Chapter 3

Catalytic Propensity of Amino Acids and the Origins of the Genetic Code and Proteins

Ádám Kun[1,2], Sándor Pongor[3,4], Ferenc Jordán[2,5], and Eörs Szathmáry[1,2,6]

Abstract The origin of the genetic code is still not fully understood, despite considerable progress in the last decade. Far from being a frozen complete accident, the canonical genetic code is full of patterns that seem to open a window on its evolutionary history. In this chapter we rethink the hypothesis that the primary selective force in favour of the emergence of genetic coding was the added value by amino acids to the RNA world in the form of increased catalytic potential. We identify a novel pattern in the genetic code suggesting that the catalytic propensity of amino acids has considerably shaped its structure. This suggestion complements older ideas arguing in favour of a driving force to build the smallest stable oligopeptide structures, such as hairpins (β-turns stabilized by small β-sheets). We outline experiments to test some of the proposals.

1 Introduction

As Crick et al. (1976) noted, 'the origin of protein synthesis is a notoriously difficult problem'. This remark refers to protein synthesis in translation using the genetic code. When thinking about difficult evolutionary transitions (cf. Maynard Smith and Szathmáry, 1995) it is rewarding to break down the problem into steps that are more readily soluble by evolution and easier to understand for us. The idea

[1] *Biological Institute, Eötvös University, Budapest*

[2] *Collegium Budapest, Institute for Advanced Study, 2 Szentháromság utca, H-1014 Budapest, Hungary*

[3] *International Centre for Genetic Engineering and Biotechnology, Padriciano 99, 34012 Trieste, Italy*

[4] *Bioinformatics Group, Biological Research Center, 6726 Szeged, Hungary*

[5] *Animal Ecology Research Group of HAS, Hungarian Natural History Museum, Budapest*

[6] *Parmenides Center for the Study of Thinking, 14a Kardinal-Faulhaber-Str, D-80333 Munich, Germany*

M. Barbieri (ed.), *The Codes of Life: The Rules of Macroevolution*.

of an RNA world (e.g. Gilbert, 1986) is important because it separates the problem of life's origin from the origin of translation. The origin of the genetic code by itself seems to be burdened by a dual difficulty: no meaningful proteins without the genetic code, and no genetic code without the appropriate proteins (especially synthetases). Fortunately, there is a way out: we know that selected RNA molecules (aptamers) can specifically bind amino acids, and charge them to RNA either in *cis* or in *trans* (discussed below). The fact that peptidyl transfer in the ribosome is catalysed by RNA rather than proteins (Moore and Steitz, 2002, Steitz and Moore, 2003) supports the view that there was a way out from the RNA world into ours, aided by RNA itself.

Yet these exciting developments leave the nature of positive selection for the genetic code obscure. Polypeptides must attain a critical size and complexity before they can serve structural and catalytic functions in a rudimentary way. Söding and Lupas (2003, p. 837) called attention to the fact that, consonant with the RNA world scenario, the first polypeptides (supersecondary structures) would have been unable to attain stable conformation by themselves, as witnessed even by contemporary ribosomal proteins: 'The peptides forming these building blocks would not in themselves have had the ability to fold, but would have emerged as cofactors supporting RNA-based replication and catalysis (the 'RNA world'). Their association into larger structures and eventual fusion into polypeptide chains would have allowed them to become independent of their RNA scaffold, leading to the evolution of a novel type of macromolecule: the folded protein.' The path from the RNA world presumably went through a marked RNA-polypeptide phase. Corollary to this is the notion that modern metabolism is a 'palimpsest' of the RNA world (Benner et al., 1989) and that evolution of modern protein aminoacyl-tRNA synthetases may shed little direct light on the very origin of the genetic code if the ancient form of the code was implemented by ribozymes rather than proteins; at most the evolution of protein synthetases could have been partly analogous to that of RNA synthetases (Wetzel, 1995).

What could have been the force that drove life out of the RNA world? The end result is clear: proteins in general are much more versatile catalysts than RNA, partly by the virtue of the greater catalytic potential of 20 amino acids as opposed to 4 nucleotides (e.g. Szathmáry, 1999). In general, replicability and catalytic potential are in conflict: they prefer smaller and larger alphabets, respectively (Szathmáry, 1991, 1992). But evolution has no foresight: one cannot rationalize a transition by noting that the end result is fitter than the starting point: a more or less smooth path on the adaptive landscape must be found.

Some time ago, one of us proposed an idea (Szathmáry, 1990) how this could have been possible by still keeping catalysis by amino acids in focus. In short, some amino acids could have been utilized as cofactors of ribozymes in a metabolically complex RNA world. According to this scenario, amino acids were linked to specific short oligonucleotides (called handles) by ribozymes, in a manner that followed the logic of the genetic code: one type of amino acids was allowed to be charged to different handles, but each particular handle with a specified sequence was charged with one type of amino acids only. Szathmáry (1990) proposed that

this arrangement was very functional in ribo-organisms because the many different ribozymes in metabolism could have specifically bound the necessary amino acid cofactors by their handles using a straightforward base-pairing mechanism, and that the burden of accurate direct amino acid recognition could have been taken on by a few ribozymes charging amino acids to their cognate handles (Fig. 1): only as many specific charging ribozymes were required as there were different types of amino acids in this system.

Needless to say, the charging ribozymes are taken to be analogous to modern aminoacyl-tRNA synthetases. The most direct experimental evidence so far for the cofactor idea was found by Roth and Breaker (1998): the very efficient role of histidine in aiding the activity of a selected DNA enzyme that cleaves RNA.

While we think this central idea holds, the original exposition suffered from several shortcomings that were rectified later (Szathmáry, 1993, 1996, 1999). Szathmáry (1990) believed that the first adaptors were simple nucleotides that grew through evolution to trinucleotides then later to tRNA molecules, missing the problem that

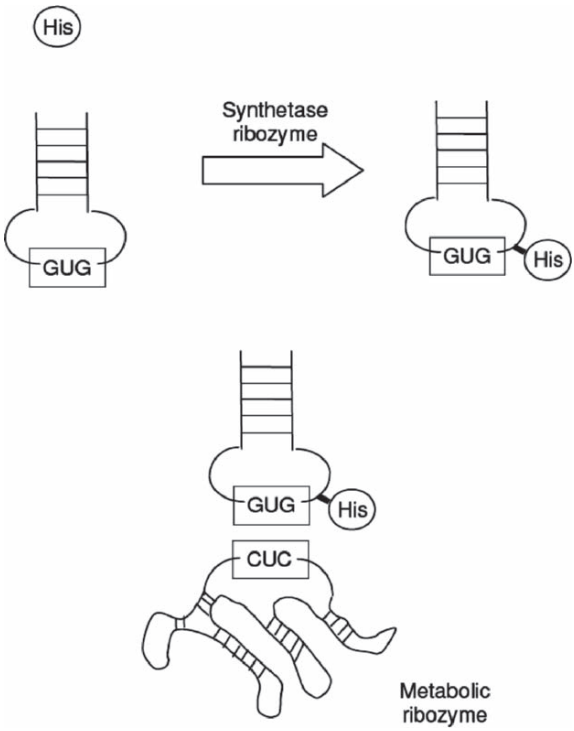

Fig. 1 Scheme of the coding coenzyme handle (CCH) hypothesis. Amino acids are *N*-linked to anticodonic hairpins by synthetase ribozymes. The products are recognized by complementary loops, embedded in ribozymes that use the linked amino acids as coenzymes. The case shown involves the metabolically prominent histidine

