944/preprints202112.0314.v1

68. Webb, B.; Sali, A., Comparative Protein Structure Modeling Using Modeller. In *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc.,: 2016; Vol. 54, pp 5.6.1-5.6.37.
69. Chaliotis, A.; Vlastaridis, P.; Mossialos, D.; Ibba, M.; Becker, H. D.; Stathopoulos, C.; Amoutzias, G. D., The complex evolutionary history of aminoacyl-tRNA synthetases. *Nucleic Acids Res.* **2017,** 45, (No. 3), 1059–1068.
70. Aziz, M. F.; Caetano-Anollés, K.; Caetano-Anollés, G., The early history and emergence of molecular functions and modular scale-free network behavior. *Scientific Reports* | **2016,** 6, 25058.
71. Eme, L.; Doolittle, W. F., Archaea. *Current Biology* **2015,** 25, (October 5), R845–R875.
72. Forterre, P., The Common Ancestor of Archaea and Eukarya Was Not an Archaeon. *Archaea* **2013,** 2013, 372396.
73. Blouin, C.; Butt, D.; Roger, A. J., Rapid evolution in conformational space: A study of loop regions in a ubiquitous GTP binding domain. *Prot. Sci.* **2004,** 13, 608–616.
74. Romeroa, M. L. R.; Yang, F.; Lind, Y.-R.; Toth-Petroczy, A.; Berezovsky, I. N.; Goncearenco, A.; Yang, W.; Wellner, A.; Kumar-Deshmukh, F.; Sharon, M.; Baker, D.; Varani, G.; Tawfik, D. S., Simple yet functional phosphate-loop proteins. *Proc. Nat. Acad. Sci. USA* **2018,** 115, (51), E11943–E11950.
75. Trifonov, E. N.; Kirzhner, A.; Kirzhner, V. M.; Berezovsky, I. N., Distinct Stages of Protein Evolution as Suggested by Protein Sequence Analysis. *J. Mol. Evol.* **2001,** 53, 394–401.
76. Berezovsky, I. N.; Grosberg, A. Y.; Trifonov, E. N., Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters* **2000,** 466, 283-286.
77. Alva, V.; Söding, J.; Lupas, A. N., A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **2015,** 4, e09410.
78. Guo, M.; Schimmel, P., Essential nontranslational functions of tRNA synthetases. *Nature Chemical Biology* **2013** 9, (march), 145-153.
79. Humphrey, W.; Dalke, A.; Schulten, K., VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996,** 14, 33-38.
80. Carter, C. W., Jr.; Carter, C. W., Protein Crystallization Using Incomplete Factorial Experiments. *Journal of Biological Chemistry* **1979,** 254, 12219-12223.
81. Yang, Z., Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *J Mol. Evol.* **1994,** 39, 306-314.
82. Gonnet, G. H.; Cohen, M. A.; Benner, S. A., Exhaustive matching of the entire protein sequence database. *Science* **1992,** 256, 1443–1445.
83. Bouckaert, R. R., DensiTree: making sense of sets of phylogenetic trees. *BIOINFORMATICS* **2010,** 26 (10), 1372–1373.
84. Rambaut, A. *FigTree*, 1.4.0; University of Edinburgh: 2010.
85. SAS *JMP: The Statistical Discovery Software*, V.16.0.0; SAS Institute, Cary NC: Cary, NC, 2021.
86. Wills, P. R., Reflexivity, Coding, and Quantum Biology. *BioSystems* **2019,** In preparation.
87. Buehner, M.; Ford, G. C.; Moras, D.; Olsen, K. W.; Rossmann, M. G., D-Glyceraldehyde 3-Phosphate Dehydrogenase: Three Dimensional Structure and Evolutionary Significance. *Proc. Nat. Acad. Sci. USA* **1973,** 70, 3052-3054.
88. Koonin, E. V.; Novozhilov, A. S., Origin and Evolution of the Genetic Code: The Universal Enigma. *IUBMB Life,* **2009,** 61, ((2) February), 99–111.