Class I Aminoacyl-tRNA Synthetase Urzyme and CP1 Modules have Distinct Genetic Origins

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: /0000-0002-2653-4452

[2]Centre for Computational Evolution, University of Auckland, PB 92019, Auckland 1142, New Zealand

ORCID ID: /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: /0000-0002-2670-7624

Running Title: Mosaic Ancestry of Class I Aminoacyl-tRNA Synthetases

Keywords: BEAST, DensiTree, protein mosaic structure, RNA World hypothesis

Abbreviations: aaRS, aminoacyl-tRNA synthetase; CP1, Connecting peptide 1; MSA, multiple sequence alignment; MCMC, Markov chain monte carlo

Abstract

Phylogenetic arguments that the adaptive radiation of the Class I aminoacyl-tRNA synthetases, aaRS, occurred after the genetic code had been extensively stabilized conflict with recent experimental data showing that ancestors of the two aaRS Classes arose on opposite strands of a single gene. The latter result necessarily implies that phylogenies for both Classes trace from a time when genetic coding used protein aaRS restricted to a far simpler amino acid alphabet. This conundrum resolves if transitions that increased the amino acid alphabet were achieved by accreting new genetic modules that today dominate the aaRS phylogenetic trees sufficiently to force convergence to a more recent date than the ancestral (i.e. Urzyme) modules. We test this hypothesis by showing that the Urzymes and comparably conserved structures from the Connecting Peptide 1 (CP1) insertions present in all Class I aaRS, and which now dominate the mass of contemporary aaRS, have significantly different degrees of conservation and inconsistent phylogenetic trees. This work represents the first time that i) sophisticated structure-based alignment has been used to curate multiple sequence alignments identifying what is conserved across an entire protein superfamily; ii) state-of-the-art phylogenetics has been used to search for high-resolution modularity in protein sequences; and iii) quantitative evidence has been adduced to show that different modules within a protein structure could have distinct ancestral genetic origins.

40 ## Introduction

41 The ancestry of the aminoacyl-tRNA synthetases, aaRS, is a quintessential chicken and
42 egg puzzle whose detailed solutions are essential for understanding the origin of coded protein
43 synthesis. How did reflexive protein machinery arise that could begin converting the symbolic
44 information dormant in nucleic acid sequences of base triplets into sequences of amino acids
45 according to a genetic coding table consistent with the Universal Genetic Code? For half a century,
46 the dominant answer has been that RNA alone was capable of inventing coded proteins (Woese,
47 C.R., Olsen, G.J., et al. 2000, Koonin, E.V. and Novozhilov, A.S. 2009, Koonin, E.V. 2011,
48 Koonin, E.V. and Novozhilov, A.S. 2017).

49 The most important phylogenetic evidence for the RNA World hypothesis is that, although
50 the accepted phylogenies of the Class I aaRS are rooted well before the Last Universal Common
51 Ancestor (LUCA) (Aravind, L., Anantharaman, V., et al. 2002, Fournier, G.P., Andam, C.P., et al.
52 2011, Caetano-Anollés, G., Wang, M., et al. 2013, Fournier, G.P. and Alm, E.J. 2015, Katsnelson,
53 M.I., Wolf, Y.I., et al. 2018), they appear also to have radiated after major bifurcations in the
54 immense superfamily (Aravind, L., Anantharaman, V., et al. 2002) containing folds based on the
55 Rossmann dinucleotide-binding fold (Buehner, M., Ford, G.C., et al. 1973). Genetic coding by
56 protein aaRS is therefore, paradoxically, thought to have replaced a prior implementation based
57 on ribozymal assignment catalysts.

58 We have advocated the opposite conclusion that protein aaRS co-evolved with the genetic
59 code from the simplest origins via translation of Class I and II ancestors from opposite strands of
60 the same bidirectional ancestral gene (Rodin, S.N. and Ohno, S. 1995, Carter, C.W., Jr. and Duax,
61 W.L. 2002, Pham, Y., Li, L., et al. 2007, Rodin, A., Rodin, S.N., et al. 2009, Li, L., Weinreb, V.,

62 et al. 2011, Chandrasekaran, S.N., Yardimci, G., et al. 2013, Li, L., Francklyn, C., et al. 2013,
63 Carter, C.W., Jr., Li, L., et al. 2014, Carter, C.W., Jr and Wills, P.R. 2018b, Carter, C.W., Jr and
64 Wills, P.R. 2018a, Wills, P.R. and Carter, C.W., Jr 2018), obviating the need for ribozymal
65 assignment catalysts. A hypothesis offered to reconcile the obvious and, we believe legitimate,
66 contradiction posed by the two origin scenarios is that contemporary protein structures are mosaics
67 composed of genetically distinct modules—some ancient and authentically related to the true
68 ancestral aaRS—to which more modern modules were assimilated at different points along their
69 phylogenetic timeline, in response to different sorts of selection pressures. Such a timeline could
70 undermine construction of reliable trees from full-length contemporary proteins.

71 In support of our view, the evolutionary history of the aaRS superfamilies has been
72 supplemented by extensive experimental deconstruction (Carter, C.W., Jr. 2017). Notably,
73 experimental (Pham, Y., Li, L., et al. 2007, Carter, C.W., Jr. 2014, Martinez, L., Jimenez-
74 Rodriguez, M., et al. 2015) and bioinformatic (Chandrasekaran, S.N., Yardimci, G., et al. 2013)
75 evidence that the two aaRS Classes descended from opposite strands of a single bidirectional gene
76 assigns a much earlier ancestry to segments constrained by bidirectional coding than to other
77 regions of contemporary aaRS structures. The aaRS consequently represent a uniquely promising
78 opportunity to test the hypothesis that contemporary enzymes are mosaic structures assembled
79 from detectably different genetic ancestors.

80 Methods

81 Algorithmic sequence alignment requires some measure of amino acid similarity to
82 optimize alignments on the assumption that it is an alias for the overall function of a particular
83 amino acid in that position of the polymer. Any MSA is therefore only as good as the assumptions

84 used to establish homologous positions in different sequences (Morrison, D.A. 2018). It is a
85 distinct step again to assume that mutation and selection are the cause of the phylogenetic variation.
86 We superposed the available bacterial Class I aaRS crystal structures to free the multiple
87 sequence alignment from empirical substitution matrices, basing it instead on three-dimensional
88 structural homology. Popinga (Popinga, A. 2019) assembled a curated extension of that alignment
89 of Class I aaRS by extensive use of HHpred (Söding, J., Biegert, A., et al. 2005) and MODELLER
90 (Webb, B. and Sali, A. 2016) to thread amino acid sequences for close relatives of each 3D
91 structure. Because previous work (Chandrasekaran, S.N., Yardimci, G., et al. 2013) suggested that
92 bacterial aaRS sequences are closer to the root than those from other domains, we included only
93 eubacterial aaRS. The resulting MSA contained roughly 20 sequences for each of 10 Class I aaRS.
94 Within the constraints of the available crystal structures, it is as representative as possible of the
95 tree of life.
96 Phylogenetic trees constructed from these structure alignments confirmed that the physical
97 chemistry differentiating the 10 amino acids in each Class (Carter, C.W., Jr. and Wolfenden, R.
98 2015, Wolfenden, R., Lewis, C.A., et al. 2015, Carter, C.W., Jr. and Wolfenden, R. 2016) underlies
99 both the aaRS phylogenies and the amino acid substitution matrices generally used to build such
100 trees (Shore, J., Holland, B.R., et al. 2019). Exclusive use of strict structural homology circumvents
101 as many problems as possible in aligning sequences that diverged to produce different substrate
102 specificities in the pre-LUCA æon. Thus, the MSA is as independent as possible of arbitrary amino
103 acid substitution matrices.

104 The full MSA in this work was assembled from the superposed crystal and threaded 3D
105 structures by eliminating all non-conserved structures and connecting loops, which have different
106 lengths in aaRS for different amino acids, leaving the strictly-conserved secondary structures,

107 which we refer to as the Class I scaffold (Wills, P.R. 2014). The idiosyncratic Class I anticodon-
108 binding domains cannot be structurally superimposed. Despite limited regions of homology,
109 Motifs in CP1 outside the scaffold, and the second connecting peptide in the largest subclass IA
110 ValRS, LeuRS, and IleRS are only slightly less idiosyncratic. As noted in detail further on,
111 evidence that these the Class I scaffold motifs are genetically homologous is substantially stronger
112 than for the remaining variable parts of the Class I superfamily, ensuring a focus on sequences
113 with the strongest evidence of divergent evolution and the most confident phylogenies.

114 The MSA for the Class I scaffold was output using VMD (Humphrey, W., Dalke, A., et al.
115 1996), and is provided in the supplement together with all subsets used in this work. The scaffold
116 fasta file was then partitioned along lines of the experimental deconstruction of the Class I aaRS
117 superfamily (Carter, C.W., Jr. 2017) into the disjointed segments of the Urzyme, separated by
118 structurally conserved segments from the variable length insertion called Connecting Peptide 1
119 (CP1 (Burbaum, J. and Schimmel, P. 1991); Fig. 1). Finally, because the Urzyme (83) and CP1
120 (20) MSAs have different numbers of amino acids, seven 20-amino acid subsets distributed
121 throughout the Urzyme were selected by a balanced, randomized procedure (Carter, C.W., Jr. and
122 Carter, C.W. 1979) to test the effect of sequence length on our phylogenetic metrics.