even trinucleotides are too short to bind well by conventional Watson-Crick base pairing to a complementary sequence in a ribozyme. The minimal, and sufficient, requirement is the interaction of two 'kissing' hairpins that provides sufficient accuracy of binding and residence time (Szathmáry, 1996). Another problem concerns the nature of the bond between amino acids and handles. This bond today is a labile anhydride bond, which is good for protein synthesis but bad for keeping amino acids bound to tRNAs for long. In contrast, stable binding of amino acids to their handles must have been a requirement in the ancient world. It is at this stage that Wong's (1991) ideas step in. He also saw the advantage of amino acids as catalytic aids in the RNA world, and argued in favour of selection for RNA peptidation. In his scenario, peptides consisting of several amino acids could have been linked to one tRNA-like molecule, which makes the logic of coding difficult to appear (T.-F. Wong, personal communication, Budapest, 1996). But Wong very correctly identified the stability problem proposing that amino acids could have been *N*-linked to some bases, as in contemporary modified bases in tRNA, for example. In fact, this goes back to the old suggestion of Woese (1972) who noted that nucleotide 37 (adjacent to the anticodon) in tRNA was always modified, and that this site could have been the one to which in ancient times amino acids were attached; but nowadays he is 'worried about the energetics of this reaction' (C. Woese, personal communication, email, 2006); we shall come back to this issue later (there seems to be a solution). Important is that Szathmáry (1999) adopted the stable *N*-linkage in his coding coenzyme handle (CCH) scenario, which comes at a price that one has to explain the origin of relocation of charging from position 37 to the 3′-end of tRNA by a different (labile) chemical bond (L. Orgel, personal communication, Stockholm, 1977).

Curiously, Szathmáry has never made a serious conjecture of amino acid entry order into the genetic code (which is surely an exciting problem; see e.g. Di Giulio and Trifonov, Chapter 4, this volume), or about the nature of the gradual building up of oligo- and polypeptides (e.g. Di Giulio, 1996). In this chapter we complement the original CCH scenario by these important considerations. We propose that amino acids were first introduced into the genetic code (by ribozymes) according to their catalytic importance (propensity). Later, other amino acids were introduced to allow the formation of β-turns (Jurka and Smith, 1987) and β-sheets (Di Giulio, 1996). Notably, α-helices arrived later: probably at around the same time when β-sheets came. The important point is that patterns (partly identified here for the first time) of the genetic code are consistent with this interpretation. The three protein features that correlated with the columnar organization of the genetic code are catalytic propensity, β-turn propensity and β-sheet propensity. We believe that Nature tells us something with this.

In this chapter we first reveal a new statistically significant pattern of the catalytic propensity of amino acids and columns of the genetic code. Considerations for the primitive ancestry of the anticodon arm of tRNA as an ancient acceptor of amino acids follows next. Then we discuss the appearance of the first oligopeptides with a novel network analysis of amino acid substitutions. Finally, we propose some experiments that could lend support to some of the evolutionary steps suggested by the CCH hypothesis.

2 Catalytic Propensity of Amino Acids and Organization of the Genetic Code

Catalytic propensity of amino acids (Fig. 2), collected from catalytic sites of known enzymes, are taken from Bartlett et al. (2002), who argued that the sample is representative.

Amino acids with the highest values gather in column A and (with smaller values) column G (Fig. 3). Is this pattern due to chance, or is it significant?

List of catalytic residues were obtained from the Catalytic Site Atlas of EMBL (Porter et al., 2004). Only literature-based entries pertaining to amino acids were used (residues inferred from sequence homologies and non-amino acid residues, such as metal ions and cofactors are left out of our analysis). In total, there were 5845 catalytic residues. The distribution of amino acids among the catalytic residues is markedly different from the frequencies of amino acids found in peptides (Bartlett et al., 2002). We performed a randomization test as follows. We took the biosynthetically restricted random set (Fig. 4) as defined by Freeland et al. (2000), which rests on the potential importance of the co-evolution theory (Wong, 1975) of the genetic code (code assignment was influenced by biosynthetic kinship of amino acids) and the observation that amino acids belonging to the same biosynthetic family tend to share the same first codon letter (i.e. they are in the same row of the table; Taylor and Coates, 1989).

Alternative tables of the genetic code were generated according to Freeland et al. (2000), limiting the number of possible alternatives to 6.48×10^6, compared to the 20, $\approx 2.43 \times 10^{18}$ totally random codes. Each of the 6.48×10^6 code tables was analysed according to the following procedure. First, the list of amino acids is ordered according to catalytic frequency in active sites. The place they occupy in

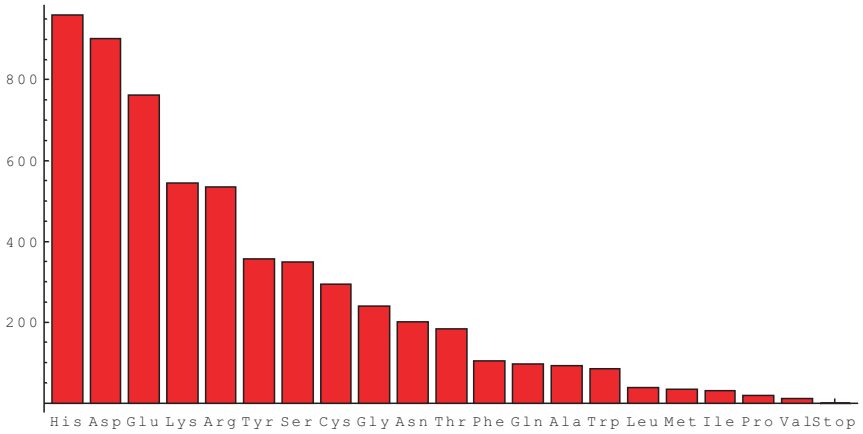

Fig. 2 Catalytic propensity of amino acids in catalytic sites of known enzymes. (From Bartlett et al., 2002.)

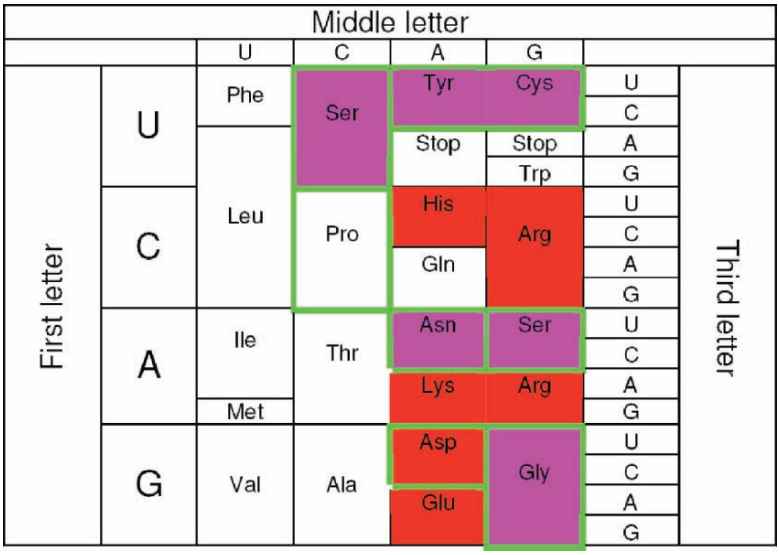

Fig. 3 Catalytic propensities and β-turn propensities superimposed on the genetic code. Only the highest values are shown (very high catalytic propensity: red, moderately high catalytic propensity: pink, highest turn propensities: green frame)

