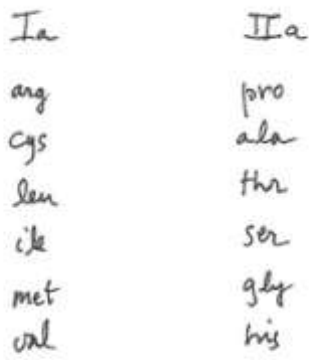
Evolution of the genetic code

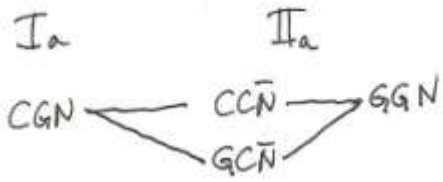
Uranya: I have tried to pursue the two base code idea a little further. To do so I have attempted to combine the table you drew based on Schimmel and Beebe's classification of aaRS's into sub-classes with the table of complementary codons I drew based on the work of the Rodins'. I start by assuming that class Ia and IIa aaRS's correspond with the earliest aa's to be incorporated into the code. The aromatic aa's associated with class Ic and IIc aaRS's are surely later additions. However, the class Ib and IIb aa's, asp, asn, glu, gln and lys, seem not to be very late additions since asp and glu are prominent in abiotic syntheses of aa's and are important in making proteinoids. However, if we take the view that matters began with the C and G bases, and that A and U were incorporated sparingly at first, then we are forced to look at classes Ia and IIa first. With this in mind I have reordered your table as follows:



To this list there corresponds a list of codons (N denotes any of the 4 bases, C, G, U and A; Y denotes pyrimidines; $\bar{N}$ denotes the complements of N; and N' denotes bases U, C and A):

Ia	IIa
CGN	CCN̄
UGY	GCN̄
CUN	ACN̄
AUN′	UCN̄
AUG	GGN
GUN	CAY

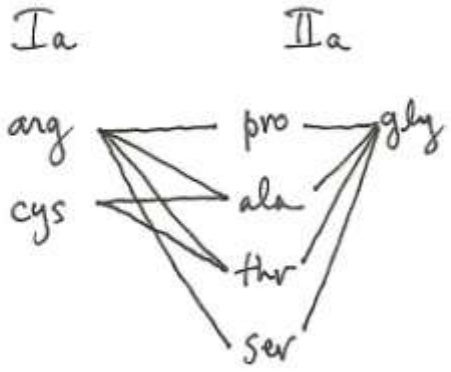
The C and G code is given by the upper portion of these tables. A connecting line connotes complementary codons:

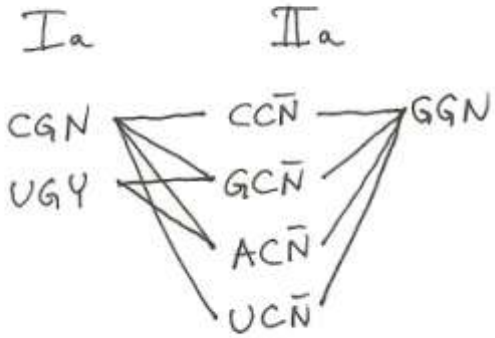


Note that 3 out of 4 cases are class IIa. This corresponds with the Rodin rules for whether or not complementary codons are from the same class or from different classes. Note also that the contemporary aaRS subunit structure is simple for arg, i.e. α, and merely α_2 for pro, but α_4 for ala and $\alpha_2\beta_2$ for gly. Perhaps this reflects a lot of evolution for the gly and ala cases and is related to their very simple aa residues.

Reynard: This is *very* nice. I suspect that you will now allow some A's and U's to come into play and increase the aa repertoire.

U: Yes, but I will only allow one A or one U at first. This leads to the tables:



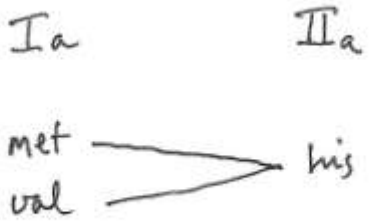


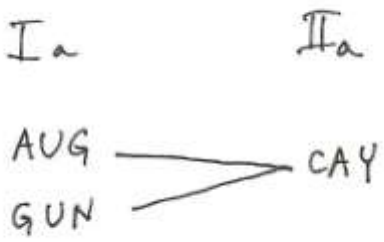
When you see either N or $\bar{N}$, only some of the possibilities may apply for a particular connecting line. For example, in the very top horizontal connection for

the codons of arg, pro and gly respectively, $\overline{N}$ is G for the lefthand side and $\overline{N}$ is C for the righthand side, but N is G in both cases. Using complementarity, all cases can be made explicit. Especially note that no change in the second base has been used, only first base changes have been made.