123 We describe a two-dimensional comparison of complementary metrics of the three
124 different partitions—protozyme, the full Class I Urzyme, and CP1—of the overall MSA. The
125 conservation quality, Q_j, computed by Clustal (Thompson, J.D., Higgins, D.G., et al. 1994) from
126 amino acids filling a single sequence position in different structural homologs. When averaged
127 over all positions within a partition of the MSA, the resulting average, $\langle Q \rangle$ gives a position-by-
128 position metric of sequence variation down each column. The mean phylogenetic support, $\langle S \rangle$, in

129 contrast, is a metric derived from row-by-row comparison with phylogenetic trees derived across
130 the entire sequence length.

131 Molecular kinships between rows in each partition were then compared by clustering the
132 sequences into clades. Phylogenetic trees were computed using BEAST (Bouckaert, R., Vaughan,
133 T.G., et al. 2019), according to multiple amino acid substitution matrices (primarily WAG and
134 LG) for structurally aligned residues after deleting indels from the MSAs. Trees were visualized
135 with DensiTree (Bouckaert, R.R. 2010) and FigTree (Rambaut, A. 2010). For the intact Class I
136 scaffold and all partial alignments, DensiTree was used to extract the estimated support for clades
137 for each amino acid generated from ~10,000 trees sampled by Markov Chain Monte Carlo
138 (MCMC) simulation. For each aaRS, this support is the percentage of all trees for which the tips
139 descend from the same most recent common ancestor with no descendants leading to a different
140 aaRS type, i.e. for which that aaRS appears to be monophyletic.

141 Columns were appended for various predictors as noted in the text to the table of $\langle Q \rangle$ and
142 $\langle S \rangle$ values to form the design matrix (Table II) used for multiple regression analysis with the JMP
143 software (SAS 2015). Multiple Regression was used to assess the statistical strength of
144 contributions to conservation metric and the mean support of the predictors (i.e., the five
145 independent variables, A, B, C, CP1 and the number of amino acids) using stepwise searches to
146 identify the best set of predictors followed by least squares analysis.

147 # Results

148 ## Deep ancestral aaRS phylogenies are better characterized, structurally and
149 experimentally, than those of any other protein superfamilies.

150 Previous phylogenetic analyses of the Class I aaRS superfamily (Wolf, Y.I., Aravind, L.,
151 et al. 1999, Aravind, L., Anantharaman, V., et al. 2002, Leipe, D.D., Wolf, Y.I., et al. 2002,
152 O'Donoghue, P. and Luthey-Schulten, Z. 2003, Roach, J.M., Sharma, S., et al. 2005, Caetano-
153 Anollés, G., Wang, M., et al. 2013, Shore, J., Holland, B.R., et al. 2019) generally agree that the
154 clades for each amino acid are monophyletic, and that the ten families fall into three subclasses
155 (Carter, C.W., Jr. 1993, Cusack, S. 1995), including IleRS, ValRS, LeuRS, and MetRS (Subclass
156 IA); GluRS and GlnRS (Subclass IB); and TyrRS and TrpRS (Subclass IC). As we restrict this
157 analysis to the bacterial aaRS, we exclude the archaeal Class I LysRS, which by consensus belong
158 to Subclass IB. There is also consensus that the remaining CysRS and ArgRS families, which were
159 assigned originally to Subclass IA (Cusack, S. 1995), are difficult to assign, sometimes appearing
160 independently, with one or the other appearing instead with Subclass IB (for example, see Roach,
161 J.M., Sharma, S., et al. 2005).

162 We test the hypothesis that Class I aaRS catalytic domains contain multiple segments with
163 distinct genetic origins (Fig. 1). Genetic deconstruction of Class I aaRS modules (Carter, C.W., Jr.
164 2014) revealed that both aaRS Classes contain a small core whose ~120-130 amino acids comprise
165 little more than an intact active site. Those constructs retain a major fraction of full-length aaRS
166 catalytic proficiency—estimated as the transition-state stabilization free energy—for both amino
167 acid and tRNA substrates (Pham, Y., Kuhlman, B., et al. 2010, Li, L., Weinreb, V., et al. 2011, Li,
168 L., Francklyn, C., et al. 2013). Because of this high catalytic proficiency, these cores were called

169 "Urzymes", from the German prefix "Ur" = original, authentic. Class I and II Urzymes retain the
170 ability to distinguish between amino acids from the two sets of substrates, Class I Urzymes
171 activating Class I amino acids in preference to Class II amino acids by ~1.0 kcal/mole, and
172 conversely (Carter, C.W., Jr., Li, L., et al. 2014, Carter, C.W., Jr 2015).

173 Even smaller subsets of 46 amino acids from both Class I and II retain nearly half of the
174 catalytic proficiency in the critical amino-acid activation reaction (Martinez, L., Jimenez-
175 Rodriguez, M., et al. 2015). These modules were called "protozymes", from the Greek προτο =
176 first, to distinguish them from Urzymes. Fully active Class I and II protozymes were expressed
177 from opposite strands of a single gene designed to encode their structures on opposite strands.
178 Experimental Michaelis-Menten parameters of all four protozymes—Class I and II; native and
179 bidirectional—are the same, within experimental error (Martinez, L., Jimenez-Rodriguez, M., et
180 al. 2015). The Class I protozyme coincides with the N-terminal crossover connection of the
181 Rossmann fold, which contains a distinctive 3D packing motif that recurs in ~25% of the proteome
182 (Cammer, S. and Carter, C.W., Jr. 2010) and imposes distinct, functionally relevant
183 conformational states tightly coupled to catalysis in TrpRS (Weinreb, V., Li, L., et al. 2012, Li, L.
184 and Carter, C.W., Jr 2013, Weinreb, V., Li, L., et al. 2014, Carter, C.W., Jr. 2019).

185 The progressive biochemical functionality of aaRS protozymes, Urzymes, and catalytic
186 domains; the fact that their structures represent universally conserved portions within both
187 families; and the strong evidence that they descended from a single bidirectional ancestral gene all
188 imply that they are legitimate experimental models for ancestral evolutionary intermediates.
189 Moreover, to date, such sequential models have been derived solely for the two aaRS
190 superfamilies. Evidence that their essential catalytic architectures arose from a bidirectional gene
191 implies, *ipso facto*, that sequences inconsistent with bidirectional coding represent accretion of

192 new genetic material. Thus, more is known about the modular evolution of the two aaRS
193 superfamilies than for any other such superfamily. Those data, in turn, predict quantitative
194 differences between MSAs for modules derived from different sources.

195 The variable CP1 insertion (Burbaum, J.J., Starzyk, R.M., et al. 1990, Burbaum, J.J. and
196 Schimmel, P. 1991) divides the Class I signature Rossmann dinucleotide binding fold into roughly
197 equal parts, always intersecting the Rossmann fold between residues ~4.5 Å apart, so that it can
198 be replaced without structural disruption with a single peptide bond to produce the Class I Urzyme
199 (Pham, Y., Li, L., et al. 2007). CP1 insertions include the large editing domains of the aaRS for
200 aliphatic amino acids, and are thus responsible for the variable lengths of Class I aaRS catalytic
201 domains. As the CP1 insertion is incompatible with bidirectional genetic coding of ancestral Class
202 I and II aaRS (Fig. 1B), its introduction would necessarily have ended bidirectional coding of Class
203 I and II aaRS. For that reason alone, it is worthwhile to examine, quantitatively, the phylogenetic
204 relationships between the Class I protozyme, Urzyme, and CP1 sequences.

205 ## Structural homology across the entire Class I superfamily extends beyond the Urzyme.

206 Independent 3D superposition of aaRS crystal structures (Pham, Y., Li, L., et al. 2007,
207 Wills, P.R. 2014) revealed considerable structural conservation within the catalytic domains of all
208 ten members of each Class—see Fig. 1 of (Wills, P.R. 2014). We call this structurally conserved
209 core the Class I scaffold. The Class I Urzyme boundaries delimited decisively by bidirectional
210 coding of Class II Urzymes constitute only about 80% of this scaffold; the remainder consists of
211 20 σ-helical residues in roughly equal segments close to each end of the CP1 insertion.

212 ## Class I CP1 sequences are significantly less well-conserved than Urzyme sequences.

213 Position-by-position conservation in comparably conserved structures from the Class I
214 scaffold is significantly higher in sequences within the Urzyme than in CP1 (Fig. 2 and Table 1).
215 The $\langle Q \rangle$ for the CP1segment is smaller by 3-6 times than the standard error of the mean. To assess
216 the significance, we estimated the regression model $\langle Q \rangle_{\text{calc}}$ = Constant + β*[CP1], where [CP1]
217 = 1 if and only if the position is within the CP1 segment and 0 otherwise. This model was selected
218 from a stepwise regression testing effects of all four segments, together with their two-way
219 interactions. It was the only legitimate model. Statistics for the model are given in Table I; the β-
220 coefficient for [CP1] is negative and its P-value is 1.3E-5. Columns of the MSA derived from CP1
221 therefore are significantly less well-conserved than those derived from the Uryzme.

222 ## Class I Urzyme- and CP1-based trees form distinctly different clades.

223 Urzyme and CP1 partitions of the MSAs produce quite different trees (Fig. 3). In particular,
224 although all ten Class I aaRS clades are monophyletic in the Urzyme trees, three clades in the CP1
225 trees—MetRS, ValRS, and TyrRS—are not monophyletic. Moreover, the Urzyme clades are
226 constrained by tight envelopes, whereas the CP1 clades have poorly defined envelopes. Thus,
227 although the 3D structures from which the two partitions were derived are apparently equally
228 homologous, the phylogenetic behaviors of their amino acid sequences differ markedly.

