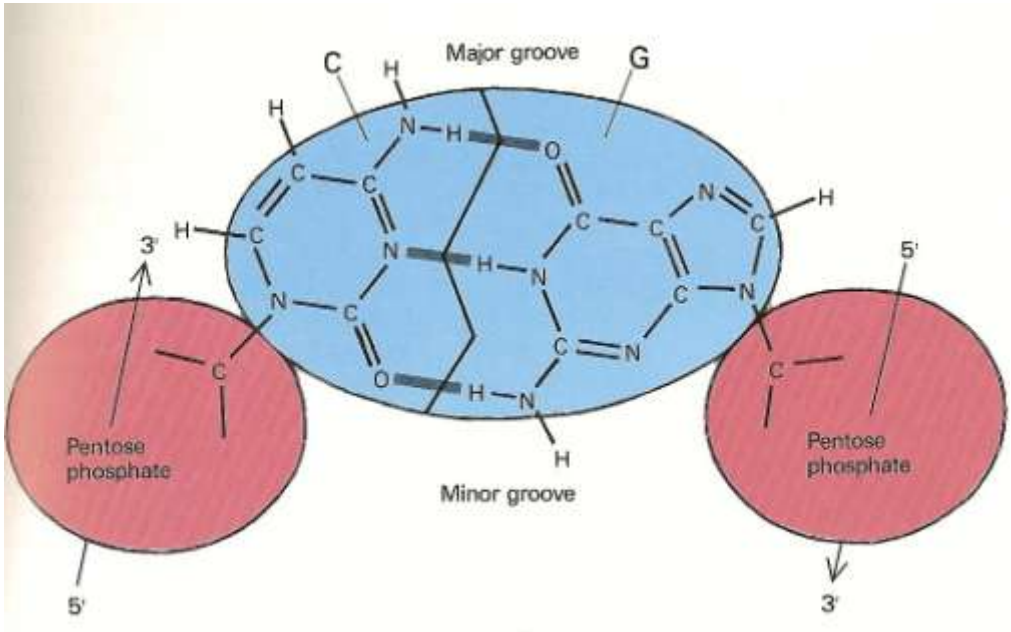
Practicing Molecular Algebra

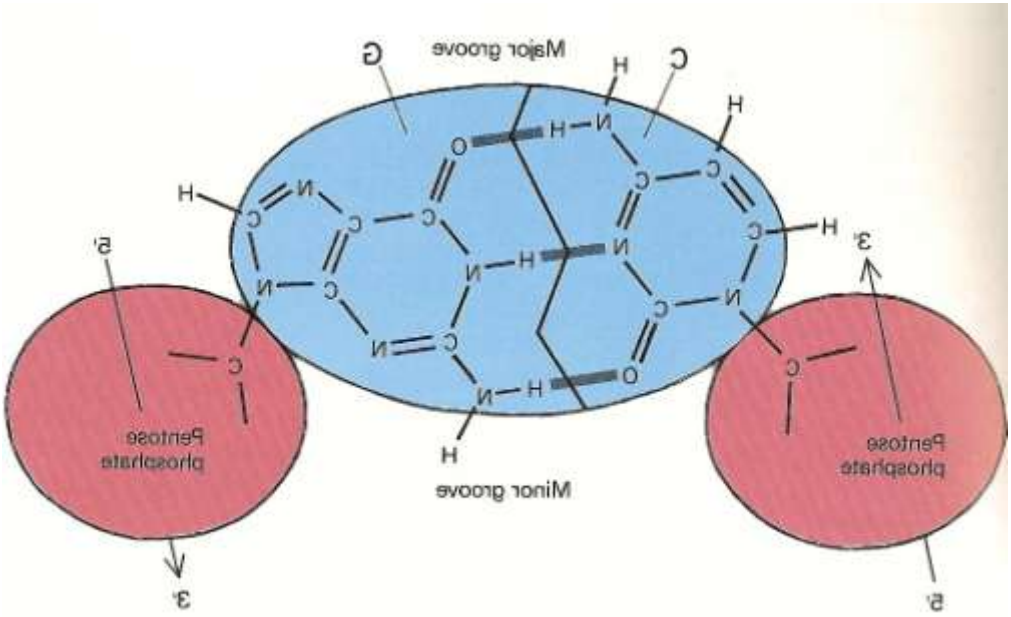
Reynard: By way of review and as an exercise in *molecular algebra*, let us look at base pairing in more detail. The modern system uses *two* sets of base pairs, CG and UA. With these *four* bases, 64 *three* base codons can be made. However, it is possible that the coding began with only one base pair, say CG, the more thermodynamically stable of the two available (three hydrogen bonds instead of two). Both the base pairing processes and the *primitive RNA translator* function as already described. The smoothness or ease of this functioning is greatly enhanced if structures of one pure chirality are used. That RNA replication and translation are much smoother for a system of pure chirality than for a racemic system is not obviated by reduction from two base pair sets to just one set. So we will analyze a one base pair system based on CG.