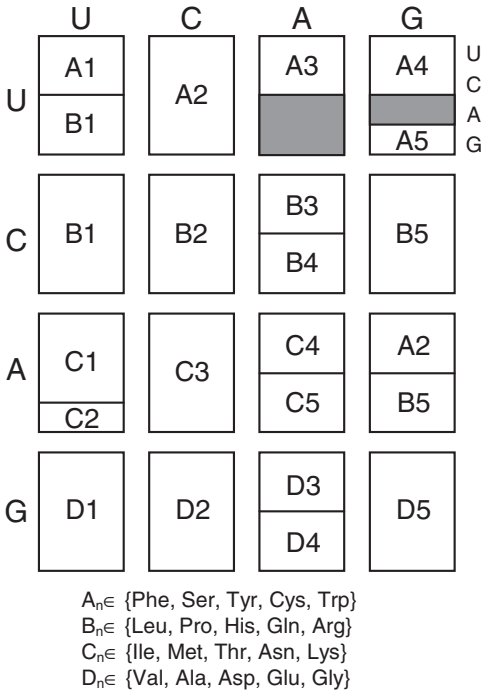

Fig. 4 The set of possible codes constrained by biosynthetic kinship (Freeland et al., 2000). In a randomized code any amino acid from set A_n can occupy any single position in the table, but only from A1 to A5. There are 6.48×10^6 possible alternative codes

this list is assigned to them as a rank. In case of a tie the average of the ranks are assigned to each amino acid having the same value. With regard to catalytic frequencies, only serine (or Phe, Tyr, Cys, and Trp in alternative codes) appears twice in the list for being in two columns, and thus has the same catalytic frequency. Sum of the ranks belonging to amino acids present in the same column of the genetic code were squared, and then summed. This procedure is identical to the calculation employed in the Kruskal-Wallis test (Zar, 1998), which is a non-parametric test employed in testing differences between multiple groups. It is similar to one-factor ANOVA, except normality of the data is not required. The pattern that amino acids segregate according to columns of the genetic code (in this order) is statistically significant ($p = 0.0107$) (Fig. 5), in agreement with the cluster analysis of catalytic propensities (Fig. 6).

We have performed a similar test of propensities for the β-turns (Prevelige and Fasman, 1989) having been taken from the EMBOSS programme (Rice et al., 2000), β-sheets (Muñoz and Serrano, 1994) and α-helices (Muñoz and Serrano, 1994): The values for the β-turns ($p = 0.0059$) and β-sheets ($p = 0.028$) have significant columnar organization *at the level of single columns*.

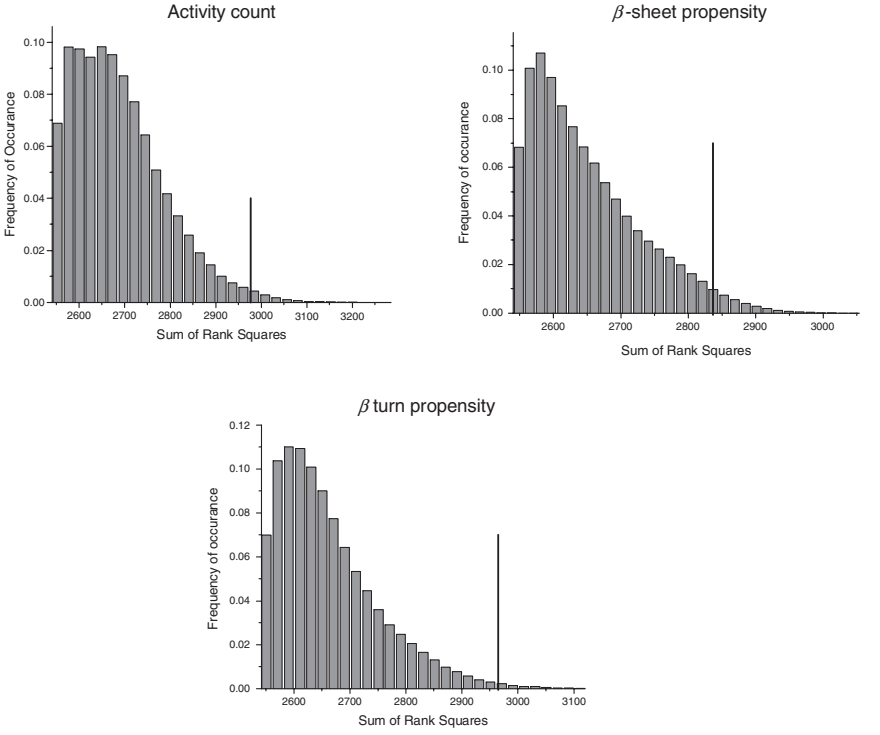

Fig. 5 Randomization test for the columnar organization of three amino acid properties in the genetic code. Thin pole indicates the position of the canonical genetic code. See text for further explanation

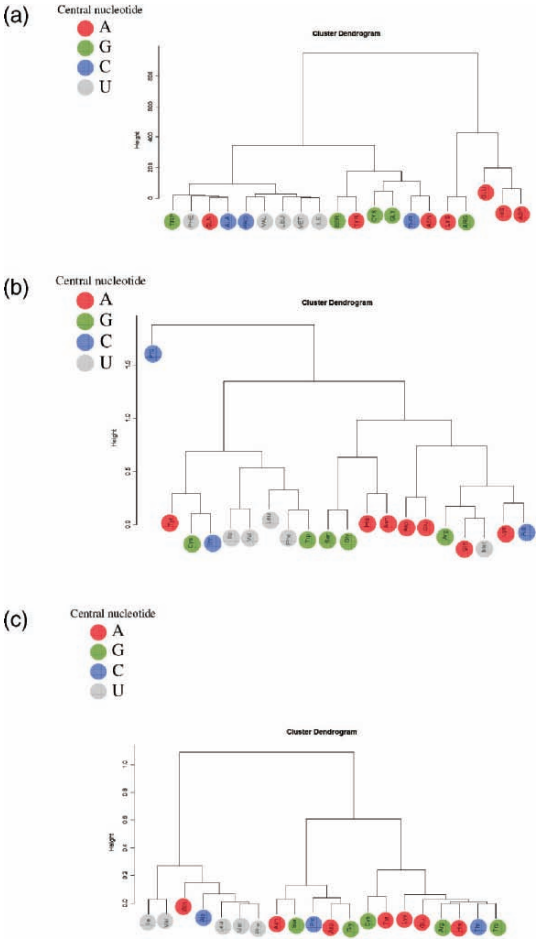

Fig. 6 Cluster analysis of amino acid propensities. (a) Catalytic propensity, (b) α-helix and β-sheet building propensities (taken together), (c) β-turn forming propensity. The hierarchical clustering was based on an Euclidean distance calculated between amino acid properties, and was carried out with the complete linkage clustering algorithm (Hartigan, 1975) as implemented in the R package, version 2.2.0 (From R Development Core Team, 2004)

It is clear that the chemically most 'exciting' amino acids are the catalytically most important ones (Fig. 7), and that the central purine bases play an exclusive role in their coding.