229 Support, S_i, the fraction of all trees for which each aaRS clade, i, is monophyletic, was
230 averaged over all Class I aaRS types to give the mean support, $\langle S \rangle = \sum S_i / 10$. That analytical
231 operation was repeated for trees built for the full scaffold sequence alignment, (Full), the Urzyme,
232 CP1, protozyme (fragment A in Fig. 1), and seven subsets consisting of twenty amino acids in

233 blocks of five residues as described in Methods. Data for these eleven rows of the design matrix
234 (Table II) were evaluated using the conventional WAG (Whelan, S. and Goldman, N. 2001)
235 substitution matrix and repeated using the more recent LG (Pokarowski, P., Kloczkowski, A., et
236 al. 2007) substitution matrix used in (Shore, J., Holland, B.R., et al. 2019).

237 Contributors to the variance of $\langle S \rangle$ were assessed using stepwise multiple regression,
238 which resulted in a single model, summarized in Figure 4 and Table III. The optimal regression
239 model was: $\langle S \rangle = 0.63 + 0.06 * A - 0.07 * CP1 + 0.05 * A * CP1$. All three coefficients are
240 significant at better than the 99% confidence level, with P-values < 0.00001. As suggested by its
241 position in the regression curve in Figure 4, trees built from CP1 residues have significantly less
242 support than those from the Urzyme. Residues from the protozyme contribute more positively to
243 the average support than those derived from elsewhere in the Urzyme. Notably, the number of
244 residues in a given block from the MSA has no detectable effect on either $\langle Q \rangle$ or $\langle S \rangle$.

245

246 Discussion

247 Roots of the two aaRS superfamilies are likely deeper than for all or most other protein
248 superfamilies (Carter, C.W., Jr. 2017), so they present a unique challenge. Results presented here
249 open a new window on deep protein phylogenetics (Gaucher, E.A., Kratzer, J.T., et al. 2010),
250 addressing questions that have been previously recognized (Trifonov, E.N. 1995, Sobolevsky, Y.,
251 Frenkel, Z.M., et al. 2007) but until now never quantitatively addressed. We are motivated here by
252 a desire to clarify the ancestry of Class I aaRS and especially the argument that they branched from
253 related parts of the proteome late, relative to the formulation of the genetic coding table. That

contention (Koonin, E.V. and Novozhilov, A.S. 2009, Koonin, E.V. and Novozhilov, A.S. 2017) is crucial to understanding the origin of coding, because it represents the strongest, and perhaps the only, phylogenetic evidence favoring the RNA World hypothesis.

Our long-range goal is to understand the evolution of aaRS•tRNA cognate pairs and thence the order of addition of entries in the Universal genetic coding table. Construction of trees for aaRS using any amino acid substitution matrix is necessarily 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, starting from 2 and ending at 20 as the Class I and II trees branch into multiple families. Thus, if different motifs within a protein have different evolutionary histories, it is uncertain—in the absence of a broader experimental and structural context—what should be inferred either from a lack of tight congruence or the presence of strong congruence of the branching pattern.

These difficulties are significantly offset by the extensive data assembled from the construction and experimental characterization of aaRS protozymes (Martinez, L., Jimenez-Rodriguez, M., et al. 2015) and Urzymes (Li, L. and Carter, C.W., Jr 2013, Li, L., Francklyn, C., et al. 2013, Carter, C.W., Jr. 2014, Carter, C.W., Jr., Li, L., et al. 2014, Weinreb, V., Li, L., et al. 2014, Carter, C.W., Jr 2015, Carter, C.W., Jr. 2017). The quantitative, parallel increases in activity with comparable additions of mass (Martinez, L., Jimenez-Rodriguez, M., et al. 2015) together with the phylogenetic evidence for bidirectional coding ancestry (Chandrasekaran, S.N., Yardimci, G., et al. 2013) establish these constructs as legitimate evolutionary intermediates. Such intermediates are manifestly unique to the aaRS at present. Substantially more data therefore constrain their deep ancestry than for any other protein superfamilies (Carter, C.W., Jr. 2017).

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

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

Conservation ($\langle Q \rangle$) and congruency ($\langle S \rangle$) metrics are essentially uncorrelated.

Regression of $\langle Q \rangle$ against $\langle S \rangle$ gives $R^2 = 0.37$ ($P = 0.05$). However, this weak correlation results entirely from the fact that the CP1 sequences are significant outliers in both metrics. Excluding CP1, regression gives $R^2 = 0.0$ ($P = 0.85$). The two metrics—conservation down columns and kinship between rows—are uncorrelated metrics, strengthening the implication that CP1 sequences are drawn from a different population.

Distinguishing features of CP1 sequences are evident without considering indels.

Figures. 2-4 and Tables I-III together illustrate a near-optimal quantitative, comprehensive two-dimensional comparison of the sequence data. The crux of what these data tell us is that, despite substantial prior evidence that CP1 is more recent than the Urzyme, its evolutionary divergence—evidenced by $\langle Q \rangle$—is much greater, and its phylogenetic consistency—evidenced by $\langle S \rangle$—is much reduced, relative to are much greater than the corresponding metrics for Urzyme

297 segments. It seems not to matter whether the scaffold for a new aaRS type is produced from its
298 parent structure by strict mutation-only or also includes indel processes.

299 The success of the mutation-only model to account for the functional expansion of the
300 aaRSs in parallel with the expansion of the alphabet (in contrast to the CP1 development)
301 underscores the adequacy of that model in corroborating the hypothesis that CP1 has quite different
302 properties that likely imply a distinct evolutionary history. That this conclusion is so evident, even
303 when considering only sequences with strict 1-1 correspondences between 3D crystal structures,
304 and excluding the large and variable-length indels that account for bulk of CP1 in most aaRS types,
305 and their consensus role in enhancing fidelity (Carter, C.W., Jr. 2017), emphasizes its importance.

306 Urzyme and CP1 sequences exhibit statistically different properties.

307 By both metrics, Urzyme-derived amino acid sequences—i.e. those that are consistent with
308 bidirectional coding ancestry—differ, statistically, from those derived from comparably conserved
309 segments within CP1. These new data suggest that Urzyme and CP1 sequences have distinct
310 evolutionary histories. The log-worth values (i.e., -log(P)) for loss of conservation and lack of
311 congruence for phylogenetic trees are 4.9 and 7.7, implying very high statistical significance that
312 the respective metrics for CP1 and Urzyme sequences are drawn from different populations. The
313 numbers of residues in different segments drawn from the MSA have no detectable impact on the
314 $\langle S \rangle$ score, implying that they have an insignificant impact on its variance.

315 If the subsets of Class I aaRS sequences we examine here were all drawn from continuously
316 replicated ancient genetic sources, subject to comparable selection history since their emergence,
317 the null hypotheses for our analyses would produce similar trees and similarly conserved
318 sequences for the Urzyme and CP1 partitions. By showing that both position-by-position and row-

319 by-row analyses of the Urzyme and CP1 partitions of the Class I scaffold MSA produce
320 significantly different statistical metrics, we corroborate the hypothesis—based on bidirectional
321 coding ancestry—that the CP1 sequences derive from a more recent acquisition of genetic
322 information than the Urzyme. Moreover, and by a similar argument, the data are also consistent
323 with the hypothesis that protozyme sequences (segment A) were, themselves, drawn from a
324 distinct, perhaps even earlier population, as previously proposed (Carter, C.W., Jr and Wills, P.R.
325 2018b).

326 Neither convergent evolution nor horizontal gene transfer are likely explanations for the 327 Urzyme/CP1 distinction.

328 The congruent clade structure of Urzyme-derived sequences from the scaffold fall into
329 consensus groupings characterizing Class I aaRS as a coherent superfamily and does not require
330 mechanisms beyond mutation and selection from a single common ancestor. The aberrant behavior
331 of sequence variations in the CP1 insertion (Figs. 2-4) does, however, appear to require appeal to
332 such processes. Treatments of Class I aaRS evolution based on full MSAs (Wolf, Y.I., Aravind,
333 L., et al. 1999, Shiba, K. 2005, Soucy, S.M., Huang, J., et al. 2015) show evidence of horizontal
334 gene transfer (HGT), genetic transpositions and large scale insertion/deletion events, of which CP1
335 is the foremost example. Wherever full-length bacterial Class I enzymes with a particular amino
336 acid substrate specificity are represented by more than one canonical structure, as has been
337 described for IleRS and MetRS (see Fig. 3 in Shiba, K. 2005), that bipartite distribution in genome
338 space is adequately explained in terms of early HGT into bacteria from an archaean/eukaryotic
339 ancestor, but not in terms of convergent evolution. CP1 sequence disparities at homologous
340 sequence positions in TyrRS, MetRS, and ValRS behave in an 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 would suggest HGT.

Descent of Class I and II aaRS from a single bidirectional ancestral gene (Rodin, S.N. and Ohno, S. 1995, Rodin, A., Rodin, S.N., et al. 2009, Chandrasekaran, S.N., Yardimci, G., et al. 2013, Carter, C.W., Jr., Li, L., et al. 2014, Carter, C.W., Jr 2015, Martinez, L., Jimenez-Rodriguez, M., et al. 2015, Carter, C.W., Jr. 2017) together with the notion (Carter, C.W., Jr and Wills, P.R. 2018b, Wills, P.R. and Carter, C.W., Jr 2018, Carter, C.W., Jr. and Wills, P.R. 2020) that the genetic coding table evolved as pre-existing aaRS genes bifurcated to produce related aaRS whose tRNA and amino acid specificities diverged to enable the daughter synthetases to differentiate between amino acids in groups that previously had functioned as synonymous underscore the likelihood that the aaRS of both families diverged, rather than converging to similar functions from different sources.

The discrepancy between CP1 and Urzyme evolutionary metrics is likely to have originated because CP1 is of later origin. Its insertion into members of the quasi-species distributions of different aaRS types perhaps relied on idiosyncratic features of receptive variants. The subsequent coevolution of CP1 and Urzyme sequences would have necessarily preserved Urzyme functionality, while the new CP1 could adapt flexibly to its developing role in different ways according to the extant context of the Urzyme's functional mode as we describe in the final section, fixing the apparent orthologs in different ways that may be related to HGT.