Uranya: The fact that the modern system uses two base pair sets is another example of *felicity*. The use of two sets depends on the "accident" that the shapes of CG and UA base pairs are the same, as far as joining to the two ribose-phosphate backbones of the double helix. In the figures, the planar angle between the two dangling bonds (attachments to ribose-phosphate *backbone*) is the same for CG and UA (and GC and AU as well).



The pentose is ribose in RNA and deoxy-ribose in DNA. The antiparallel pentose phosphate *backbone pair* runs perpendicular to the page. Note that

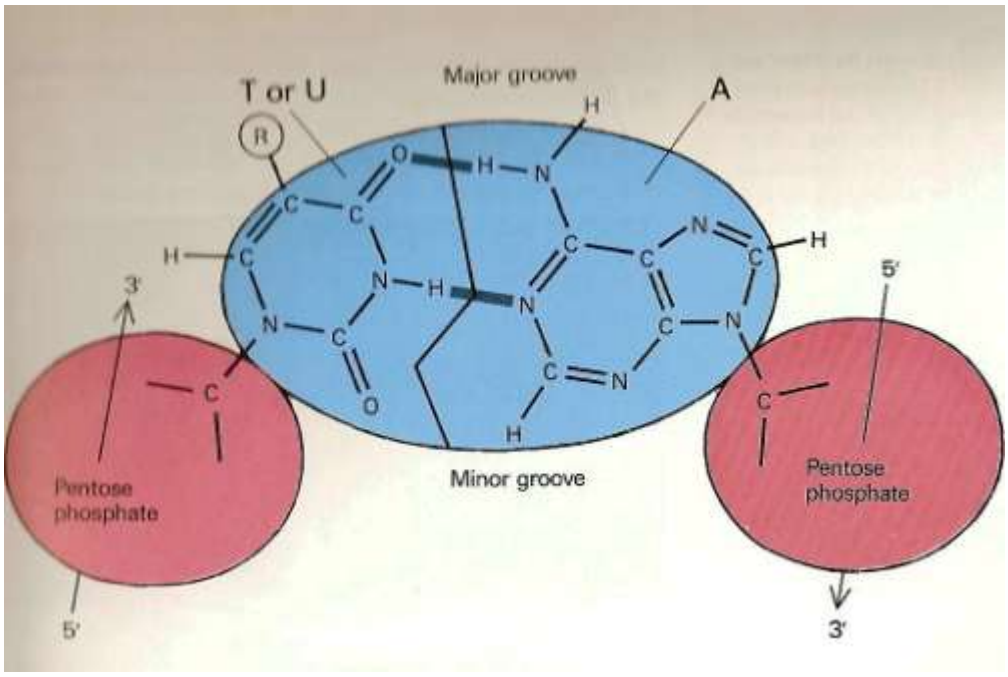
the CN linkage to pentose points northeast on the left and points north-northwest on the right. This difference in angle is 67.5 degrees (3/8 π).



