aaRS evolution

Uranya: I did the research you requested regarding what is currently known about the aaRS's. (*a*mino*a*cyl-t*R*NA *s*ynthetases)

Reynard: Thank you. You have a better chance to hear about good recent sources than I do. Tell me about what you have learned.

U: First of all, a good recent source is the book *The Aminoacyl-tRNA Synthetases*, (Eds. Ibba, Francklyn and Cusack, Landes Bioscience, 2005). It has chapters about each amino acid aaRS as well as other chapters. Secondly there is a series of papers by Andrei Rodin and Sergei Rodin (and earlier with Susumu Ohno) that show codon patterns associated with the two classes.

R: Are these Rodins the Rodin brothers?

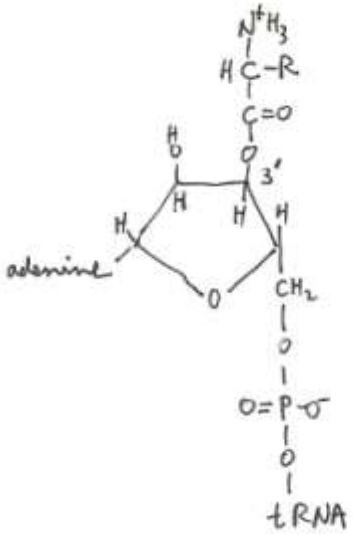
U: Andrei is Andrei Sergeyevich so I suspect he is the son, not the brother, of Sergei.

R: I too have looked for patterns relating amino acids to their aaRS's. The results surprised me. The link [polymer biosynthesis] contains the excerpt below.

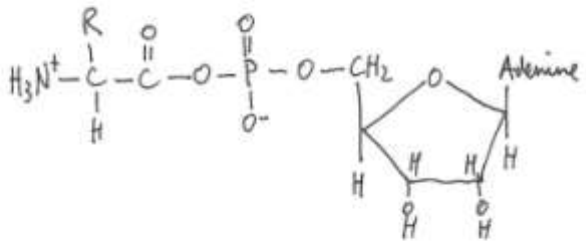
The tRNA's contain anti-codons at one end that read the codons by anti-parallel base pairing. At the other end of the tRNA is the amino acid attachment site on the tRNA 3' terminus. The aaRS's catalyze the attachment of the amino acids to the tRNA's at the expense of ATP.

$$tRNA + aa + ATP \rightarrow aa - tRNA + AMP + PP$$

in which AMP denotes adenine monophosphate and PP denotes pyrophosphate. The linkage between the carboxyl group of the amino acid and the 3' -OH group of the terminal ribose of the tRNA is a high energy ester without either P or S. This is uniquely unusual.



The intermediate is the aminoacyl adenylate, aa-AMP



The aaRS recognizes and binds the amino acid and the cognate tRNA, that for amino acid aa is labelled $tRNA^{aa}$. It also binds ATP and forms the intermediates aa-AMP and PP. The PP is released and subsequently hydrolysed, and the aa-AMP reacts with the bound $tRNA^{aa}$ to form aa-$tRNA^{aa}$. All of this occurs enzyme bound on the aaRS, but free aminoacyl adenylates can be prepared in the laboratory as free chemical species.