361 Primordial aaRS were quasispecies that covered progressively smaller regions of sequence
362 space, the further back in time toward the "root".

363 AARS evolution cannot have been an ordinary mutation/selection process. All solutions to
364 the problem normally framed as the "chicken and egg problem"(Tyson, N.d. 2013) imply a
365 continuity principal: at any stage during the emergence of coding, some prior system must have
366 been interpreted extant protogenes. A key aspect of such prior systems is that as the alphabet size,
367 diversity, and consequent fidelity increased, they would, *ipso facto*, have created more narrowly
368 targeted selection pressures, strongly coupling mutation and selection in early stages of genetic
369 coding (Fig. 5). Their behaviors would have depended on the relationship between the shapes of
370 fitness landscapes and error rates of catalysis by the extant quasispecies (Lauring, A.S. and
371 Andino, R. 2010). The earliest alphabets, including at least the first two amino acid types, were
372 also tightly coupled by the additional constraint of bidirectional coding (gray shaded background,
373 Figure 5). These characteristics are, of course, the fundamental unknowns we wish, ultimately, to
374 infer, underscoring the primary importance of this work.

375 We have argued (Carter, C.W., Jr and Wills, P.R. 2018b, Wills, P.R. and Carter, C.W., Jr
376 2018) that in that historical context a single functional island in sequence space (i.e. quasispecies)
377 would invariably have been a strong attractor, irrespective of the detailed shape of the fitness
378 landscape that stabilized it, in that all mutations would move the system slightly off its optimum
379 phenotype but then be subject to strongly restorative selection pressures. On the other hand, if
380 ancestral protozymes from a bidirectional gene had broad, relatively flat (and necessarily
381 codependent) fitness landscapes, matched to the correspondingly high error rates, that could have
382 favored the emergence of bifurcated quasispecies that enhanced genetic coding by recruiting new

383 amino acid types and simultaneously increasing the precision with which child specificities could
384 be encoded (Fig. 5).

385 High-resolution structural modularity is consistent with functionally significant
386 discontinuities in the evolution of genetic coding.

387 The shortest CP1 insertions have ~75 residues in TrpRS and TyrRS. That segment recurs
388 essentially intact in the longer CP1 insertions present in the remaining eight Class I aaRS (Pham,
389 Y., Li, L., et al. 2007, Wills, P.R. 2014). It appears to be required to form an annulus around the
390 Urzyme (Fig. 1A,C), so it seems likely that the initial insertion was that long. For that reason the
391 original hypothesis as to its origin (Pham, Y., Li, L., et al. 2007) was that it resulted from the
392 essentially simultaneous insertion of some form of mobile genetic element in all extant Class I
393 Urzymes, in which case its actual root sequence would have been both more recent than that of the
394 Urzymes, and earlier than the earliest contemporary full-length aaRS enzymes. That, alone, could
395 generate distinctive behavior of CP1 segments in the scaffold MSA.

396 Different background colors in Figure 5 denote how branching of the tree to allow
397 introduction of the nth amino acid into the alphabet enables a highly cooperative re-optimizing of
398 the n-1 aaRS types functions already present. As each new, refined amino acid type emerges, all
399 extant gene sequences adapt to opportunities introduced by progressively finer discrimination
400 between amino acid side chain physical chemistry. In turn, adaptation to a more diverse alphabet
401 sharpens the new aaRS fidelities (i.e., the contraction of branch thicknesses. That cooperativity
402 creates a Lamarckian correlation between selection pressure and its outcome—the result of any
403 mutation being nearly synonymous with the selection pressure it faced, especially at the beginning
404 of the differentiation of the code.

405 Insertion of the CP1 motif necessarily ended bidirectional coding, thus dividing aaRS
406 evolutionary histories qualitatively into distinct phases (between green and yellow backgrounds).
407 A variety of evidence supports the notion that the chief selective advantage represented by the CP1
408 insertion might have been to transcend the fundamental limitation on specificity posed by the
409 Urzyme architecture, allowing substantial increases in the fidelity of the assignment catalysts.
410 Amino acid specificities of Class I and II Urzymes (Carter, C.W., Jr., Li, L., et al. 2014, Carter,
411 C.W., Jr 2015) suggest that even these highly evolved assignment catalysts could activate an amino
412 acid from the correct amino acid Class only 80% of the time, or four times out of five. Urzyme
413 constructs capable of 10^9-fold rate accelerations nonetheless had quite modest fidelity and had
414 only just begun the ascent to 99.99% specificity. In turn, this suggests a fundamental limitation to
415 the precision of which bidirectional coding was ultimately capable.

416 The most obvious contribution of CP1 to fidelity is that the editing domains present in the
417 larger subclass IA aaRS for aliphatic amino acids Ile, Val, and Leu are all elaborations of the CP1
418 motif present in the simplest subclass IC aaRS for Tyr and Trp. Such a function may, however,
419 have developed from the annulus that CP1 forms around two parts of the Urzyme (Fig. 6),
420 constraining the size and shape of the amino acid binding pocket. Several groups have been unable
421 to change specificities of subclass IB Glu vs Gln (Bullock, T., Uter, N., et al. 2003, Perona, J.J.
422 and Gruic-Sovulj, I. 2013) or subclass IC Trp vs Tyr (Sever, S., Rogers, K., et al. 1996) via
423 comprehensive mutation of side chains in the immediate vicinity of the amino acid substrate, all
424 of which are within the Urzyme architecture. In the former work (Bullock, T., Uter, N., et al. 2003),
425 the authors could not alter specificity without making wholesale mutations outside the Urzyme
426 and within the CP1 domain that form the second boundary layer of the amino acid binding site.

427 Following up the latter work (Sever, S., Rogers, K., et al. 1996), a modular thermodynamic

428 cycle consisting of full-length TrpRS, the Urzyme and the Urzyme plus either CP1 or the anticodon
429 binding domain (ABD) showed that rejection of Tyrosine and selection of the indole side chain of
430 tryptophan by *G. stearothermophilus* TrpRS requires cooperation between CP1 and the ABD (Li,
431 L. and Carter, C.W., Jr 2013, Weinreb, V., Li, L., et al. 2014). CP1 must coordinate its movements
432 with those of the ABD to perform the allosteric communication necessary to enhance side chain
433 selectivity beyond the modest capabilities of the Urzyme. In other words, bidirectional coding
434 likely limited Urzyme specificity. Fine tuning specific recognition of amino acid substrates
435 required insertion of CP1 and the ABD, which in turn dramatically transformed the Class I aaRS
436 fitness landscape.

437 Figure 7 summarizes in color several important stages in the development of genetic coding
438 we have identified here that were inaccessible prior to our work on evolutionary intermediates.
439 These stages complement—without necessarily contradicting—the work of Koonin (Koonin, E.V.
440 2011). Importantly, our results support a plausible alternative to the late radiation of assignment
441 catalysts, by breaking code development into steps we can document. The events on which we
442 hope to shed new light lie far deeper in time than all or most other phylogenetic analyses, and yet
443 pose among the most fundamental and challenging questions. That both metrics lend support to,
444 and help make sense of, a substantial body of experimental data relevant to Class I aaRS evolution
445 validates this effort.

446

447 Acknowledgments. This work was supported by NIGMS R01-78227. We are grateful to Corbin
448 Jones and Greg Fournier for helpful feedback in preparing the manuscript and Loren Williams for
449 suggesting that ribosomal protein sequences exhibit a similar genetic mosaicity.

450

451 ## References

452 Aravind L, Anantharaman V, Koonin EV. 2002. Monophyly of Class I Aminoacyl
453 tRNASynthetase, USPA, ETFP, Photolyase, and PP-ATPase Nucleotide-Binding Domains:
454 Implication for Protein Evolution in the RNAWorld. PROTEINS: Struct. Funct. Gen., 48:1–14.

455 Bouckaert R, Vaughan TG, Sottani JB, Duchene S, Fourment M, Gavryushkina A, Heled
456 J, Jones G, Kühnert D, De Maio N*, et al.* 2019. BEAST 2.5: An Advanced Software Platform for
457 Bayesian Evolutionary Analysis. PLoS Comput Biol, In Press.

458 Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees.
459 BIOINFORMATICS, 26 1372–1373.

460 Buehner M, Ford GC, Moras D, Olsen KW, Rossmann MG. 1973. D-Glyceraldehyde 3-
461 Phosphate Dehydrogenase: Three Dimensional Structure and Evolutionary Significance. Proc.
462 Nat. Acad. Sci. USA, 70:3052-3054.

463 Bullock T, Uter N, Nissan TA, Perona JJ. 2003. Amino Acid Discrimination by a class I
464 aminoacyl-tRNA synthetase specified by negative determinants. J. Mol. Biol., 328:395-408.

465 Burbaum J, Schimmel P. 1991. Structural Relationships and the Classification of
466 Aminoacyl-tRNA Synthetases. The Journal of Biological Chemistry, 266:16965-16968.

467 Burbaum JJ, Schimmel P. 1991. Assembly of a Class I tRNA Synthetase from Products
468 of an Artificially Split Gene. Biochem., 30:319-324.

469 Burbaum JJ, Starzyk RM, Schimmel P. 1990. Understanding Structural Relationships in
470 Proteins of Unsolved Three-Dimensional Structure. PROTEINS: Struct. Funct. Gen., 7:99-111.

471 Caetano-Anollés G, Wang M, Caetano-Anollés D. 2013. Structural Phylogenomics
472 Retrodicts the Origin of the Genetic Code and Uncovers the Evolutionary Impact of Protein
473 Flexibility. Plos One, 8:e72225.