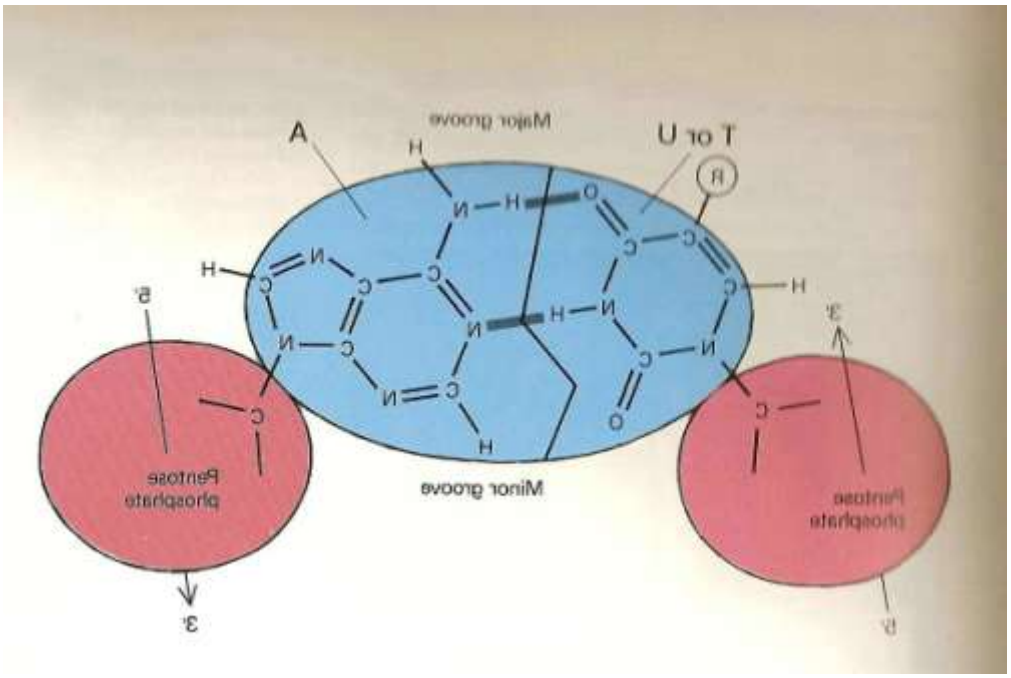
When reversed the CG base pair becomes the GC base pair, but with different angles of attachment (Since this image was produced by reflection, the polarities of the backbones are reversed from what they should be if we had simply drawn the GC figure from scratch' i.e. from the blue part but with each letter reversed to normal form, and from the red parts but with the arrows pointing the same as in the first CG figure.) The CN linkage on the left is steeper than that on the right, the reverse of the situation in the CG picture. This means that the attachment to the ribose requires some accommodation by the ribose. These are C1 linkages to ribose, the plane of which is oriented similarly to that of the base pair plane (in the page) but translated perpendicularly to this plane. With ribose-*pucker* freedom an accommodation is easy to imagine.

By the way, I have redrawn these figures from a figure in De Duve's brilliant book *A Guided Tour of the Living Cell* (W.H. Freeman and Co., New York, 1984,) In the picture below, R is H in (U)racil and is CH3 in (T)hymine.

The next picture shows the U(T)A base pair. The similarity to CG is manifest. Even the hydrogen bonds are parallel, and two of each overlap with their hydrogen bond and covalent bond stretches interchanged (the N to O directions are reversed.), when the shapes are superimposed.



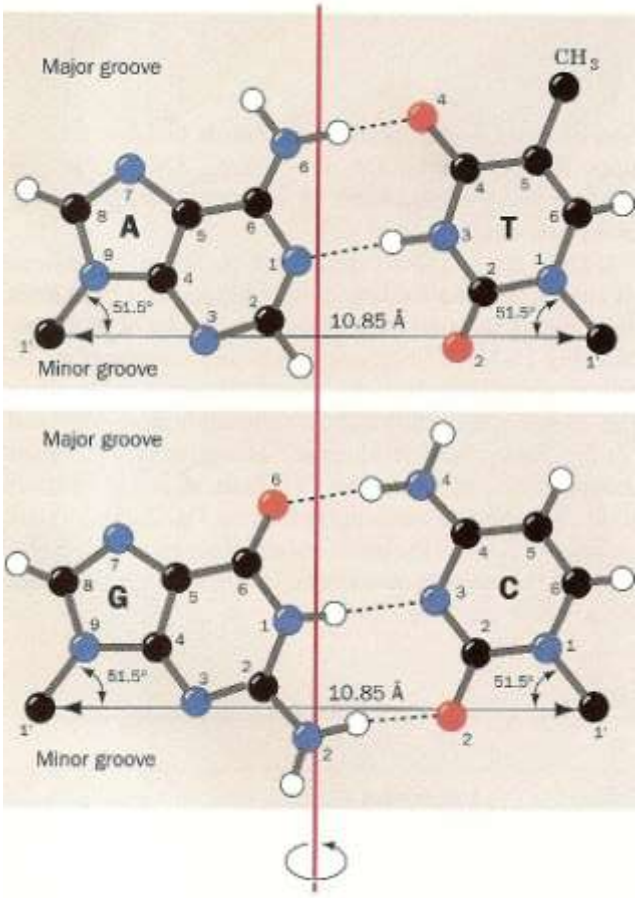
The AU version is given below. Note, each of the symbols AU(T) is inversion invariant ! :-) So look carefully.