Our suggestion is that *the introduction of the first amino acids into the genetic code coincided with the order of decreasing catalytic importance*. This presumes that in the ribo-organisms the highly catalytic amino acids were available either in

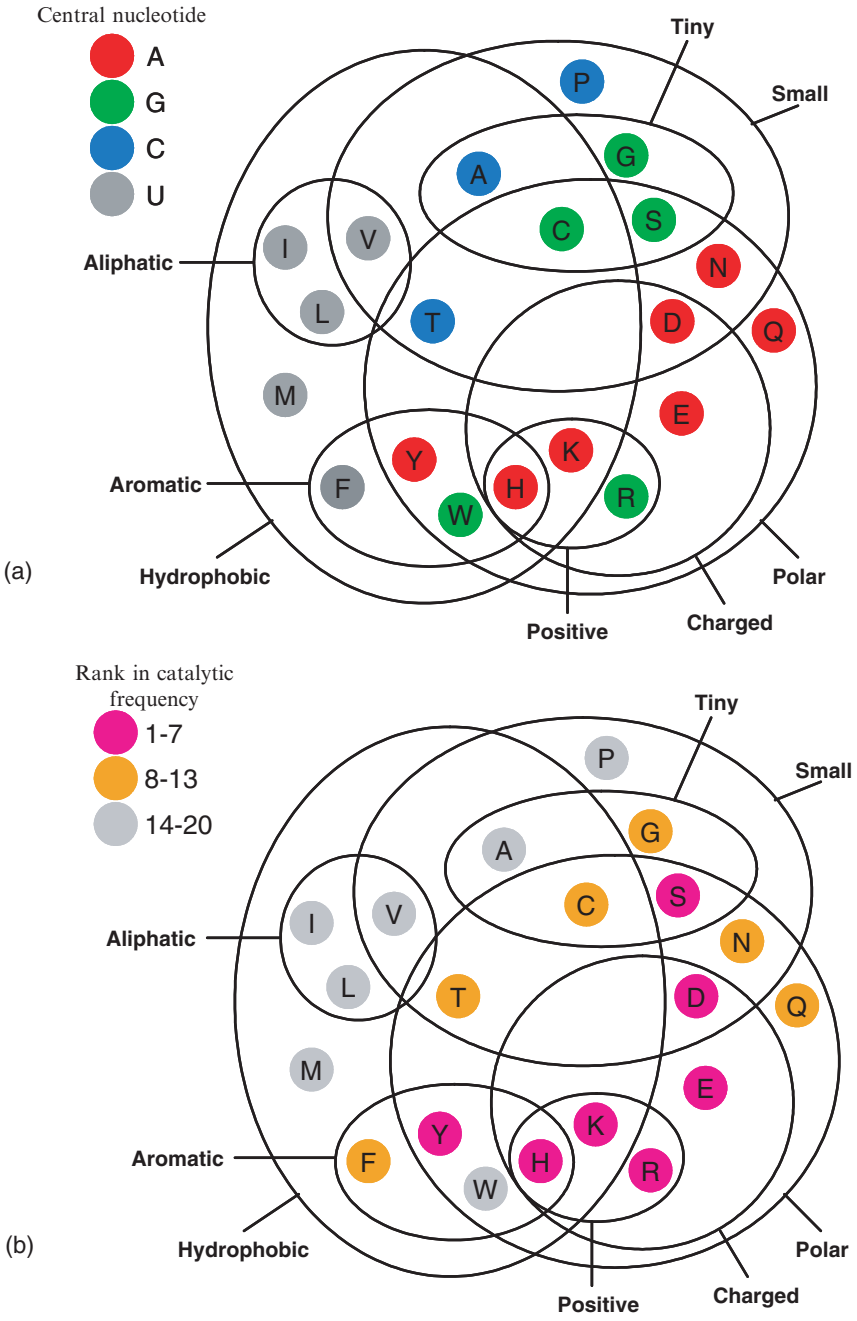

Fig. 7 Venn diagrams of amino acids (chemical sets from Taylor, 1986). (a) Distribution of amino acids based on the middle letter of the genetic code, (b) distribution of amino acids according catalytic frequency ranks and chemical properties

the medium or as a result or internal synthesis. As noted by Wong and Bronskill (1979), ideas about amino acid availability in the 'primordial soup' are inadequate when one considers origin of the genetic code. Indeed, if the RNA world was metabolically complex (which seems likely: Benner et al., 1989) then a protracted period of co-evolution of ribozymes, membranes, and metabolism is likely to have taken place (Szathmáry, 2007). Nevertheless it is important that, with the exception of lysine and arginine, all catalytically important amino acids seem to have at least some prebiotic plausibility (Miller, 1986), including histidine (Shen et al., 1990). Lysine has two different, complicated biosynthetic routes in modern organisms (Berg et al., 2003), so for the time being it is safer to assume that it is a very late invention. We propose that its role and position in an interim genetic code could have been taken by arginine (see section on protein appearance). We believe that arginine goes back to the RNA world, supported by its recognition by RNA aptamers with codonic binding sites (Knight and Landweber, 2000).

As discussed above, the ancient charging enzymes are assumed to have been ribozymes. Specific aptamers for the amino acids Arg, Ile, Tyr, Gln, Phe, His, Trp, and Leu have been selected by now. According to the 'escaped triplet theory', triplets overrepresented in aptamer binding sites for amino acids became part of the modern genetic code (Yarus et al., 2005). Noteworthy in this regard is that *in vitro* generated RNA aptamers contain – in a statistically important way – anticodonic and, to a lesser degree, codonic binding sites for these amino acids (Caporaso et al., 2005). Although aptamers for Asp and Glu have not yet been selected, it is likely that divalent metal ions could neutralize the repulsion between RNA and these negatively charged amino acids and, ultimately, aptamers will be selected with success (R. Knight, personal communication, email, 2007).

Finally, it is important to mention that RNA molecules can charge amino acids either in *cis* (Illangasekare et al., 1995) or *trans* (Lee et al., 2000). Even the phosphate anhydride activation reaction of amino acids is feasible by RNA (Kumar and Yarus, 2001).

3 The Anticodon Hairpin as the Ancient Adaptor

The simplest form of evolutionary continuity is provided by conservation of the anticodon hairpin (stem and loop) of tRNAs as the most ancient adaptor, which was charged at position 37 of the modern tRNA molecule, adjacent to the 3′-end of the anticodon (Woese, 1972). The CCH hypothesis has adopted this view (Szathmáry, 1996, 1999). As discussed in those papers, such a hairpin offers an ideal transient binding by ribozymes through complementary structures (e.g. in the form of 'kissing' hairpins). Here we deal with two questions: the nature and synthesis of the chemical bond between the adaptor and the amino acid, and the growth of the tRNA molecule to its present form.

As suggested by Woese (1972) and Wong (1991), the link must have been established by a stable *N*-bond. Inspection of relevant contemporary modified bases (Fig. 8) suggests that the nature of this primordial form was through a carbamoyl group.