474 Cammer S, Carter CW, Jr. 2010. Six Rossmannoid Folds, Including the Class I
475 Aminoacyl-tRNA Synthetases, Share a Partial Core with the Anticodon-Binding Domain of a
476 Class II Aminoacyl-tRNA Synthetase. Bioinformatics, 26:709-714.

477 Carter CW, Jr. 2015. What RNA World? Why a Peptide/RNA Partnership Merits
478 Renewed Experimental Attention. Life, 5:294-320.

479 Carter CW, Jr, Wills PR. 2018a. Did Gene Expression Co-evolve with Gene Replication?
480 In: Pontarotti P editor. Evolutionary Biology: Genome and Phenotype Evolution. Paris,
481 Springer-Nature p. 293-313.

482 Carter CW, Jr, Wills PR. 2018b. Interdependence, Reflexivity, Fidelity, and Impedance
483 Matching, and the Evolution of Genetic Coding. Molecular Biology and Evolution, 35:269-286.

484 Carter CW, Jr. 1993. Cognition, Mechanism, and Evolutionary Relationships in
485 Aminoacyl-tRNA Synthetases. Ann. Rev. Biochem., 62:715-748.

486 Carter CW, Jr. 2014. Urzymology: Experimental Access to a Key Transition in the
487 Appearance of Enzymes. J. Biol. Chem., 289:30213–30220.

488 Carter CW, Jr. 2017. Coding of Class I and II aminoacyl-tRNA synthetases. Advances in
489 Experimental Medicine and Biology: Protein Reviews, 18:103-148.

490 Carter CW, Jr. 2019. Escapement mechanisms: efficient free energy transduction by
491 reciprocally-coupled gating. Proteins: Structure, Function, and Bioinformatics, 88:forthcoming.

492 Carter CW, Jr., Carter CW. 1979. Protein Crystallization Using Incomplete Factorial
493 Experiments. Journal of Biological Chemistry, 254:12219-12223.

494 Carter CW, Jr., Duax WL. 2002. Did tRNA Synthetase Classes Arise on Opposite
495 Strands of the Same Gene? Molecular Cell, 10:705-708.

496 Carter CW, Jr., Li L, Weinreb V, Collier M, Gonzales-Rivera K, Jimenez-Rodriguez M,
497 Erdogan O, Chandrasekharan SN. 2014. The Rodin-Ohno Hypothesis That Two Enzyme
498 Superfamilies Descended from One Ancestral Gene: An Unlikely Scenario for the Origins of
499 Translation That Will Not Be Dismissed. Biology Direct, 9:11.

500 Carter CW, Jr., Wills PR. 2020. The Roots of Genetic Coding in Aminoacyl-tRNA
501 Synthetase Duality Annual Review of Biochemistry, 89; In Preparation.

502 Carter CW, Jr., Wolfenden R. 2015. tRNA Acceptor-Stem and Anticodon Bases Form
503 Independent Codes Related to Protein Folding. Proc. Nat. Acad. Sci. USA, 112 7489-7494.

504 Carter CW, Jr., Wolfenden R. 2016. Acceptor-stem and anticodon bases embed amino
505 acid chemistry into tRNA. RNA Biology, 13:145–151.

506 Carter CWJ. 1993. Cognition, Mechanism, and Evolutionary Relationships in
507 Aminoacyl-tRNA Synthetases. Annual Review of Biochemistry, 62:715-748.

508 Chandrasekaran SN, Yardimci G, Erdogan O, Roach JM, Carter CW, Jr. 2013. Statistical
509 Evaluation of the Rodin-Ohno Hypothesis: Sense/Antisense Coding of Ancestral Class I and II
510 Aminoacyl-tRNA Synthetases. Molecular Biology and Evolution, 30:1588-1604.

511 Cusack S. 1995. Eleven down and nine to go. Nat. Str. Biol., 2:824-831.

512 Fournier GP, Alm EJ. 2015. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA
513 Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. J. Mol. Evol.,
514 80:171-185.

515 Fournier GP, Andam CP, Alm EJ, Gogarten JP. 2011. Molecular Evolution of Aminoacyl
516 tRNA Synthetase Proteins in the Early History of Life. Orig Life Evol Biosph, 41 621–632

517 Gaucher EA, Kratzer JT, Randall RN. 2010. Deep Phylogeny—How a Tree Can Help
518 Characterize Early Life on Earth. Cold Spring Harb. Perspect. Biol., 2:a002238.

519 Humphrey W, Dalke A, Schulten K. 1996. VMD - Visual Molecular Dynamics. Journal
520 of Molecular Graphics, 14:33-38.

521 Katsnelson MI, Wolf YI, Koonin EV. 2018. Towards physical principles of biological
522 evolution. arXiv, orgabs1709.00284.

523 Koonin EV. 2011. The Logic of Chance: The Nature and Origin of Biological Evolution.
524 Upper Saddle River, NJ, Pearson Education; FT Press Science.

525 Koonin EV, Novozhilov AS. 2009. Origin and Evolution of the Genetic Code: The
526 Universal Enigma. IUBMB Life, , 61:99–111.

527 Koonin EV, Novozhilov AS. 2017. Origin and Evolution of the Universal Genetic Code.
528 Annu. Rev. Genet., 51:45–62.

529 Lauring AS, Andino R. 2010. Quasispecies Theory and the Behavior of RNA Viruses.
530 PLoS Pathogens 6 e1001005.

531 Leipe DD, Wolf YI, Koonin EV, Aravind L. 2002. Classification and Evolution of P-loop
532 GTPases and Related ATPases. J. Mol. Biol., 317:41-72.

533 Li L, Carter CW, Jr. 2013. Full Implementation of the Genetic Code by Tryptophanyl-
534 tRNA Synthetase Requires Intermodular Coupling. J. Biol. Chem., 288:34736–34745.

535 Li L, Francklyn C, Carter CW, Jr. 2013. Aminoacylating Urzymes Challenge the RNA
536 World Hypothesis. J. Biol. Chem., 288:26856-26863.

537 Li L, Weinreb V, Francklyn C, Carter CW, Jr. 2011. Histidyl-tRNA Synthetase Urzymes:
538 Class I and II Aminoacyl-tRNA Synthetase Urzymes have Comparable Catalytic Activities for
539 Cognate Amino Acid Activation. J. Biol. Chem., 286:10387-10395.

540 Martinez L, Jimenez-Rodriguez M, Gonzalez-Rivera K, Williams T, Li L, Weinreb V,
541 Chandrasekaran SN, Collier M, Ambroggio X, Kuhlman B*, et al.* 2015. Functional Class I and II
542 Amino Acid Activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. J. Biol.
543 Chem., 290:19710–19725.

544 Morrison DA. 2018. Multiple Sequence Alignment is not a Solved Problem.
545 arXiv:1808.07717.

546 O'Donoghue P, Luthey-Schulten Z. 2003. On the Evolution of Structure in Aminoacyl-
547 tRNA Synthetases. Microbiol. Mol. Biol. Rev., 67:550–573.

548 Perona JJ, Gruic-Sovulj I. 2013. Synthetic and Editing Mechanisms of Aminoacyl-tRNA
549 Synthetases. Topics in Current Chemistry DOI: 10.1007/128_2013_456.

550 Pham Y, Kuhlman B, Butterfoss GL, Hu H, Weinreb V, Carter CW, Jr. 2010.
551 Tryptophanyl-tRNA synthetase Urzyme: a model to recapitulate molecular evolution and
552 investigate intramolecular complementation. J. Biol. Chem., 285:38590-38601.

553 Pham Y, Li L, Kim A, Erdogan O, Weinreb V, Butterfoss G, Kuhlman B, Carter CW, Jr.
554 2007. A Minimal TrpRS Catalytic Domain Supports Sense/Antisense Ancestry of Class I and II
555 Aminoacyl-tRNA Synthetases. Mol Cell, 25:851-862.

556 Pokarowski P, Kloczkowski A, Nowakowski S, Pokarowska M, Jernigan RL, Kolinski A.
557 2007. Ideal amino acid exchange forms for approximating substitution matrices. Proteins,
558 69:379–393.

559 Popinga A. 2019. From the Origins of Life to Epidemics: Bayesian Inference, Stochastic
560 Simulation, and Dynamics of Bioinformatic Systems. Computer Science. Auckland, NZ,
561 University of Auckland: Supplementary Data. http://github.com/alexpopinga/aaRS-Pipeline,
562 accessed 11 April 2019.

563 Rambaut A. 2010. FigTree. University of Edinburgh.

564 Roach JM, Sharma S, Kapustina M, Carter CW, Jr. 2005. Structure alignment via
565 Delaunay tetrahedralization. PROTEINS: Struct. Funct. Bioinf., 60:66-81.

566 Rodin A, Rodin SN, Carter CW, Jr. 2009. On Primordial Sense-Antisense Coding.
567 Journal of Molecular Evolution, 69:555-567.

568 Rodin SN, Ohno S. 1995. Two Types of Aminoacyl-tRNA Synthetases Could be
569 Originally Encoded by Complementary Strands of the Same Nucleic Acid. Orig. Life Evol.
570 Biosph., 25:565-589.

571 SAS. 2015. JMP: The Statistical Discovery Software. Cary, NC, SAS Institute, Cary NC.

572 Sever S, Rogers K, Rogers MJ, Carter CW, Jr., Söll D. 1996. *Escherichia coli*
573 tryptophanyl-tRNA synthetase mutants selected for tryptophan auxotrophy implicate the dimer
574 interface in optimizing amino acid binding. Biochemistry, 35:32-40.

575 Shiba K. 2005. Glycyl-tRNA Synthetases. In: Ibba M, Francklyn C, Cusack S editors.
576 The Aminoacyl-tRNA Synthetases Austin, Landes Bioscience.