Given CG (UA), the shape of GC (AU) is not exactly the same as regards the two bonds required to connect the base pair to the ribose-phosphate helices! The relative linear displacement of the two bonds is the same as must be so in a mirror reversal, but the orientations of the attachment

angles are somewhat different by about π/8 radians, a small but non-trivial amount. This simple and obvious point is often overlooked. Once UA is added and has the same dangling bonds geometry as CG, so then can AU be added to parallel the inclusion of GC.

R: I must protest your presentation. Much of what you said hinges on a diagram, the one you re-drew from De Duve's book. The difference in the attachment angles was the key. However, this is an illusion re-enforced by the figure. The orientation of the base pairs can be represented so that the attachment angles are identical. Then the mirror reflections are invariant (dyad symmetry) ! I have a schematic re-drawn from *Biochemistry* (D. Voet and J.G.Voet, John Wiley, New York, 1995). The attachment angle is 51.5 ^{o} .

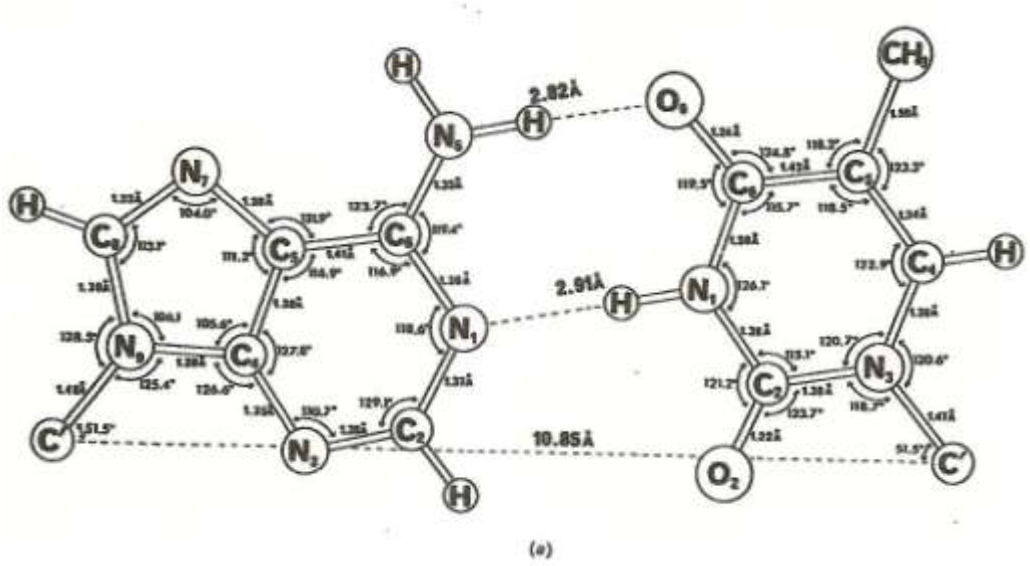


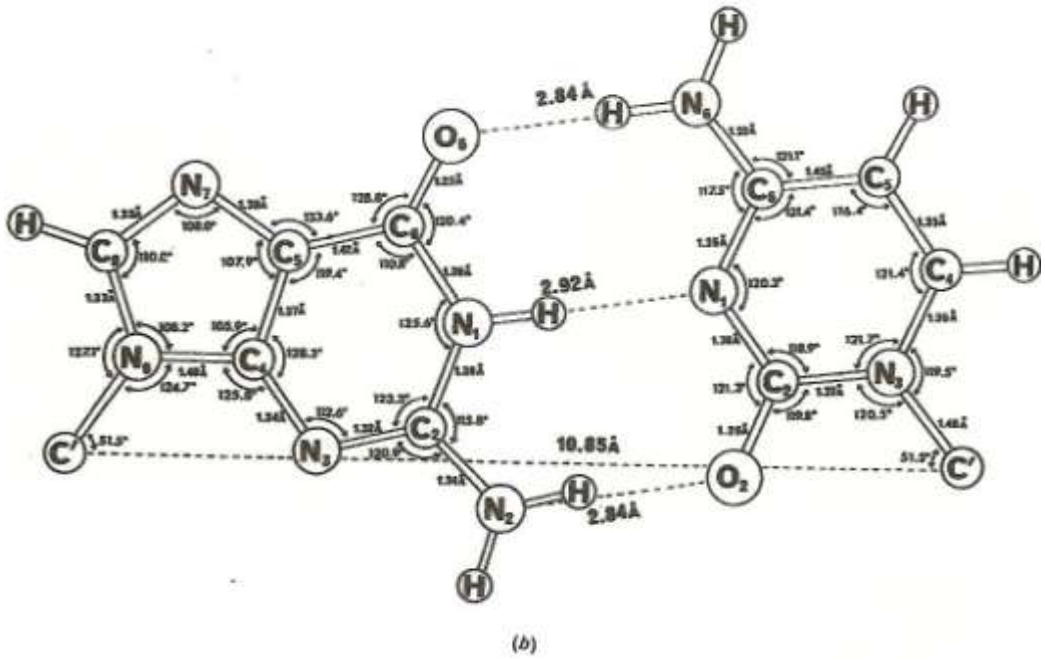
So the felicitous accident is solely the identical shapes and sizes of CG and UA, and GC and AU. Each ribose in the helix is attached to a base pair with the same

conformation and orientation. Rotation by π around the symmetry axis in the figure does not alter the attachment angles or require "ribose accommodation." That this is so is a very *felicitous* accident. Perhaps a quote directly from the research article (Struther Arnott, et al. *Acta Cryst. B.* (1969), 25, 2192, p. 2195):

"It is noteworthy that, in spite of the rigorous demands of contemporary values of covalent and hydrogen bond lengths and covalent bond angles, the nucleic acid purines and pyrimidines lend themselves very readily to the construction of base pairs with perfect diads relating the glycosidic links and with identical distances between these links, so that all the pairs can replace one another without distorting the regularity of the sugar-phosphate backbone. "

Reread this a few times. It is a single sentence ! By "all the pairs" 4 choices are permitted. I re-draw below figures from the Arnott et al. paper, upon which the figure above is based. In (a) is AT and in (b) is GC.