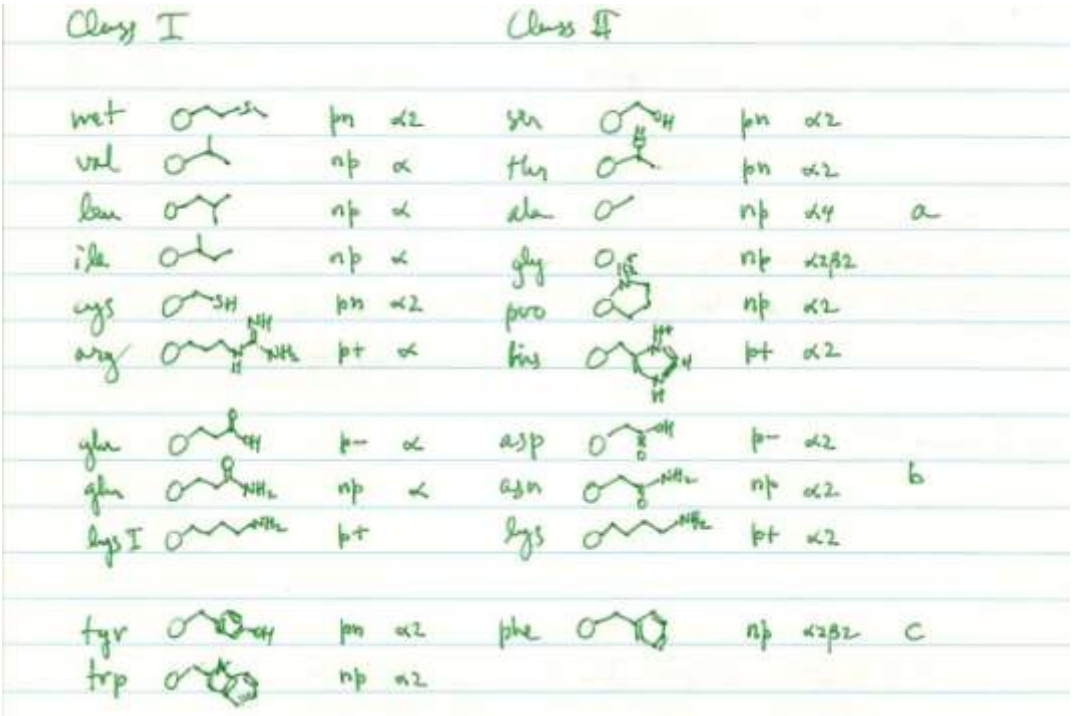
There is at least one aaRS for each of the 20 amino acids. The structural similarity of all tRNAs suggests that the aaRS's should all be closely related since they perform the same tasks. This is not so, they are a very diverse group of enzymes. Four types of subunit structure are found: α, α_2, α_4 and $\alpha_2\beta_2$. This means that some have just one subunit of type α, the α's. Others have multiple subunits of type α, the α_2's and α_4's. Some have two copies of two types of subunits, the $\alpha_2\beta_2$'s. For different amino acids, the subunits involved are of different sequence and structure, even though each is referred to as, say, α. Subunit sizes range from 334 to 1112 amino acids residues. Little sequence similarity exists from one aaRS species to another. Two major classes of aaRS's are found: class I and class II. Each has the same 10 members in all organisms. Class I aaRS's require anti-codon recognition to aminoacylate their cognate tRNA's. Class II aaRS's do not interact with the anti-codons. Class I aaRS's aminoacylate their tRNA's 3' terminal 2'-OH group whereas class II aaRS's charge the 3' terminal 3'-OH group instead.

Class I	Class II	p+ ; polar +
glu p-	asp p-	p- ; polar -
arg p+	his p+	pn ; polar neutral
cys pn	lys p+	np ; non-polar
met pn	ser pn	
tyr pn	thr pn	
gln np	asn np	
ile np	gly np	

leu np	ala np	
val np	pro np	
trp np	phe np	
lys* p+		

The extra lysRS of Class I simply balances the distribution of amino acids of like types between the two classes, i.e. each class has one negatively charged group, two positively charged groups, five non-polar groups and 3 (I) or 2 (II) polar neutral groups. Amino acid functions are distributed fairly between the two classes.

R: Nevertheless, a bit of a pattern emerges when the aaRS sub-classes are identified (Schimmel and Beebe, "From the RNA World to the theatre of proteins" in *The RNA World* Eds. Gesteland, Cech and Atkins, Cold Spring Harbor Laboratory Press, 2006). I sketch their results below in a table that also shows the aa residues and lists both the aaRS sub-unit conformations and the amino acid functionalities (polar plus, p+; polar negative, p-; polar neutral, pn; and non-polar, np).



The sub-classes are related in amino acid sequence. Ic and IIc exclusively hold the aromatic aa's, trp, tyr and phe. Ib and IIb respectively hold glu & gln and asp & asn, as well as lys.

U: This really shows how little relation there is to aa type and aaRS class. Perhaps the fact that the two classes yield aa sets that each contain all functionality types will prove significant. This table you have drawn is the modern complex system.

As if to emphasize this with its structure, the smallest aa, glycine, has the most complex subunit structure.