577 Shore J, Holland BR, Sumner JG, Nieselt K, Wills PR. 2019. The Ancient Operational
578 Code is Embedded in the Amino Acid Substitution Matrix and aaRS Phylogenies. J. Mol. Evol.,
579 88:136–150.

580 Sobolevsky Y, Frenkel ZM, Trifonov EN. 2007. Combinations of Ancestral Modules in
581 Proteins. J. Mol. Evol. , 65:640–650.

582 Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein
583 homology detection and structure prediction. Nucl. Acids Res., 33:W244–W248

584 Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of
585 life. Nat. Rev. Gen., 16:472

586 Thompson JD, Higgins DG, Gibson TJ. 1994. Clustal W: improving the sensitivity of
587 progressive multiple sequence alignment through sequence weighting, positions-specific gap
588 penalties and weight matrix choice. Nucleic Acids Res., 22:4673-4680.

589 Trifonov EN. 1995. Segmented Structure of Protein Sequences and Early Evolution of
590 Genome by Combinatorial Fusion of DNA Elements. J. Mol, Evol. , 40:337-342.

591 Tyson Nd. 2013. "Just to settle it once and for all: Which came first the Chicken or the
592 Egg? The Egg – laid by a bird that was not a Chicken. 2013-01-28:Tweet.

593 Webb B, Sali A. 2016. Comparative Protein Structure Modeling Using Modeller. Current
594 Protocols in Bioinformatics, John Wiley & Sons, Inc.,, p. 5.6.1-5.6.37.

595 Weinreb V, Li L, Carter CW, Jr. 2012. A Master Switch Couples Mg^{2+}-Assisted
596 Catalysis to Domain Motion in B. stearothermophilus Tryptophanyl-tRNA Synthetase. Structure,
597 20:128-138.

598 Weinreb V, Li L, Chandrasekaran SN, Koehl P, Delarue M, Carter CW, Jr 2014.
599 Enhanced Amino Acid Selection in Fully-Evolved Tryptophanyl-tRNA Synthetase, Relative to
600 its Urzyme, Requires Domain Movement Sensed by the D1 Switch, a Remote, Dynamic Packing
601 Motif J Biol Chem 289:4367-4376.

602 Whelan S, Goldman N. 2001. A General Empirical Model of Protein Evolution Derived
603 from Multiple Protein Families Using a Maximum-Likelihood Approach. Molecular Biology and
604 Evolution, 18:691-699.

605 Wills PR. 2014. Genetic Information, Physical Interpreters and Thermodynamics; The
606 Material-Informatic Basis of Biosemiosis Biosemiotics 7:141–165.

607 Wills PR, Carter CW, Jr. 2018. Insuperable problems of an initial genetic code emerging
608 from an RNA World. BioSystems, 164:155-166.

609 Woese CR, Olsen GJ, Ibba M, Soll D. 2000. Aminoacyl-tRNA Synthetases, the Genetic
610 Code, and the Evolutionary Process. Microbiol. Mol. Biol. Rev., 64:202–236.

611 Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999. Evolution of Aminoacyl-tRNA
612 Synthetases—Analysis of Unique Domain Architectures and Phylogenetic Trees Reveals a
613 Complex History of Horizontal Gene Transfer Events. Genome Research, 9:689–710.

614 Wolfenden R, Lewis CA, Yuan Y, Carter CW, Jr. 2015. Temperature dependence of
615 amino acid hydrophobicities. Proc. Nat. Acad. Sci. USA, 112 7484-7488.

616

Table I. Regression model for $\langle Q \rangle_{\text{calc}}$ as a function of inclusion in CP1. R^2 for this model is 0.17 for the 104 columns in the MSA. Regression using the (alternate) design matrix in Table II yields a similar result, for which $R^2 = 0.32$ with $P = 0.007$.

Term	β	Std Error	t Ratio	Prob>|t|
Intercept	52.45	1.80	29.22	<.0001
CP1	-18.77	4.09	-4.59	1.3 E-5

Table II. Design matrix for regression analysis. $\langle Q \rangle$ is the mean Conservation Quality averaged over a given partition of the MSA. $\langle S \rangle$ is the mean support value for all ten amino acids defined in Methods from clade congruencies. A, B, C, and CP1 are independent variables, defined from the proportional representation of residues (approximately as multiples of 10) in each of the segments for each row, in accordance with Figure 1C. LG encodes which amino acid substitution matrix, LG (1) or WAG (0), was used to construct trees. NUMB is the number of amino acids in the alignment.

Subset	$\langle S \rangle$	$\langle Q \rangle$	A	B	C	CP1	LG	NUMB
Full	0.90	48.8	4	4	1	2	0	103
Urzyme	0.87	52.4	4	4	1	0	0	83
Protoz	0.85	52.3	4	0	0	0	0	46
CP1	0.34	33.7	0	0	0	2	0	20
Urz_0	0.76	50.1	1	1	0	0	0	20
Urz_1	0.71	57.8	0	1	0.5	0	0	20
Urz_2	0.72	55.0	2	0	0	0	0	20
Urz_3	0.62	45.6	0	4	0	0	0	20
Urz_4	0.82	58.5	2	1	0.5	0	0	20
Urz_5	0.78	45.6	4	0	0	0	0	20
Urz_6	0.62	56.6	0	2	1	0	0	20
Full	0.86	na	4	4	1	2	1	103
Urzyme	0.82	na	4	4	1	0	1	83
Protoz	0.85	na	4	0	0	0	1	46

CP1	0.35	na	0	0	0	2	1	20
Urz_0	0.73	na	1	1	0	0	1	20
Urz_1	0.68	na	0	1	0.5	0	1	20
Urz_2	0.70	na	2	0	0	0	1	20
Urz_3	0.62	na	0	4	0	0	1	20
Urz_4	0.77	na	2	1	0.5	0	1	20
Urz_5	0.80	na	4	0	0	0	1	20
Urz_6	nd	na	0	2	1	0	1	20

630 Table III. Regression model for $\langle S \rangle$. R^2 was 0.94 for 20 observations, and the F-ratio was 89 with

631 a P-value <0.0001.

632

Term	β	Std Error	t Ratio	Prob>|t|
Intercept	0.63	0.01	44.45	<.0001
A	0.06	0.01	11.63	1.3E-10
CP1	-0.07	0.01	-6.1	1.8E-8
A*CP1	0.05	0.01	7.86	4.6E-7

633

634

635 ## Figure Legends

636

637 Figure 1. Structural biology of the hypothesis and the organization of data adduced as
638 evidence. A. The hypothesis. The Class I Urzyme has an ancestry distinct from that of the internal
639 CP1 insertion, which together make up the Class I catalytic domain. The activated
640 aminoacyladenylate is also shown as sticks. Cartoons were prepared for the tryptophanyl-tRNA
641 synthetase, the smallest of the Class I aaRS. The entire structural array appears intact in each of
642 the remaining nine Class I aaRS. The anticodon-binding domains are substantially more
643 idiosyncratic, and are not shown in the scheme. Previous work established that the CP1 insertion
644 has little impact on the enzymatic properties of the Urzyme (Li, L. and Carter, C.W., Jr 2013).
645 Colors match those in the schematic (C). The amino terminal crossover connection (blue) is the
646 protozyme. B. Overlapping of the portions of Class I and II aaRS as envisioned in the ancestral
647 bidirectional gene by Rodin and Ohno (Rodin, S.N. and Ohno, S. 1995) coincide with the
648 respective Class I and II Urzymes (Pham, Y., Li, L., et al. 2007, Carter, C.W., Jr., Li, L., et al.
649 2014). Vertical lines denote ancestral base-pairing. Grey segments were presumably more recent
650 additions. In particular, insertion of CP1 is incompatible with bidirectional coding. C. Schematic
651 of the location of CP1 between two roughly equal fragments of the Urzyme, colored to allow
652 identification of structural fragments in A. Intensely colored segments are highly conserved
653 secondary structures in all 10 Class I aaRS and compose the Class I scaffold. Subclass I A enzymes
654 contain one or more additional insertions (red). C. Sampled alignments consisting of different
655 segments totaling 20 amino acids, selected as described in Methods from the three Urzyme
656 segments to have the same number of amino acids as the CP1 scaffold alignments.

657 Figure 2. Mean conservation quality scores, $\langle Q \rangle$ (Thompson, J.D., Higgins, D.G., et al.
658 1994), for the four segments of the Class I scaffold alignment. Class I signature peptides (Carter,
659 C.W.J. 1993) are in parentheses. Error bars show the standard error of the mean over all positions
660 within the segment. The numbers of amino acids in each segment are given for each histogram.

661 Figure 3. DensiTree representations of the Urzyme and CP1 segments of the Class I
662 structural scaffold. Each aaRS is monophyletic in the Urzyme alignment, whereas several of the
663 clades in the CP1 alignment, highlighted in color, have separated ancestries.

664 Figure 4. Regression model for $\langle S \rangle_i$. The R^2 and P value of the model's F-ratio are shown
665 under the X-axis. Coefficients and their statistics for the model are given in Table II. Individual
666 $\langle S \rangle$ values are labeled to enable comparison with Fig. 1C.

667 Figure 5. Coupling effects entailed in aaRS evolution. Progressive increases in the alphabet
668 size enabled by introducing new aaRS into the context of the ancestral bidirectional gene (grey
669 background and dashed connecting lines) simultaneously introduce new enhanced specificity and
670 fundamental changes in selection pressure. The scale of each aaRS quasispecies is indicated by the
671 diameter of the tube associated with each alphabet as their evolution proceeds upward. Different
672 colored backgrounds signify altered selection pressures that apply to all extant aaRS at a given
673 stage of the coding alphabet. Increasing cardinality of the alphabet induces (i) sequence space
674 inflation as a greater number of distinguishable sequences can be specified; and (ii) progressive
675 restriction in the diameters of the quasispecies as they approach fully coded sequences in the final
676 20-letter alphabet. Bold face entries (CP1 and ID) denote qualitative changes in aaRS architecture
677 shown experimentally to enrich the possible mechanisms for specificity. Note that this scheme and
678 that in Figure 7 represent evolutionary events before the LUCA.