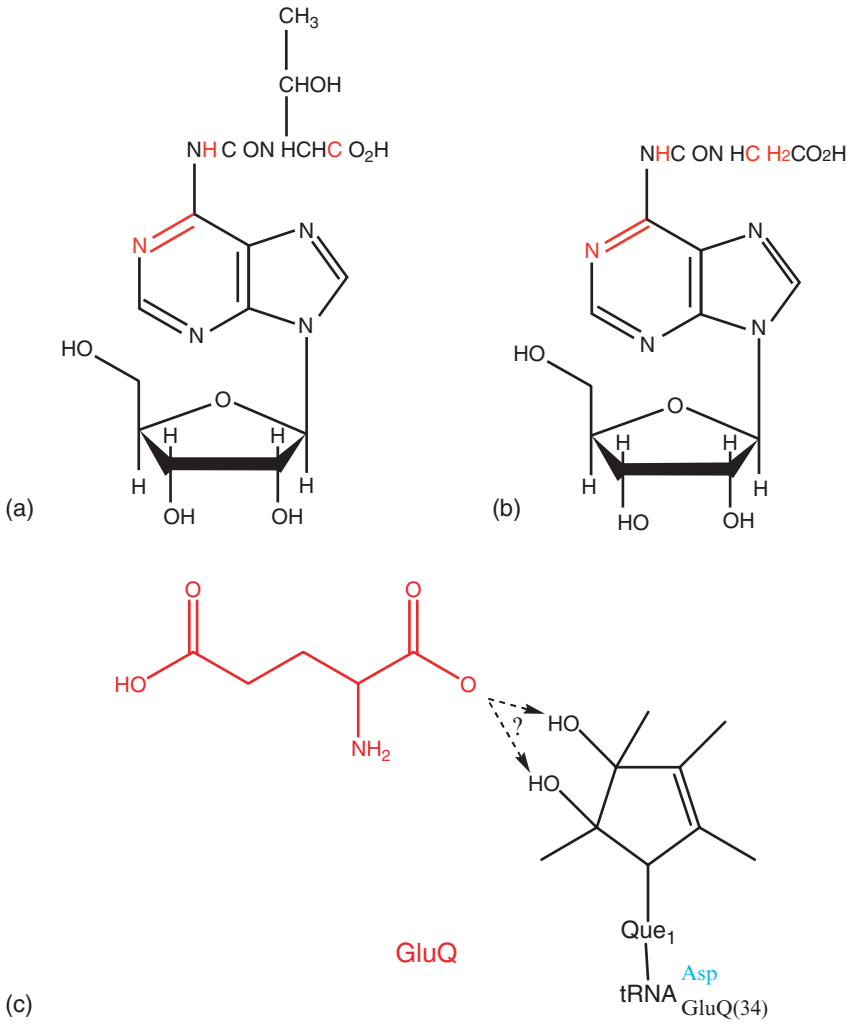

Fig. 8 Nucleosides modified by amino acids in tRNA molecules (from Grosjean et al., 2004). (a) *N*6-threonylcarbamoyladenosine (hn6 A: *N*6-hydroxynorvalylcarbamoyladenosine); (b) *N*6-glycylcarbamoyladenosine; (c) glutamylqueuosine

Regarding the early plausibility of this link, we call attention to the experimental investigations of Taillades et al. (1998) who suggested that *N*-protected *N*-carbamoyl-α-amino acids rather than free α-amino acids formed in the primitive hydrosphere, which serve as eminent starting points also for peptide synthesis. This type of metabolism may have been present even in the RNA world. Remarkably, a formal peptide bond arises by the coupling between the amino acid and adenine through the carbamoyl link!

If the original charging site was in the anticodon loop, then there must have been a stage in evolution when amino acids were still charged to the old position in this loop

and, *at the same time*, already to the new 3′-end of the full tRNA molecule (Szathmáry, 1999). Some of the initial charging persisted as modern tRNA modifications, provided translation evolved using tRNA molecules still charged in the loop (Woese, 1972). Put differently, translation has adapted to these molecules being charged, which partly explains why removal of these modifications disturbs translation today. It is remarkable that the prediction of dual charging of tRNA by synthetases turned out to be correct for $tRNA^{Asp}$ in *Escherichia coli* (Dubois et al., 2004). A paralog of a glutamyl-tRNA synthetase charges Glu to $tRNA^{Asp}$ at position 34 (wobble) to queuosine (Fig. 8c). This example is quite suggestive, even if the chemistry and exact location is different from what we consider important for the CCH scenario. We predict that other synthetase paralogs with tRNA modifying activity will be found in the future. In agreement with the view advocated here, Grosjean et al. (2004, p. 519) comment that 'this modified nucleoside might be a relic of an ancient code'.

We wish to comment further on position 37. It is adjacent to the third base of the anticodon, which is complementary to the first base of the codon. As mentioned above, there is good correlation between first codon base identity and amino acid biosynthetic family membership (Taylor and Coates, 1989). If it is true that position 37 was an important ancient charging site, then the third anticodon base was closest to it. If some of the amino acid transformations (analogous to the modern tRNA: synthetase system) took place in the context of the ancient adaptor–ribozyme synthetase relation, then in agreement with the co-evolution theory (Wong, 1975) we suggest that some amino acids could have been biosynthetically transformed while bound to the ancient adaptor (the anticodon hairpin). The immediate and amino acid-specific neighbourhood on the adaptor would have been position 36, to which the transforming ribozymes would have been sensitive. The correlation of the third anticodon base with biosynthesis could be a relic of the most ancient genetic code and adaptor charging.

We should also explain how the ancient adaptor could have grown to its current size (tRNA). Again, position 37 seems to convey a message. It is between position 37 and 38 that tRNA genes for Glu, His, Gln, and initiator Met has been found in the archeon *Nanoarchaeum equitans* (Randau et al., 2005a,b). The splicing endonuclease splices other intron-coding tRNA genes as well as the transcripts of the two halves of the split tRNA genes (Randau et al., 2005c). The latter is made possible by overhangs at the 3′ and at the 5′-ends of the two transcripts, respectively. This archeon remains the only known organism that functions with split tRNA genes, so this is likely to be a derived (rather than ancestral) phenomenon, in contrast to the interpretation of Di Giulio (2006).

Nevertheless, an adjacent position (between 36 and 37) seems to be important. In Eubacteria there is a Group I self-splicing intron in the same position in several tRNA genes (Reinhold-Hurek and Shub, 1992; Haugen et al., 2005). It is this type of intron that can bind arginine with codonic binding sites (Yarus, 1989). Szathmáry (1993) accepted the idea of Ho (1988) that Group I introns had once been primordial synthetases. This idea still seems promising to us, and experimental attempts at demonstrating such a function in some (mutant) version would be welcome. Preservation of these introns could be due to fortuitous self-insertion (reverse self-splicing) of these molecules into some anticodon loops (Szathmáry, 1993).

Regarding the evolutionary growth of the anticodon hairpin to a full-blown tRNA, we think that ideas resting simply on single, major hairpin duplication (e.g. Widmann et al., 2005) are remote from what we want because half a tRNA is much bigger than a single anticodon loop. This question warrants careful investigation; here we merely call attention to the fact that Bloch et al. (1985) and Nazarea et al. (1985) found statistical evidence for tandem repeats of units of length 8–10 (centred around 9) in both tRNAs and rRNAs. Note that the Group I intron splits the anticodon hairpin into one piece of 10 nucleotides and another one of seven nucleotides. We do not yet know what this could mean. Yet, the most direct evidence for the primitive ancestry of the anticodon arm is that the anticodon arms of tRNAs with complementary anticodons are also complementary, which is not true for the acceptor stem (Rodin et al., 1993).

Last but not least, we point out that the proposed ancestral mechanism for adaptor charging can explain a strange feature found in the acceptor stem. There seems to be a vestigial anticodon–codon pair in the 1-2-3 position, and opposite to in the tRNA acceptor stem (Rodin et al., 1996). The same investigation revealed that there are several tandem repeats in tRNAs, e.g. those of the –DCCA motif (D is the overhanging 'discriminator base' at the 3′end) and its complementary sequence. Szathmáry (1999) presented a somewhat artificial scenario for the evolutionary growth of the anticodon arm to a tRNA with Rodin's anticodon–codon pair in the acceptor stem. Finally, we mention a resolution of the apparent conflict between the primitive ancestry of the anticodon arm and the idea that the anticodon-binding part of present day synthetases is regarded younger than the part binding the acceptor stem (Woese et al., 2000). Whereas ribozymes charged tRNAs both at the old (anticodon) and at the new (acceptor stem) positions by the cognate amino acids, most of the emerging protein synthetases charged them only at the new site.

4 Towards the Appearance of Proteins

One can imagine two ways to build up proteins: (i) to start with a more or less structural role of maybe otherwise 'boring' oligo/polypeptides, which later became complemented by slowly emerging catalytic potential; or (ii) to introduce catalytically highly promising amino acids which later become complemented by structural supports that would ultimately fold without the help of RNA. We prefer the second alternative, since (as explained above) it offers a straightforward way of the appearance of coding before translation, and it provides a substantial and immediate selective advantage in the RNA world.