R: Perhaps the relationship between aa's and aaRS class is not illuminated by this table. However, it does suggest some evolutionary possibilities. For example, sub-classes Ic and IIc contain the only aromatic aa residues. Their complexity would suggest that they are late acquisitions in the evolution of the code. Typically there are only two phe codons and only two tyr codons, consistent with the idea that these codons were used for other amino acids (or stops) earlier and then recruited for phe and tyr later. Trp has only one codon typically (there are a few exceptions that have two codons for trp [Polymer biosynthesis]) and this is also consistent with it's putative late recruitment. The cases of sub-classes Ib and IIb are even more interesting. Both asn and gln can be made indirectly from asp-tRNA^{asn} and glu-tRNA^{gln} respectively by most archaebacteria and eubacteria. For example, a tRNA-dependent glu-tRNA^{gln} amidotransferase (glu-AdT) amidates the complex to form gln-tRNA^{gln}. A parallel process produces asn-tRNA^{asn}. These two step mechanisms are thought to be the primitive pathways. The lys aaRS story is more complex. Most bacteria and all eukaryotes contain class II lysRS. Most archaea and a few bacteria contain class I lysRS. This is the only aa for which there are aaRS's of each class. While the structural properties of all class Ib aaRS's are similar, and the same is true for all class IIb aaRS's, the two lysRS's show little sequence similarity, and yet still function for lys.

We have implicitly assumed that a mechanism used by a so-called primitive organism, such as an archaea, is itself primitive. Of course the more primitive organisms have had longer to make evolutionary refinements and their mechanisms in some cases may actually be more modern. However, for glu/gln and asp/asn it does seem natural to suspect that glu and asp were the early aa's and gln and asn came along later. Both glu and asp are common products in abiotic aa syntheses and important in the synthesis of proteinoids.

U: There is another case of indirect charging of tRNA that bears on the issue you just raised. That is the case of indirect charging of cys-tRNA^{cys}. In the first step O-phosphoseryl-RS (sepRS) charges tRNA^{cys} with O-phosphoserine (sep) a precursor of cysteine in modern metabolism. This yields sep-tRNA^{cys}. In the second step, sep-tRNA:cys-tRNA synthase (sepcysS) converts sep to cys in a tRNA-dependent reaction yielding cys-tRNA^{cys}. This rare indirect pathway is found in *Euryarchaea* which suggests it may be of recent origin rather than more primitive than the direct cys charging process. P. O'Donoghue et al. (PNAS, **102**, 19003-19008, 2005) argue that for the indirect modes used for asn and gln the evidence is strongly in favor of the indirect mode being more primitive, whereas sepRS and cysRS appear to have evolved in parallel.

R: All of this goes to show that once again it is unsafe to make absolute statements about aaRS's, such as *there are 20 aaRS's, one for each amino acid*. Clearly, before asnRS and glnRS evolved there were perhaps only 18. In addition there are

both sepRS and cysRS. In the case of selenocystyl-tRNA^{sec} (sec-tRNA^{sec}), the the serRS charges tRNA^{sec} with serine, just as it would charge tRNA^{ser} with serine, and then in a two step process sec-tRNA^{sec} is made from the ser-tRNA^{sec}. This is a case of one aaRS leading to two differently charged tRNAs. For the highly degenerate aa codons, one aaRS charges several different tRNA's with the same aa. So this is different from the sec-tRNA^{sec} case. Selenocysteine is called the 21^{rst} amino acid. There is also a 22nd amino acid, pyrrolysine. Each is encoded by what is usually a stop codon, UGA and UAG respectively. These are no doubt modern additions and need not play a role in the primitive mechanisms. Nevertheless, one must be careful to not be too dogmatic about the status of aaRS's, codons and coded amino acids.

U: I have found the following ideas in the Rodins' papers. I make a table of codon complementarity (after Andrei Rodin, Sergei Rodin [and Susumu Ohno]). This complementarity is an antiparallel base pairing of a codon and another codon, the complementary codon. Thus the complementary codon is similar in sequence to the anti-codon of the tRNA but without wobble rules. The codon's *third* base is degenerate, either completely or partially, whereas its complement is the *first* base of the complementary codon, which is absolutely non-degenerate. By arranging the table according to the second base, the complementarity of the second bases is highlighted.

1	2	3	1	2	3
	U			A	
U	Phe	U	A	Lys	A
U	Phe	C	G	Glu	A
U	Leu	A	U	Stop	A
U	Leu	G	C	Gln	A
C	Leu	U	A	Lys	G
C	Leu	C	G	Glu	G
C	Leu	A	U	Stop	G
C	Leu	G	C	Gln	G
A	Ile	U	A	Asn	U
A	Ile	C	G	Asp	U
A	Ile	A	U	Tyr	U
A	Met	G	C	His	U
G	Val	U	A	Asn	C
G	Val	C	G	Asp	C
G	Val	A	U	Tyr	C
G	Val	G	C	His	C
1	2	3	1	2	3
	G			C	
U	Cys	U	A	Thr	A
U	Cys	C	G	Ala	A
U	Stop	A	U	Ser	A
U	Trp	G	C	Pro	A
C	Arg	U	A	Thr	G
C	Arg	C	G	Ala	G
C	Arg	A	U	Ser	G
C	Arg	G	C	Pro	G
A	Ser	U	A	Thr	U
A	Ser	C	G	Ala	U
A	Arg	A	U	Ser	U
A	Arg	G	C	Pro	U
G	Gly	U	A	Thr	C
G	Gly	C	G	Ala	C
G	Gly	A	U	Ser	C
G	Gly	G	C	Pro	C