679 Figure 6. Structural relationship between the TrpRS Urzyme, CP1, and the anticodon-
680 binding domain (ABD). The CP1 motif forms an annulus that constrains the motions of the
681 specificity determining helix (sand) and the Protozyme (yellow), thereby imposing constraints on
682 the effective size of the amino acid binding pocket. As described in the text, there is experimental
683 evidence that these constraints contribute to the ability of full-length TrpRS to reject tyrosine in
684 the transition state complex for tryptophan activation (Li, L. and Carter, C.W., Jr 2013, Weinreb,
685 V., Li, L., et al. 2014).

686 Figure 7. Modified Class I aaRS phylogeny consistent with bidirectional coding ancestry
687 in a peptide/RNA world. Adapted from Fig. 12-2A of ref (Koonin, E.V. 2011). Green circles
688 represent key stages in the emergence of genetic coding, beginning with ATP-dependent amino
689 acid activation by the peptides coded by the bidirectional protozyme gene (Martinez, L., Jimenez-
690 Rodriguez, M., et al. 2015). Magenta shading indicates the progression illustrated in detail in Fig.
691 5.

692

693

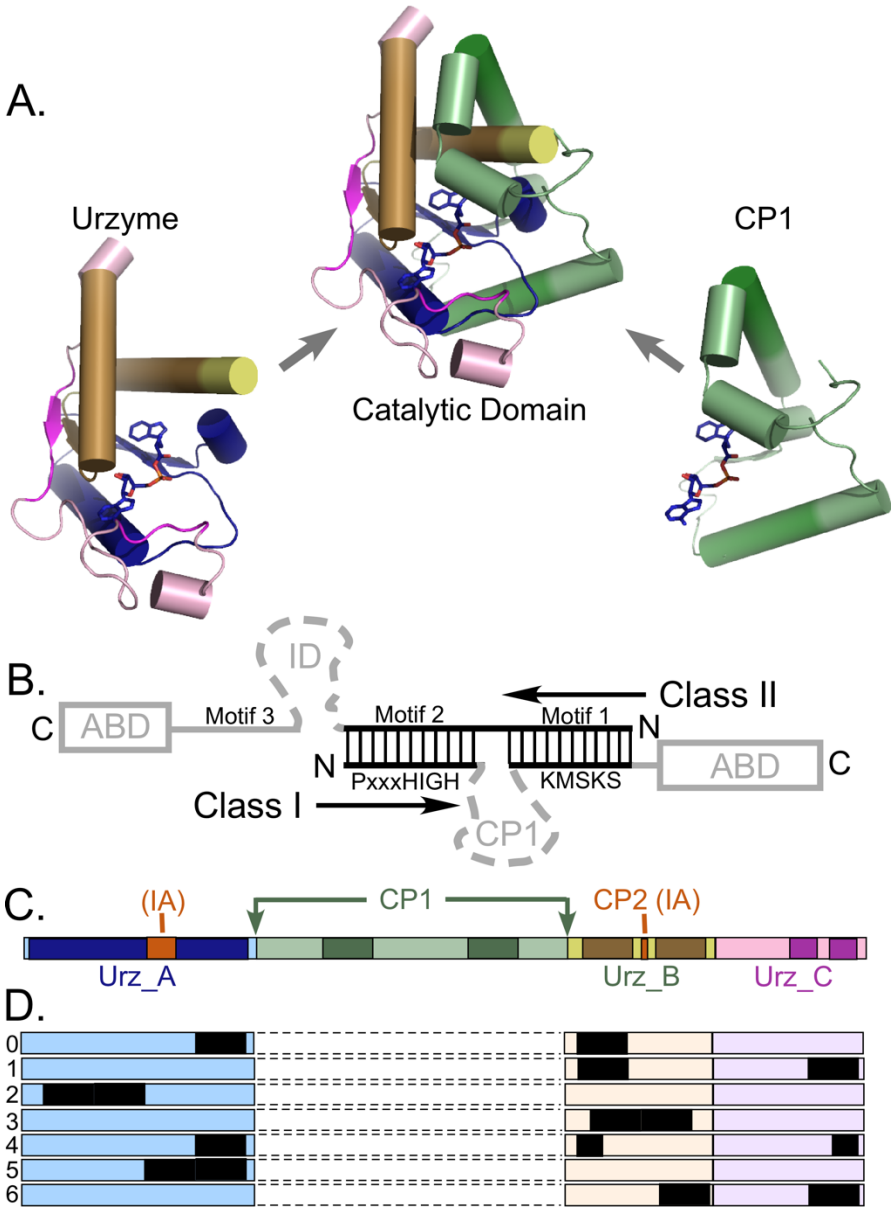

A.
Urzyme
Catalytic Domain
CP1
B.
ID
C
ABD
Motif 3
Motif 2
Motif 1
Class II
N
N
PxxxHIGH
KMSKS
ABD
C
Class I
CP1
C.
(IA)
CP1
CP2 (IA)
Urz_A
Urz_B
Urz_C
D.
0
1
2
3
4
5
6

Figure 1.

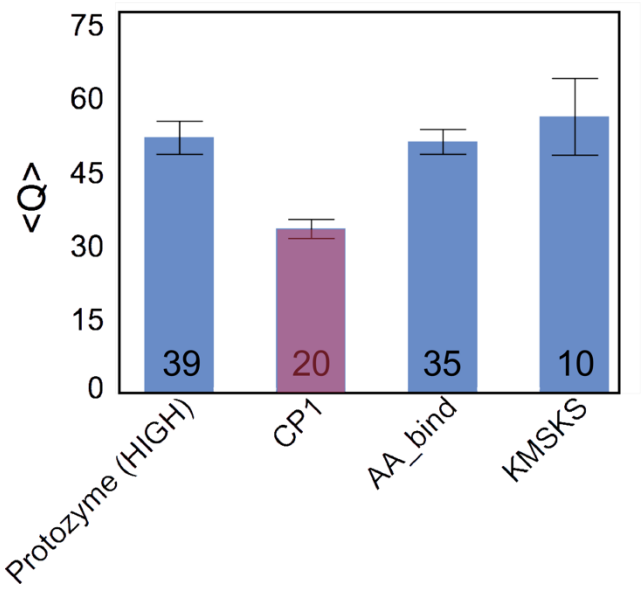


Figure 2.

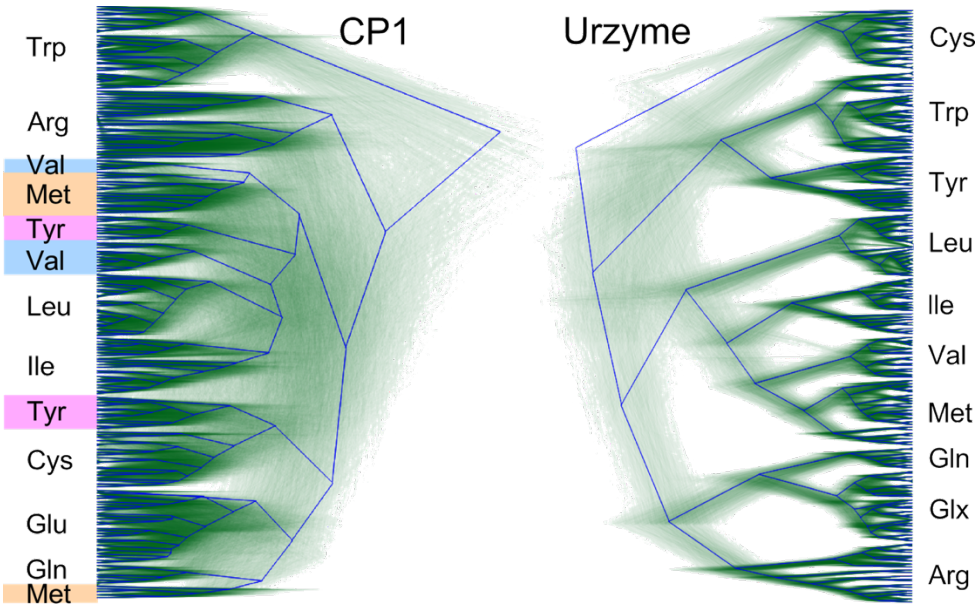

CP1
Urzyme
Trp
Arg
Val
Met
Tyr
Val
Leu
Ile
Tyr
Cys
Glu
Gln
Met
Cys
Trp
Tyr
Leu
Ile
Val
Met
Gln
Glx
Arg

Figure 3.