But it is a good question to ask how single amino acids could have grown to polypeptides in this scenario. Presumably, first dipeptides would appear that would be kept in place after formation of the peptides bond between two adjacent CCH molecules (cf. Szathmáry, 1999), which would then result in a dipeptide bound to one of the adaptors, an intermediate that is identical to Wong's (1991)

peptidyl-tRNA; the other adaptor would be liberated. Further growth can be envisaged under selection for improved enzymatic activity, but then the burning question arises: what would keep the growing polypeptide in a stable/useful conformation? One possibility is that it is binding to the RNA 'scaffold' of the ribozyme, as mentioned previously (Söding and Lupas, 2003). Another would be the *build-up of the smallest possible foldable structures, which are the β-turns stabilized by short β-sheets* (Lesk, 2001). Knowing the opportunistic nature of evolution, we would not exclude either possibility. Following a suggestion by Orgel (1977), Jurka and Smith (1987) argued that *the first β-turns were encoded by RRN codons, which include, with exception of His, all the catalytically most important amino acids!* But the second most important group of amino acids for the β-turns is the YRN group, which includes His, and the two add up to NRN, the last two columns of the code (Fig. 3). Dendrogram Fig. 6c confirms this idea strongly. Amino acids with the NRN pattern are also *mediocre* α-helix *and* β-sheet builders (Fig. 6b), so experimentation in that direction was not totally excluded either. Data (Prevelige and Fasman, 1989) show that proline is the third strongest loop builder, which makes it the only real exception to the NRN rule (Ser has also an AGY codon).

It is known that one can make enzymes with fewer than 20 amino acids. Walter et al. (2005) managed to evolve a chorismate mutase built of nine amino acids only: Arg, Asp, Glu, Asn, Lys, Phe, Ile, Leu, and Met. Noting the redundancies Asp/Glu and Ile/Leu, the enzyme could be probably even more simplified in the future. Remarkably, its active site (Fig. 9) is built of RRN codon-type

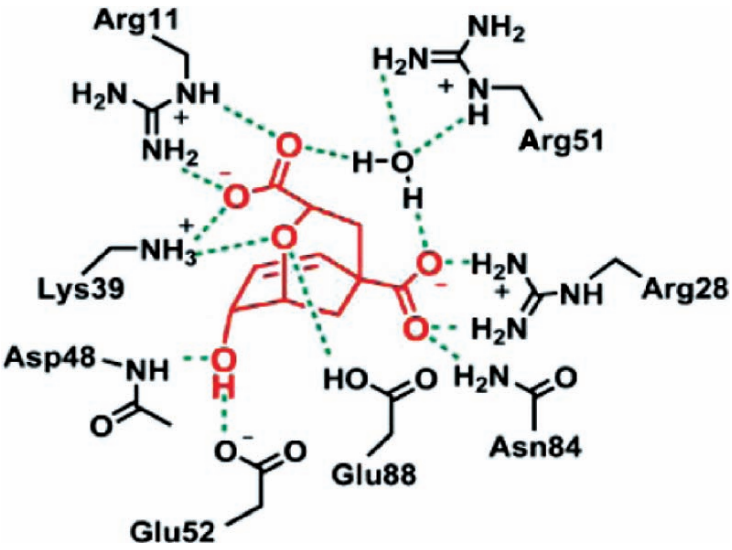

Fig. 9 Proposed active site of a chorismate mutase built of just nine amino acids. (Walter et al., 2005)

amino acids only! Adding amino acids of the NUN codon-type allows the set to build proper α-helices; amino acids with NCN codons are not required at all. Evolution of this enzyme is in rather good agreement with the theoretical estimate that the minimum number of amino acids to fold a protein is around ten (Fan and Wang, 2003).

The next amino acids, to further stabilize the β-turns with β-sheets were amino acids with NYN codons (Jurka and Smith, 1987). The proposal by Di Giulio (1996) that the genetic code was driven by the need to form β-sheets is in our view secondary to the catalytic propensity/β-turn-driven primary evolution, but independently important to build the scaffolds for the catalytic structures. But selection for structures in this order gives α-helices for free since the code is virtually complete. We suggest that *the multiplicity of catalytically boring amino acids in the genetic code is explained by selection for fine-tuning of the 3D structure of the scaffolds to optimize the geometric arrangements of the active sites*.

This scenario is supported by the correlation coefficients between the amino acids properties (Table 1).

There is a weak negative correlation between catalytic and α-helix propensity, so the latter structures cannot arise based on amino acids selected for catalysis; it is easiest to go for the β-turns. From these, one can go by evolution (amino acid vocabulary extension) in the direction of either the β-sheets (slightly favoured) or the α-helices, but presumably not in both.

We performed a network analysis of the BLOSUM amino acid substitution matrix (Henikoff and Henikoff, 1992) in order to see along which lines amino acid vocabulary extension/replacement could have been most likely (Fig. 10).

The most common substitutions occur within the (Lys, Arg), (Ile, Val), and (Phe, Tyr, Trp) sets. The firs dyad is clearly a catalytically very important one, and this is another reason why we suggest that Arg replaced Lys before LUCA. Note the remarkable role of His in these plots also. The catalytically most important amino acid His builds the bridge via the Tyr-His-(Asn, Gln) link between the catalytically unimportant and important (internally well connected) clusters (Fig. 10c,d)! We have specific RNA aptamers (Caporaso et al., 2005) for the whole bridge (none for Asn, but the bridge is still functional via Gln), so we suggest that it has been built in the RNA world.

Table 1 Correlation between pairs of amino acid properties related to the construction of active sites and scaffolds

Trait pair	Correlation coefficient
Activity, α-helix	–0.15049
Activity, β-sheet	0.31893
Activity, β-turn	0.38407
α-helix, β-sheet	0.42399
α-helix, β-turn	0.64948
β-sheet, β-turn	0.67883