R: This is now getting really interesting. You are suggesting that cys is an early acquisition for the code! This differs markedly from what others, such as Edward Trifinov (*Early Molecular Evolution*), have concluded using other attractive models. Because so many abiotic simulations of aa formation do not include sulfur compounds in the mixture, cys is not a product. Adding molecules containing S to the mixture is perhaps worth a try. Several reports in the literature bear this out, and in addition to cysteine they also produce methionine. In addition, in [Part 2, Harnessing Energy] it was argued that the thiol group plays a central role in energy processing evolution so it makes sense that an aa with a thiol residue would be an early component. Methionine plays a very special role in translation-initiation in the contemporary system. Perhaps this role was acquired early on as well. How do you see the emergence of met?

U: Are you feeding me an easy questions for a reason? Pedagogy ? Just look at the remaining portion of the Ia and IIa classes. By allowing A and U, we get a disjoint pair of complementarity diagrams:





Not only does this include met but it also brings in his. Having both cys and his in the picture permits the early products of the *primitive RNA translator* to possess the ability to bind metal ions such as Z^{2+}. An especially interesting contemporary example is the case of the zinc-finger motif.

Among many other functions, is the function of zinc fingers as essential components of the RNA polymerase complex. To say that their function is essential but distinct from the polymerase component proper, is to say the principle function of transcription/replication for the ur-RNA's resides partly in the fingers, at least. In structure the fingers have an antiparallel double stranded β-sheet connected to a small α-helix by a Z^{2+} connection. From the α-helix part two his residues combine with two cys residues from the β-sheet portion to contain one tetrahedrally liganded zinc ion.

http://en.wikipedia.org/wiki/Zinc_finger

This structure contains 30 aa's, making it much bigger than the ur-proteins we have argued for up to now. Its ur-gene requires at least 90 base pairs, far more than the 6 or 7 in the polymer products energy evolution generates. Thus ligase as well as transcriptase/replicase activity is needed early. In short, zinc fingers are part of a much more evolved state of the system. Nevertheless, the "agreements" are remarkable.

R: By "agreements" you mean consistencies between the evolved and the putative primitive mechanisms?

U: Yes. For example, the zinc fingers for the modern replicase complex bind the major groove of the DNA and cover three base pairs in length. They bind only one strand, a G rich sequence. 5 of 6 contacts between a finger and DNA are by one aa. Can you guess which one?

R: From the earlier discussions centering on the arg residue interaction with ribophosphates, and because it is all that is available for the purpose (*felicity*), I will bet on arg !!

U: And so it is. Don't overlook the fact that it is G that matters in the zinc fingers. G is the second base of the CGN codons, known to interact with arg by SELEX, and suspected to do so determined mostly by the second base. Indeed the usual second-base-ordered tables of the code are frequently interpreted to show that it is the second base that counts most, followed by the first and not much influenced

by the third. Note in addition that after A's are allowed into the code two more arg codons are possible, AGR. G is still second base, as if to emphasize the point. I have ordered the table for the code so that it readily reflects the arguments made above. Thus the upper lefthand quarter is based on C and G only. The two leftmost columns contain arg, pro, ala, thr, ser, gly and cys. The extra trp in the codon table is surely a late arrival and may have initially been a second stop.

Genetic code

1rst base	2nd base: C	G	U	A	3rd base
C	pro	arg	leu	his	C
C	pro	arg	leu	his	U
C	pro	arg	leu	gln	G
C	pro	arg	leu	gln	A
G	ala	gly	val	asp	C
G	ala	gly	val	asp	U
G	ala	gly	val	glu	G
G	ala	gly	val	glu	A
U	ser	cys	phe	tyr	C
U	ser	cys	phe	tyr	U
U	ser	trp	leu	stop	G
U	ser	stop	leu	stop	A
A	thr	ser	ile	asn	C
A	thr	ser	ile	asn	U
A	thr	arg	met	lys	G
A	thr	arg	ile	lys	A