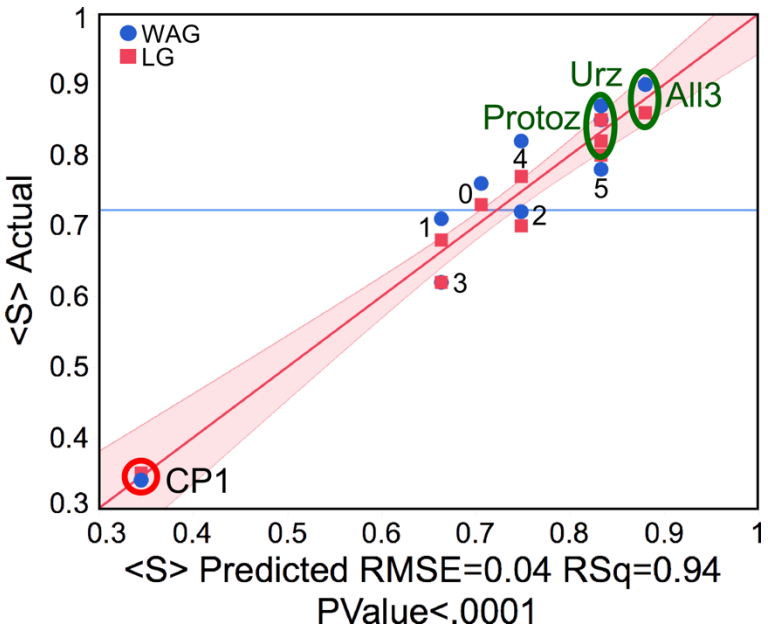

WAG
LG
Urz
All3
Protoz
CP1
0
1
2
3
4
5
<S> Actual
<S> Predicted RMSE=0.04 RSq=0.94
PValue<.0001
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 4.

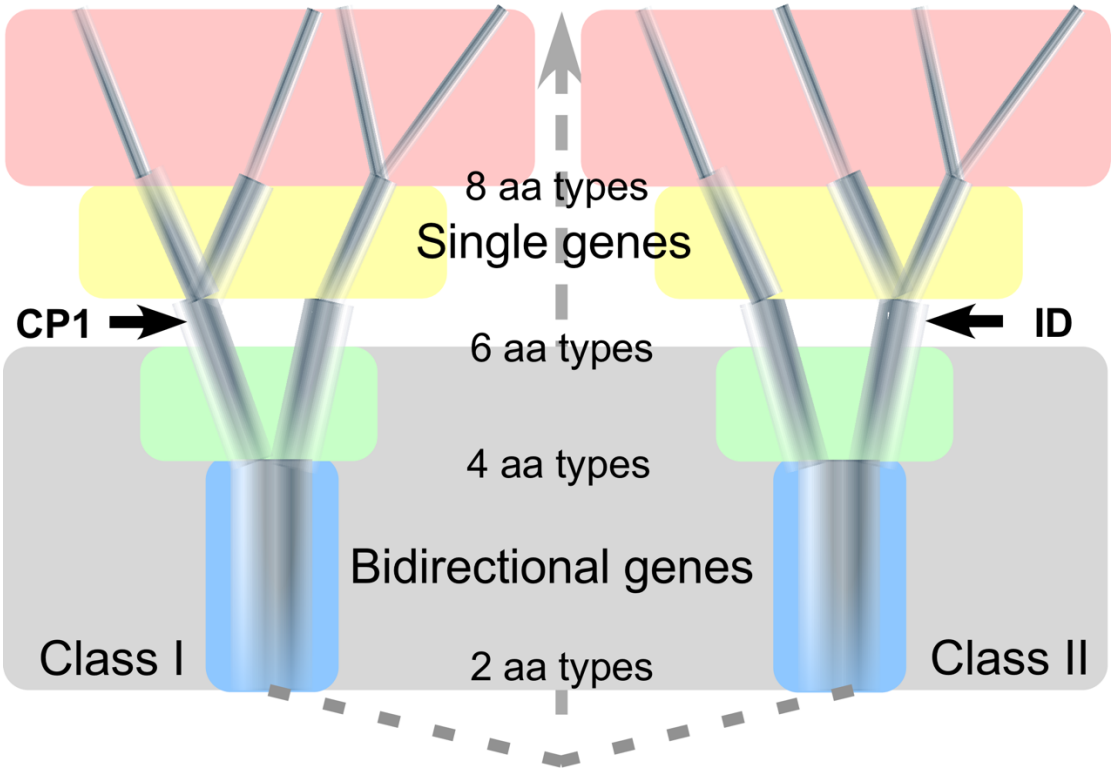
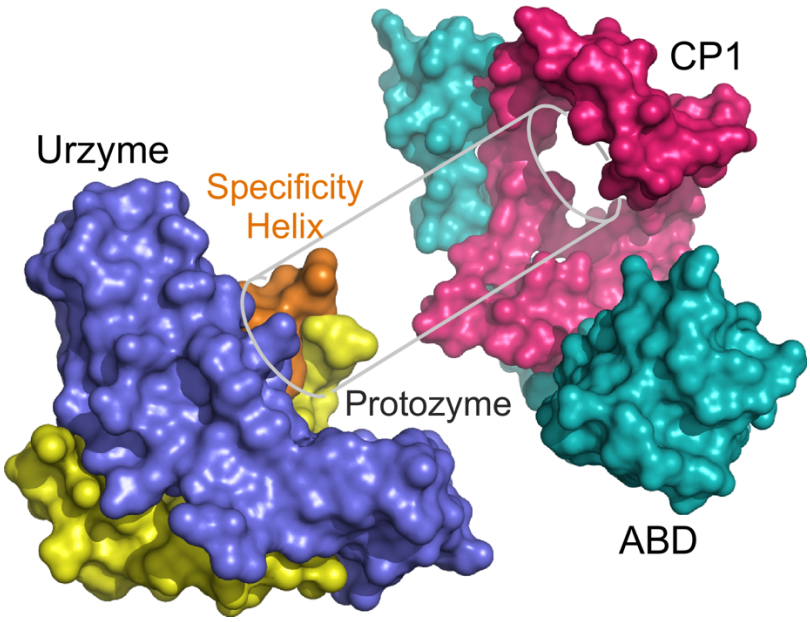

8 aa types
Single genes
CP1
ID
6 aa types
4 aa types
Bidirectional genes
Class I
2 aa types
Class II

Figure 5.



712

713 Figure 6.

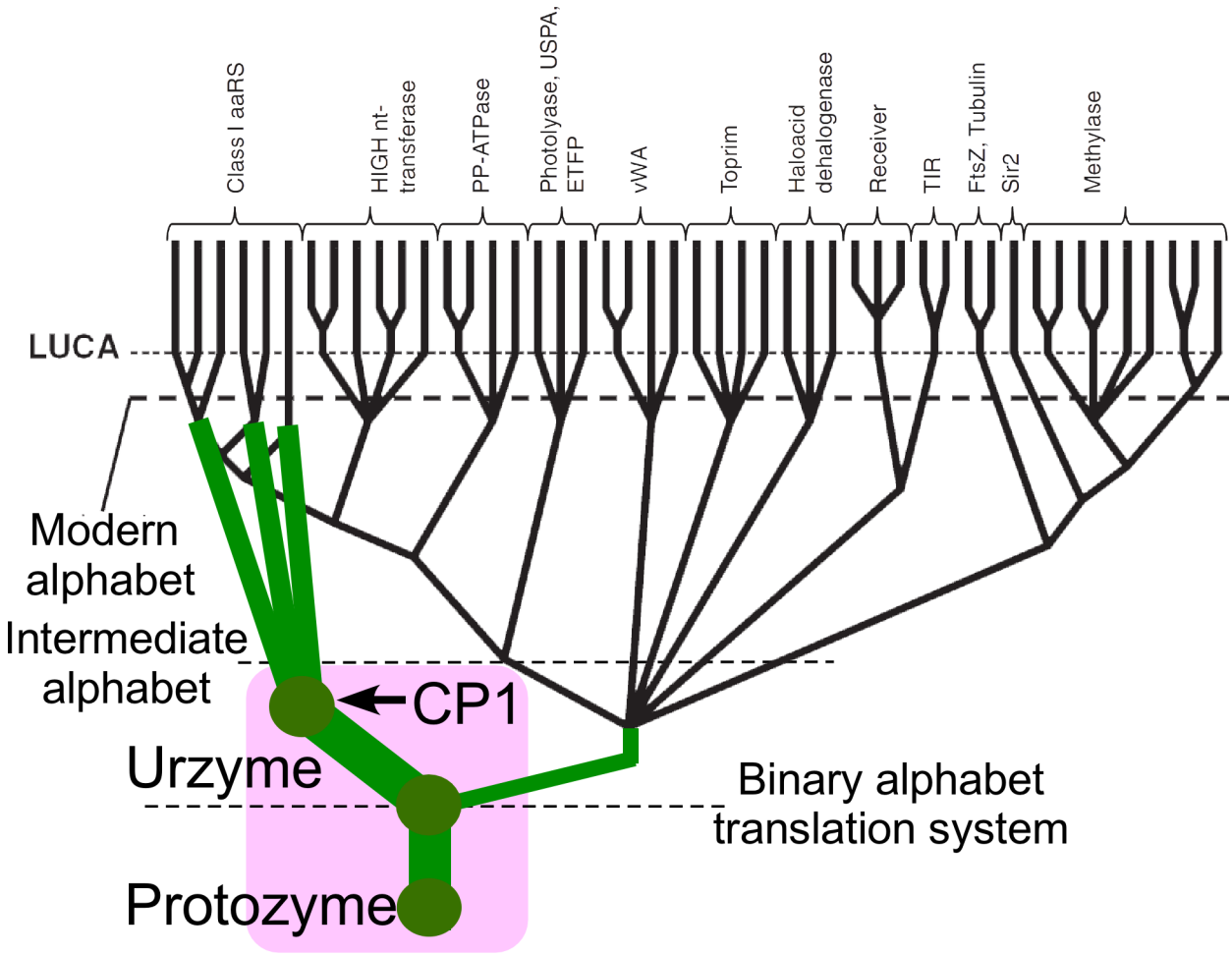

Class I aaRS
HIGH nt-transferase
PP-ATPase
Photolyase, USPA, ETFP
vWA
Toprim
Haloacid dehalogenase
Receiver
TIR
FtsZ, Tubulin
Sir2
Methylase
LUCA
Modern alphabet
Intermediate alphabet
CP1
Urzyme
Binary alphabet translation system
Protozyme

Figure 7