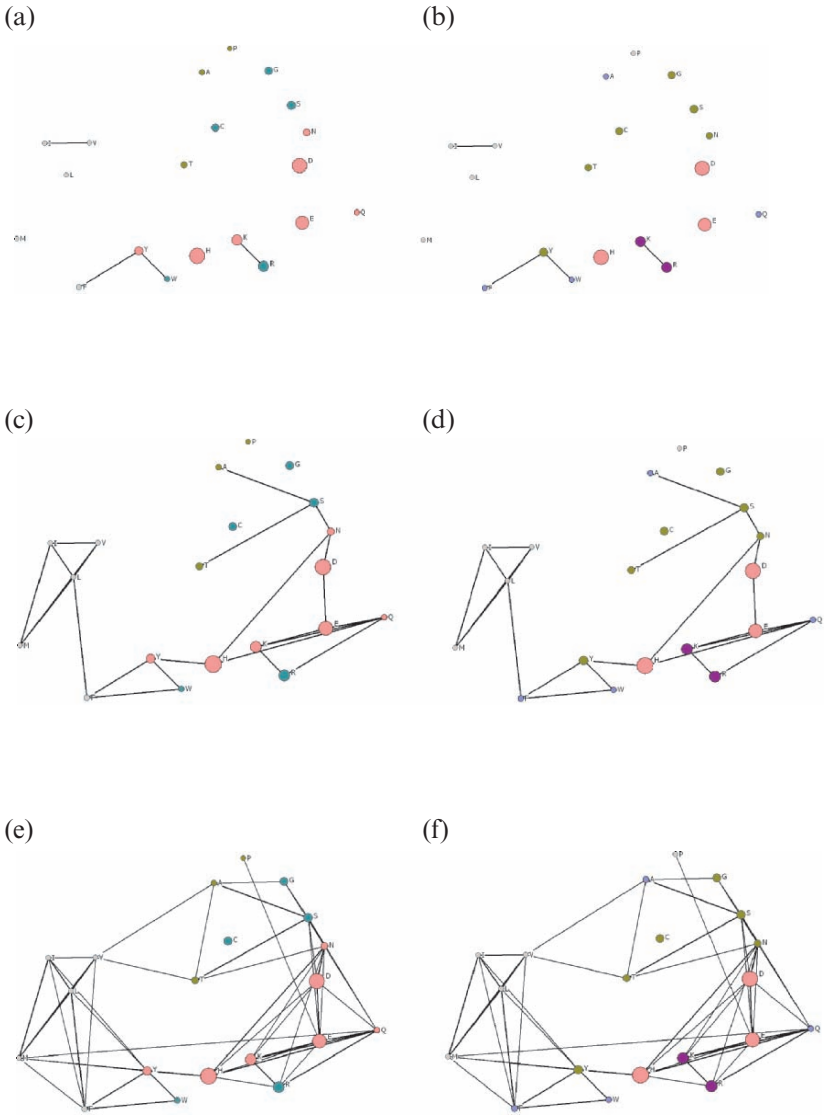

Fig. 10 Connectivity of the amino acid substitution network based on the BLOSUM 62 matrix (Henikoff and Henikoff, 1992). Colour codes of amino acids refer to those of Fig. 7a (a, c, e) and Fig. 7b (b, d, f). Substitution data have been transformed: we added 6 to the originals given by Henikoff and Henikoff (1992). For clarity, loops given in the main diagonal are not shown. We illustrate the network with different lower thresholds (minimal frequency values) for the transformed substitution data: 9 in a and b, 7 in c and d, and 6 in e and f (9 is the strongest value, i.e. it was equal to 3 in the original data matrix). Note that the network is undirected. Drawn by UCINET. (From Borgatti et al., 2002)

5 Towards an Experimental Test of the CCH Hypothesis with Catalytically Important Amino Acids

As noted before, the strongest experimental support in favour of the 'amino acids as cofactors' idea is the successful histidine-dependent RNA-cleaving enzyme made of DNA (Roth and Breaker, 1998). One would like to demonstrate the usefulness of the CCH hypothesis by the evolved usage by ribozymes of amino acids linked to handles (i.e. the free anticodon arm).

To this end one could repeat the above experiment with histidine linked to position 37 of $tRNA^{His}$ by the carbamoyl bond. For this and the forthcoming experiments we suggest usage of the *in vitro* compartment selection technique of Griffiths (Agresti et al., 2005).

A perhaps more appealing test would be to go for a metabolic reaction. The successful in vitro selection of an NADH-dependent alcohol dehydrogenase ribozyme by Tsukiji et al. (2004) is an excellent starting point (Fig. 11); it is also highly relevant for the idea of a metabolically complex RNA world (Benner et al., 1989).

First, a variant of the ribozyme that would work in *trans* should be selected in compartments for a different substrate, such as malate. Malate dehydrogenase has Asp and His in its active centre. In the second step ribozymes that could utilize

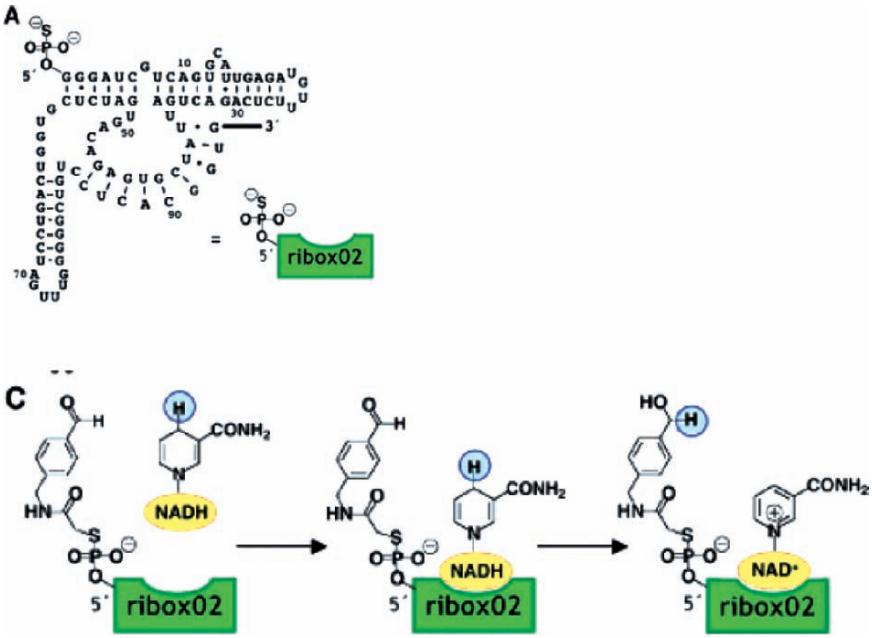

Fig. 11 Structure and action (on benzaldehyde) of an alcohol dehydrogenase ribozyme. (From Tsukiji et al., 2004)

either of these amino acids in free form should be selected. Finally, ribozymes that could use the CCH-amino acid should be selected. If successful, all the ribozymes should be compared for structure and activity. Ultimately a trial using both amino acids should be attempted. This would be an exciting experiment for the origin of the genetic code and the increase in biological complexity.

Acknowledgements This work was supported by the Hungarian Scientific Research Fund (D048406) and by the National Office for Research and Technology (NAP 2005/KCKHA005). We thank L. Patthy for discussion. The help of Paolo Sonego with hierarchical clustering is gratefully acknowledged. FJ is fully supported by Society in Science: The Branco Weiss Fellowship, ETH Zürich, Switzerland.

References

Agresti JJ, Kelly BT, Jaschke A, Griffiths AD (2005) Selection of ribozymes that catalyse multiple turnover Diels-Alder cycloadditions by using *in vitro* compartmentalization. Proc Natl Acad Sci USA 102:16170–16175

Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324:105–121

Benner SA, Ellington AD Tauer A (1989) Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci USA 86:7054–7058

Berg JM, Tymoczko JL, Stryer L (2003) Biochemistry, Fifth Edition. W. H. Freeman, San Francisco

Bloch DP, McArthur B, Mirrop S (1985) tRNA-rRNA sequence homologies: evidence for an ancient modular format shared by tRNAs and rRNAs. BioSystems 17:209–225

Borgatti SP, Everett MG, Freeman LC (2002) Ucinet for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard

Caporaso JG, Yarus M, Knight R (2005) Error minimization of coding triplet/binding site associations are independent features of the canonical genetic code. J Mol Evol 61:597–607

Crick FHC, Brenner S, Klug A, Pieczenik G (1976) A speculation on the origin of protein synthesis. Orig Life 7:389–397

Di Giulio M (1996) The β-sheets of proteins, the biosynthetic relationships between amino acids, and the origin of the genetic code. Orig Life Evol Biosph 26(6):589–609

Di Giulio M (2006) *Nanoarchaeum equitans* is a living fossil. J Theor Biol 242:257–260

Dubois DY, Blaise M, Becker HD, Campanacci V, Keith G, Giege R, Cambillau C, Lapointe J, Kern D (2004) An aminoacyl-tRNA synthetase-like protein encoded by the *Escherichia coli yadB* gene glutamylates specifically $tRNA^{Asp}$. Proc Natl Acad Sci USA 101:7030–7035