R: Some general remarks seem to be in order. Unless the transcription/replication rate of good ur-genes can keep pace with the division rate of the growing ur-cells, there is no chance for genetics to be active. Catalysts seem to be essential. The evolution of the aaRS's increases the rate of RNA translation, some of the products of which are hydrophobic, even collagen-like, polypeptides that can self-assemble into the bounding membrane. This enables the membrane to grow and to eventually bud or divide [Part 3, Compartmentalization], [Part 4]. This dynamics puts pressure on evolution to produce an ur-transcriptase/replicase so that copies

of the ur-gene's message can keep pace with ur-cell numbers rather than be diluted by a fast ur-cell division rate and a slow RNA replication rate. Perhaps the zinc fingers have a simpler precursor, based around zinc ion chelation, that helps promote transcription/replication. Both cys and his are relevant. Until the RNA ur-polymerase problem is solved the ur-aaRS problem isn't operative. This is another chicken-egg sort of conundrum. As usual co-evolution is the answer. Zinc ur-fingers surely still chelate using cys and perhaps his. Even among contemporary zinc fingers there are those that chelate using 2 cys and 2 his and those, putatively more primitive, that use 4 cys (other variations also exist such as 3 cys and 1 his, or 6 cys and 2 Z^{2+}). Thus, in the early stages of your description of the code evolution, cys is available first and then cys and his become available together not long afterwards. Arg is there from the start, possibly because of primitive metabolic reactions underlying the early evolution of the urea cycle. Note, also that in your diagrams, one row differs from an adjacent row by single base changes only.

The classes Ib and IIb invite comment. Since the first two bases of the asp and glu codons contain second base A and first base G, they are not part of the initial set of coded for aa's based on C and G alone. Asn, however, has AA for its first two bases and gln has CA for the first two. This is consistent with the idea that the code began with C and G only and added in A and U gradually, first just one and then perhaps two in the first two codon positions. Glu, (gln) and asp would have preceded asn and (gln). This is consistent with what was argued earlier [Part 8] for the corresponding aaRS's. As for asp and glu, they are connected to leu and ile by complementarity. Thus, (leu, ile, asp and glu) can be added as a group, just as were (met, val and his). The problem this poses is whether the simpler system that has capacity to evolve requires aa residues having carboxyl groups as do asp and glu. It is not easy to imagine that a system using only positively charged and neutral polar residues (as well as the hydrocarbon residues) can do all that is required. The added versatility a negatively charged residue contributes strongly suggests that classes Ib and IIb emerged at the same time as classes Ia and IIa. Nevertheless, the minimal (arg, pro, ala and gly) system is thought to have functioned successfully up to some point. What are the details that characterize this system at this point and also characterize the transitions from (arg, pro, ala and gly) to (arg, pro, ala, gly, cys, thr and ser), and from (arg, pro, ala, gly, cys, thr and ser) to (arg, pro, ala, gly, cys, thr, ser, met, val and his) as you have suggested with your diagrams. When do (leu, ile, asp and glu) enter the picture?

The bacterial RNA polymerase today has in its phosphodiester bond forming active site 2 Mg^{2+}'s that are coordinated by several asp residues. Only one Mg^{2+} is tightly bound in the active center while the other likely arrives coordinated to the nucleotide triphosphate that will be added to the chain with the release of pyrophosphate. If these asp residues are essential for the ur-polymerase as well then asp had to be part of the compliment of early aa's. However, it is possible to

imagine alternatives to the asp residue so that asp would not have to be present as early as an ur-polymerase would have to be. Indeed, the ur-polymerase may have been a simple Mg^{2+} ion. However, as we have seen with the minimal model (arg, pro, ala and gly) we do expect arg rich polypeptides that act as ur-ligases and ur-transcriptases. Perhaps the arg positive charge does the job of the Mg^{2+} positive charges in an ur-polymerase that later evolves zinc ur-fingers and bound Mg^{2+} ions.Today the mechanism of RNA polymerase is fundamentally the same in all cell types. In eukaryotes there are many extra protein factors and the complex also engages in many regulatory interactions. But the main point is that the active site mechanism that makes a new phosphodiester bond is very highly conserved and features a Mg^{2+} ion. This is in marked contrast to the aaRS's.

Both zinc and magnesium are alpha-multiples [Elements of life]. Magnesium is quite high in cosmic abundance. Both are natural candidates for metal catalysts that are still in use today.