The green entries are Class I and the red entries are Class II. Second base complementarity does a pretty good job of separating the amino acids into the two aaRS classes. If we change the Class definition *from* depending on the structure of the ATP binding site (Class I Rossmann fold or Class II 7-stranded β-sheet) and whether attachment of the amino acid is to the 2'-OH (Class I) or to the 3'-OH (Class II) of ribose *to* depending on whether the aaRS approaches the tRNA from the major or minor groove side of the tRNA, that leads to exchange of classification of Phe and Tyr. Note that pheRS attaches Phe to the 2'-OH and

tyrRS attaches Tyr to the 3'-OH so it is only the ATP binding site that is "wrong" in each case. This yields

1	2	3	1	2	3
	U			A	
U	Phe	U	A	Lys	A
U	Phe	C	G	Glu	A
U	Leu	A	U	Stop	A
U	Leu	G	C	Gln	A
C	Leu	U	A	Lys	G
C	Leu	C	G	Glu	G
C	Leu	A	U	Stop	G
C	Leu	G	C	Gln	G
A	Ile	U	A	Asn	U
A	Ile	C	G	Asp	U
A	Ile	A	U	Tyr	U
A	Met	G	C	His	U
G	Val	U	A	Asn	C
G	Val	C	G	Asp	C
G	Val	A	U	Tyr	C
G	Val	G	C	His	C
1	2	3	1	2	3
	G			C	
U	Cys	U	A	Thr	A
U	Cys	C	G	Ala	A
U	Stop	A	U	Ser	A
U	Trp	G	C	Pro	A
C	Arg	U	A	Thr	G
C	Arg	C	G	Ala	G
C	Arg	A	U	Ser	G
C	Arg	G	C	Pro	G
A	Ser	U	A	Thr	U
A	Ser	C	G	Ala	U
A	Arg	A	U	Ser	U
A	Arg	G	C	Pro	U
G	Gly	U	A	Thr	C
G	Gly	C	G	Ala	C
G	Gly	A	U	Ser	C
G	Gly	G	C	Pro	C

Two more changes can be made. In some archaebacteria there is a Class I aaRS for Lys. Let us suppose this is the primitive state and make the change from Class II to Class I in the above table. Moreover in some mitochondria AGG and AGA code for Gly or Ser, both Class II, instead of Arg that is in Class I. If we also

make this change in the above table then we get (I made stop green to emphasize the beauty in the Rodins' discovery).

1	2	3	1	2	3
	U			A	
U	Phe	U	A	Lys	A
U	Phe	C	G	Glu	A
U	Leu	A	U	Stop	A
U	Leu	G	C	Gln	A
C	Leu	U	A	Lys	G
C	Leu	C	G	Glu	G
C	Leu	A	U	Stop	G
C	Leu	G	C	Gln	G
A	Ile	U	A	Asn	U
A	Ile	C	G	Asp	U
A	Ile	A	U	Tyr	U
A	Met	G	C	His	U
G	Val	U	A	Asn	C
G	Val	C	G	Asp	C
G	Val	A	U	Tyr	C
G	Val	G	C	His	C
1	2	3	1	2	3
	G			C	
U	Cys	U	A	Thr	A
U	Cys	C	G	Ala	A
U	Stop	A	U	Ser	A
U	Trp	G	C	Pro	A
C	Arg	U	A	Thr	G
C	Arg	C	G	Ala	G
C	Arg	A	U	Ser	G
C	Arg	G	C	Pro	G
A	Ser	U	A	Thr	U
A	Ser	C	G	Ala	U
A	G/S	A	U	Ser	U
A	G/S	G	C	Pro	U
G	Gly	U	A	Thr	C
G	Gly	C	G	Ala	C
G	Gly	A	U	Ser	C
G	Gly	G	C	Pro	C

This information can be rendered in an equivalent manner by interchanging the positions of the second base. That is, instead of writing them in the order UA and GC we write them in the order AU and CG:

1	2	3	1	2	3
	A			U	
U	Tyr	U	A	Ile	A
U	Tyr	C	G	Val	A
U	Stop	A	U	Leu	A
U	Stop	G	C	Leu	A
C	His	U	A	Met	G
C	His	C	G	Val	G
C	Gln	A	U	Leu	G
C	Gln	G	C	Leu	G
A	Asn	U	A	Ile	U
A	Asn	C	G	Val	U
A	Lys	A	U	Phe	U
A	Lys	G	C	Leu	U
G	Asp	U	A	Ile	C
G	Asp	C	G	Val	C
G	Glu	A	U	Phe	C
G	Glu	G	C	Leu	C
1	2	3	1	2	3
	C			G	
U	Ser	U	A	S/G	A
U	Ser	C	G	Gly	A
U	Ser	A	U	Stop	A
U	Ser	G	C	Arg	A
C	Pro	U	A	S/G	G
C	Pro	C	G	Gly	G
C	Pro	A	U	Trp	G
C	Pro	G	C	Arg	G
A	Thr	U	A	Ser	U
A	Thr	C	G	Gly	U
A	Thr	A	U	Cys	U
A	Thr	G	C	Arg	U
G	Ala	U	A	Ser	C
G	Ala	C	G	Gly	C
G	Ala	A	U	Cys	C
G	Ala	G	C	Arg	C

If each stop codon is considered green, a simple pattern has emerged: complementary codons either use the same type of aaRS or the opposite type in a regular alternation by pairs. The pattern can be described in terms of pyrimidines (Y) and purines (R). Look at the left hand sides with second base A or C. Consider the codons with second and third bases of the form RY, RR, YY and YR. The combinations RY and YR always have opposite class aaRS's for the complementary codon amino acids, and the combinations RR and YY always have same class aaRS's for the complementary codon amino acids. The same

conclusion arises from looking at the right hand sides with second base U or G. Now, however, it is the first and second bases in the form RY, YY, RR and YR that matter. Again, RR and YY always have the same class aaRS's for the amino acids of complementary codons and RY and YR always have opposite class aaRS's. These rules are just the complements of the previous rules and are built into the table.

This analysis provides some structure to the pattern of aaRS's and codons. However, it says nothing about the amino acids. There does not appear to be a comparable pattern for aaRS's and their corresponding amino acids as we noted earlier. Any model regarding origins must explain how these patterns arose.

R: Very interesting ! Can the *primitive RNA translator* [Part 6] evolve into a mechanism with the properties just described, involving ribosomes, tRNA's and aaRS's ? That is a lot of complexity to evolve. Even if we restrict ourselves to Class Ia and Class IIa sub-classes, twelve aaRS's are required all together. Can these emerge at once, together ? I think genome consolidation is the issue. Joining of genome pieces, the *genes*, creates combined functionalities after the *primitve RNA translator* works. Longer polypeptides can be made from longer RNA's. Any feedback functionality that increases either membrane growth and replication or RNA transcription and translation, will be selected because a population of microspheres carrying this gene will arise. Replication of the gene keeps pace with the microsphere division rate. Robust Brownian motion segregates the gene copies into the two daughter microspheres so that each has the same genome.

Let me introduce new terminology for the primitive components of this evolving model. Earlier I have used *proto-* to connote a more primordial form or state. From now on I will use *ur-*. I do this in your honor, Uranya. For example, the microspheres with growing membrane that can divide will be called ur-cells. The collection of RNA genes, is now the ur-genome. One such ur-gene is surely the ur-ligase based on arginine and CGN codons. As genes are joined the longer RNA's eventually are so long that the *primitive RNA translator* mechanism doesn't work reliably from one end to the other. Evolution of a linear ur-mRNA reading mechanism may need to occur. The same RNA molecule that serves as an ur-gene in the *primitive RNA translator* mechanism serves as the ur-mRNA of the linear mechanism. No doubt some sort of ur-ribosome is needed to support the linear reading of the ur-mRNA.

These thoughts address the question of "why did the more complex molecular machinery evolve at all ?"

U: So you want a model that explains the evolution of ur-tRNA's and ur-aaRS's ?? I can see the ur-tRNA's in the ligased hexameric-polynucleotide populations that you discussed at the end of [Part 4]. But I don't yet see how the ur-aaRS's arise.

R: Origin of the aaRS's is the core problem. That it is perhaps a difficult problem is not equivalent to [IC]. The appeal **should be to** *discover the answers*, **not to** *there aren't any plausible answers; the existence of the aaRS's is miraculous*.
As an extant open problem, it is the *Mysterium Tremendum.*