Fan K, Wang W (2003) What is the minimum number of letters required to fold a protein? J Mol Biol 328:921–926

Freeland SJ, Knight RD, Landweber LF, Hurst LD (2000) Early fixation of an optimal genetic code. Mol Biol Evol 17:511–518

Gilbert W (1986) The RNA world. Nature 319:618

Grosjean H, De Crécy-Lagard V, Björk GR (2004) Aminoacylation of the anticodon stem by a tRNA-synthetase paralog: relic of an ancient code? Trends Biochem Sci 29:519–522

Hartigan JA (1975) Clustering Algorithms, Wiley, New York

Haugen P, Simon DM, Bhattacharya D (2005) The natural history of group I introns. Trends Genet 21:111–119

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Ho CK (1988) Primitive ancestry of transfer RNA. Nature 333:24

Illangasekare M, Sanchez G, Nickles T, Yarus M (1995) Aminoacyl-RNA synthesis catalyzed by an RNA. Science 267:643–647

Jurka J, Smith TF (1987) β-turn driven early evolution: the genetic code and biosynthetic pathways. J Mol Evol 25:151–159

Porter CT, Bartlett GJ, Thornton JM (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acid Res. 32:D129–D133

Knight RD, Landweber LF (2000) Guilt by association: the arginine case revisited. RNA 6:499–510

Kumar RK, Yarus M (2001) RNA-catalyzed amino acid activation. Biochemistry 40:6998–7004

Lee N, Bessho Y, Wei K, Szostak JW, Suga H (2000) Ribozyme-catalyzed tRNA aminoacylation. Nat Struct Biol 7:28–33

Lesk AM (2001) Introduction to Protein Architecture. Oxford University Press, Oxford

Maynard Smith J, Szathmáry E (1995) The Major Transitions in Evolution. Freeman, Oxford

Miller SL (1986) Current status of the prebiotic synthesis of small molecules. Chem Scr 26B:5–11

Moore PB, Steitz TA (2002) The involvement of RNA in ribosome function. Nature 418:229–235

Muñoz V, Serrano L (1994) Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices. Comparison with experimental scales. Proteins 20:301–311

Nazarea AD, Bloch DP, Semrau AC (1985) Detection of a fundamental modular format common to transfer and ribosomal RNAs: second-order spectral analysis. Proc Natl Acad Sci USA 82:5337–5341

Orgel LE (1977) *β*-Turns and the evolution of protein synthesis. In: Bradbury EM, Javaherian K (eds) The Organization and Expression of the Eukaryotic Genome. Academic Press, London, pp. 499–504

Prevelige JP, Fasman G (1989) Chou-Fasman prediction of the secondary structure of proteins: the Chou-Fasman-Prevelige algorithm. In: Fasman G (ed.) Prediction of Protein Structure and the Principles of Protein Confoimation. Plenum, New York, pp. 391–416

Randau L, Munch R, Hohn MJ, Jahn D, Söll D (2005a) *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5′- and 3′-halves. Nature 433:537–541

Randau L, Pearson M, Söll D (2005b) The complete set of tRNA species in *Nanoarchaeum equitans*. FEBS Lett 579:2945–2947

Randau L, Calvin K, Hall M, Yuan J, Podar M, Li H, Söll D (2005c) The heteromeric *Nanoarchaeum equitans* splicing endonuclease cleaves noncanonical bulge-helix-bulge motifs of joined tRNA halves. Proc Natl Acad Sci USA 102:17934–17939

R Development Core Team (2004) A Language And Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria

Reinhold-Hurek B, Shub DA (1992) Self-splicing introns in tRNA genes of widely divergent bacteria. Nature 357:173–176

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Rodin S, Ohno S, Rodin A (1993) Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors?

Proc Natl Acad Sci USA 90:4723–4727

Rodin S, Rodin A, Ohno S (1996) The presence of codon-anticodon pairs in the acceptor stem of tRNAs. Proc Natl Acad Sci USA 93:4537–4542

Roth A, Breaker RR (1998) An amino acid as a cofactor for a catalytic polynucleotide. Proc Natl Acad Sci USA 95:6027–6031

Shen C, Yang L, Miller SL, Oró J (1990) Prebiotic synthesis of histidine. J Mol Evol 31:167–174

Söding J, Lupas AN (2003) More than the sum of their parts: on the evolution of proteins from peptides. BioEssays 25:837–846

Steitz TA, Moore PB (2003) RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem Sci 28:411–418

Szathmáry E (1990) Useful coding before translation: the coding coenzymes handle hypothesis for the origin of the genetic code. In: Lukács B. et al. (eds) Evolution: from Cosmogenesis to Biogenesis. KFKI-1990-50/C, Budapest, pp. 77–83
Szathmáry E (1991) Four letters in the genetic alphabet: a frozen evolutionary optimum? Proc R Soc Lond B 245:91–99
Szathmáry E (1992) What determines the size of the genetic alphabet? Proc Natl Acad Sci USA 89:2614–2618
Szathmáry E (1993) Coding coenzyme handles: A hypothesis for the origin of the genetic code. Proc Natl Acad Sci USA 90:9916–9920
Szathmáry E (1996) Coding coenzyme handles and the origin of the genetic code. In: Müller A, Dress A, Vögtle F (eds) From Simplicity to Complexity in Chemistry – and Beyond. Part I. Vieweg, Braunschweig, pp. 33–41
Szathmáry E (1999) The origin of the genetic code: amino acids as cofactors in an RNA world. Trends Genet 15:223–229
Szathmáry E (2007) Coevolution of metabolic networks and membranes: the scenario of progressive sequestration. Phil Trans R Soc B DOI: 10.1098/rstb.2007.2070
Taillades J, Beuzelin I, Garrel L, Tabacik V, Bied C, Commeyras A (1998) *N*-carbamoyl-alpha-amino acids rather than free alpha-amino acids formation in the primitive hydrosphere: a novel proposal for the emergence of prebiotic peptides. Orig Life Evol Biosph 28:61–77
Taylor FJ, Coates D (1989) The Code within the codons. Biosystems 22:177–187
Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119:205–218
Tsukiji S, Pattnaik SB, Suga H (2004) Reduction of aldehyde by a NADH/Zn^{2+}-dependent redox active ribozyme. J Am Chem Soc 126:5044–5045
Walter KU, Vamvaca K, Hilvert D (2005) An active enzyme constructed from a 9-amino acid alphabet. J Biol Chem 45:37742–37746
Wetzel R (1995) Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code. J Mol Evol 40:545–550
Widmann J, Di Giulio M, Yarus M, Knight R (2005) tRNA creation by hairpin duplication. J Mol Evol 61:524–530
Woese CR (1972) The emergence of genetic organization. In: Ponnamperuma, C. (ed.) Exobiology. North-Holland Publishing, Amsterdam, The Netherlands, pp. 301–341
Woese CR, Olsen GJ, Ibba M, Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64:202–236
Wong JT (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912
Wong JT (1991) Origin of genetically encoded protein synthesis: a model based on selection for RNA peptidation. Orig Life Evol Biosph 21:165–176
Wong JT Bronskill PM (1979) Inadequacy of prebiotic synthesis as origin of proteinous amino acids. J Mol Evol 13:115–125
Yarus M (1989) Specificity of arginine binding by the *Tetrahymena* intron. Biochemistry 28:980–988
Yarus M, Caporaso JG, Knight R (2005) Origins of the genetic code: the escaped triplet theory. Annu Rev Biochem 74:179–198
Zar JH (1998) Biostatistical Analysis (4th Edition). Prentice Hall