Especially note that because these five and six membered rings contain both C and N in asymmetric places, the angles are not perfect angles for regular pentagons and hexagons. This just makes the outcome all the more amazing.

U: I stand corrected. Figures in textbooks frequently mislead the reader. There might not have been an UA that matched up with CG in this way. If it weren't for this *felicitous* "accident" the ur-cell would have had to make do with the *one* base pair, or *two* base system described here. That each base pair is pyrimidine-purine in structure partially explains why we have more than one pair at work today, since there are several pyrimidines and purines from which to choose. Given this accident a richer code is possible with two pairs rather than one. Making a study of the one base system may make clear some of the advantages to having more than 8 codons.

R: When you say "accident" do you mean like the case of the apparently equal radii of the sun and the moon as seen from the earth today? Long ago the moon's apparent radius was much different. The beautiful solar eclipses (coronal) we get to see now, are esthetic pleasures and *felicitous* "accidents." For the base pairs, it isn't their sizes that can change but instead the fact that there is more than one pair with the same attachment angle (singular). But it is more, it is also the fact that even one base pair exists since that one pair must still possess the dyad symnmetry, the underlying *felicitous* circumstance.

U: Yes, I mean "accident" exactly like that. Note that we still have a three base spacing for the code because that still is how the *primitive RNA translator* works. Only the codon's first two bases matter for the coding and we can assume that the primitive genetic code is given by (N denotes any base, either C or G or Y or R)

codon	amino acid
CGN	arg+
CCN	pro
GCN	ala
GGN	gly

This has one charged residue in arg+ and three hydrophobic residues in pro, ala and gly. One should note that if instead of CG we use UA for the two base system, then we get the table

codon	amino acid
UUU	phe
UUA	leu
AAA	lys+
AAU	asn
UAU	tyr
UAA	stop
AUU	ile
AUA	ile

This contains one positive aa and three hydrophobic aa's, as in the first table, but it also contains "modern" aa's such as asn, phe and tyr, as we discussed when we talked about aaRS's. It does not have the look of a primitive table.