Another perspective regarding asp and glu comes from what you said about zinc fingers. The key to their emergence in the system is cys. Metal ion chelation, particularly zinc in this case, depends on the advent of cys. Apparently this can occur before the advents of asp and glu. Better RNA ur-polymerases are needed right away and may be based partly on zinc ur-fingers. Thus some improvement in RNA ur-polymerase (pure arginine to asp?, glu?, zinc and magnesium) could precede the development of complex aaRS's. When asp and glu finally enter the scene, much more sophisticated catalysts become possible. In all of this one has to keep in mind polymer size. Polymers are not very long initially without RNA ur-ligases. Once these arise naturally, longer RNA's become possible, and thereby longer polypeptides!! Note that the longer RNA's precede longer polypeptides in this model. No polypeptide *ur-ligase* is required and indeed none exists in today's organisms. Longer polypeptides are always made by addition of more monomeric aa's. These longer polypeptides may finally have enough length to have, say, 4 cys residues that can chelate zinc ion. However, now we see why there might have been evolutionary pressure to have RNA translation make the transition from the *primitive RNA translator* to the linear mRNA reading ur-tRNA, ur-ribosome based complex, the *ribosomal* system.

U: How the *primitive RNA translator* became the *ribosomal* system remains to be demonstrated step by step. As you say, **why it evolved** seems to be connected to the physical difference in mechanism. The *primitive RNA translator* uses a helix-coil transition to translate the RNA whereas the *ribosomal* system uses tRNA's, aaRS's, … and RNA polymerases, to read the RNA as an ur-mRNA. The RNA ur-gene of the *primitive RNA translator* system that is read directly off of the RNA becomes the ur-mRNA for the *ribosomal* system that is read, linearly over

long polynucleotides, by a ribosomal complex. You seem to see the evolution of this stage as genuine *RNA World* mechanism. Am I right?

R: Not the *RNA World* again??
You are right that this issue needs to be addressed, again.

For the ur-gene trapping behavior of the *primitive RNA translator* mechanism inside a racemic proteinoid ur-cell to exist, ur-genes, i.e. RNA strands, need to be replicated fast enough to avoid dilution during ur-cell growth and replication. Hence the need for an ur-polymerase. Without it there is no point to discussing aaRS evolution because there is no way to *lock in* the base sequences for catalysts (enzymes and ribozymes). If the rate of RNA replication (perhaps as a double transcription) can keep up with the rate of ur-cell multiplication, a propagating genome gets *locked in*. The ultimate ribozyme in my view is the *primitive RNA translator*. This RNA is both the ur-gene and the ur-mRNA at the same time in the *primitive RNA translator*. *Functioning* as an ur-gene requires replication. Perhaps the first ur-gene is 5'N(CGN)n [see the end of Part 5], and its arg rich products are the ur-polymerase, most likely in concert with Mg^{2+}. Both arg and Mg^{2+} interact strongly with RNA ribophosphate backbones at short range (Debye length). In the currently accepted mechanism of Mg^{2+} catalysis in contemporary polymerases the metal ion is the core factor (*The structure of bacterial RNA polymerase*, Kati Geszvain and Robert Landick In *The bacterial chromosome* (ed. N.P.Higgins), pp. 283–296. American Society for Microbiology, Washington, D.C.). A faster polymerase gives a genome the chance to spread more widely than a genome without. Thus, much evolution of this function is expected. Today the molecular complex in which this function resides is an enormous pure protein complex. Any protein components require some type of translation, starting with the *primitive RNA translator*. So the ur-polymerase and the translation process will co-evolve. Apparently the (Mg^{2+}, arg, pro, ala and gly) system functions at first in energy rich chemical soups containing pyrophosphate, P~P. It supports the first ur-polymerase in arg rich translation products. It also makes lots of hydrophobic sequences, based on (gly, ala and pro), that I like to call ur-collagen, that self-assemble into the ur-cell membrane. Membrane growth and ur-cell division create daughter ur-cells that each contain the same genomes, provided the ur-polymerase rate is large enough that by the time an ur-cell divides, an ur-gene has been copied at least once, and, better yet, many times. Robust Brownian motion distributes the multiple copies of any molecular species as *fairly* as any process can. Can this system evolve aaRS's? Do we need the addition of (Z^{2+}, cys, thr and ser) before it will work? For aaRS's to work ur-tRNA's are also needed. An ur-ligase in needed to get RNA's up to length 20, the length at which ur-tRNA's can form. Can we work out the steps of this evolution?

U: No, this is not classic *RNA World* mechanism.