R: Are you assuming a physical-chemical basis for the code, as was done earlier for CGN and arg ?

U: Yes, for all four cases. Nothing absolutely and exclusively precise but rather as preferences, i.e. CGN and arg are connected, i.e. *cognate* most of the time. Since GGN has the two biggest first-two-bases of all codons, it is not surprising that gly, with the smallest aa residue, dominates the others in acting cognate to GGN by forming the aminoacyl-ribose ester bond.

The question arises whether pro can be used by the *primitive RNA translator* mechanism. Does its odd shape stop the translator mechanism? In the affirmative case there are only three aa's with which to build molecules. In the negative case there are four amino acids including pro. Amino acids gly and ala are often the dominant two amino acids in pre-biotic simulations. Arginine may be more common from ur-metabolism that exists because of ammonia processing and involves ornithine, citrulline, carbamoyl-phosphate and urea (see the urea cycle). Anyway, chains of these amino acids would produce polypeptides that interact with polynucleotides, and polypeptides that would integrate into the ur-cell membrane, depending on arginine content. The positive charge of the arginine residue makes it ideal for interacting with polynucleotide phosphates, and its guanidino group can hydrogen bond with base groups. This may explain arginine's preference for CGN over GCN, leaving GCN for ala.

The presence of pro makes possible a very good ur-collagen, the ancestor of the most abundant protein in vertebrates today. With precisely, gly, pro and an occasional ala or arg, a very representative collagen can be made. Collagen is repeating triplets of gly-X-Y where X is often pro and Y is often hydroxyproline, a post-translational modification of pro. Perhaps the primitive system had no hydroxyproline and just used pro instead. An occasional ala or arg, etc. for Y is typical of modern day collagens. Without hydroxyproline, the ur-collagens may have to be single strands.

R: It amazes me that you can be talking about collagen so early in the one base pair analysis. To make even a single triplet repeat, i.e. a hexameric collagen, an RNA of order 20 residues is needed. Where do you get such long RNA's? Is urcollagen a good candidate for a polypeptide that self-assembles into the membrane, promoting membrane growth and division? Will hexamers with 2 or 3 arg's be candidates for ur-ligases and ur-transcriptases (note that an RNA transcript is not a replica unless the RNA sequence is a palindrome). Two iterations of the ur-transcriptase can produce an RNA complement and the complement of the complement, the replica, if there are no errors. So we see that the early ur-replicase is the same as the ur-transcriptase. An RNA coding for an ur-transcriptase would replicate itself faster than an RNA that did not. If the system is rich in pyrophosphate generated by ur-metabolism, then it is likely to be making plenty of ur-collagen and copies of its ur-replicase ur-gene (an RNA). Microsphere growth and division, with distribution of equal shares of the genome copies, would constitute a *living genetic system*. Am I seeing things the same way you see them?

U: A modern replicase need not use this simple arg based mechanism. Much evolution has taken place. Arg could be replaced by lys or his or even inorganic cations, e.g. zinc and magnesium ions, coordinated with the polypeptides. The modern mechanism may be based on interactions other than electrostatic. What is known is that the modern mechanism involves a complex of many components and many complex steps. How the modern replicase complex evolved from a much simpler system, and what that simpler system was, are hard questions, on a par with from where did the aaRS's come. RNA *complementation* (transcription, replication) and translation benefit from catalysts of several sorts. How the ur-polymerases for RNA arose is a problem on a par with the origin and evolution of protein biosynthesis.

R: Any process having a complex molecular apparatus, acting as a catalyst, whether enzyme or ribozyme, is latent in an ancestral form that is much simpler in structure. Finding the simpler form, and explaining how it is rich enough for evolution to generate the complex form, is the challenge. The history of RNA polymerase starting with a simple ur-polymerase is a long complex process. Available in the two base, one base pair, model are polypeptides containing arg's and the ability to interact with RNA. Within these must lie the ur-polymerase capacity (*felicity*). The modern day *E. Coli* RNA polymerase holoenzyme has a mass of 449 kD, about 4,000 aa's long. This is a big, highly evolved polypeptide complex.

U: I am still waiting to hear how aaRS's arose to connect aa's with their cognate codons, and presumedly much about ur-tRNA's.

R: Patience! How far towards this goal will the two base model take you?