J Mol Evol (2001) 53:724–732
DOI: 10.1007/s002390010259

JOURNAL OF MOLECULAR EVOLUTION

A Blind Empiricism Against the Coevolution Theory of the Origin of the Genetic Code

Massimo Di Giulio

International Institute of Genetics and Biophysics, CNR, Via G. Marconi 10, 80125 Naples, Napoli, Italy

Received: 21 February 2001 / Accepted: 22 May 2001

Abstract. Ronneberg et al. (Proc Natl Acad Sci USA 97:13690–13695, 2000) recently suggested abandoning the coevolution theory of genetic code origin on the basis of two pieces of evidence. They (1) criticize the use of several pairs of amino acids in a precursor–product relationship to support this theory and (2) suggest a new set of codes in which to investigate the statistical bases of the coevolution theory, reaching the conclusion that this theory is not statistically validated in this set. In this paper I critically analyze the robustness of these conclusions. Observations and arguments lead to the belief that the pairs of amino acids in a precursor–product relationship originally used by the coevolution theory are such, or may at least be interpreted as such, and are therefore a manifestation of this theory. Furthermore, the new set of codes that Ronneberg et al. suggest is open to criticism and is thus substituted by the set of amino acid permutation codes, in which even the pairs of amino acids they favor end up by supporting the coevolution theory. Overall, the analysis seems to show that the paper by Ronneberg et al. is of minor scientific value while the coevolution theory seems to be one of the best theories at our disposal for explaining the evolutionary organisation of the genetic code and is, contrary to their claims, statistically well validated.

Key words: Coevolution — Biosynthetic relationships between amino acids — Evolution of the codon subdomain in the code — Hypergeometric distribution

Correspondence to: Dr. M. Di Giulio: *email:* digiulio@iigb. na.cnr.it

The Coevolution Theory and the Aim of the Analysis

In 1975 Wong introduced a clear idea with which to read the evolution of genetic code organization. He hypothesized (Wong 1975) that in the early stage of genetic code origin, only a few amino acids (precursors) were codified within the code. As other amino acids (products) evolved from these along biosynthetic pathways, part or all of the precursor amino acid's codon domain was conceded to the product amino acids. The mechanism by which codons were conceded by the precursor to the product amino acid was assumed to have been mediated by tRNA-like molecules on which the theory envisages that the biosynthetic transformation of amino acids must have taken place (Wong 1975). In other words, if the metabolic transformation from precursor to product amino acid took place on a tRNA-like molecule, then this tRNA, which clearly recognized part of the precursor amino acid's codon domain, was able to be naturally conceded to the product amino acid in the evolving genetic code.

If this effectively was the mechanism that led to the evolutionary structuring of genetic code organization (Wong 1975), then the code itself must reflect, in general, the biosynthetic relationships between amino acids and, in particular, many pairs of amino acids that are in a clear, unambiguous precursor–product relationship. The literature contains a large number of papers linking genetic code organization to the biosynthetic relationships between amino acids (Dillon 1973; Wong 1975, 1976, 1988; McCledon 1986; Miseta 1989; Taylor and Coates 1989; Di Giulio 1996, 1997a, 1999, 2000;

Morowitz 1992) and, in particular, to the amino acid pairs in a precursor–product relationship (Wong 1975; Danchin 1989; de Duve 1991; Di Giulio 1991, 1996; Tumbula et al. 2000).

Recently Ronneberg et al. (2000) (1) criticized several pairs of amino acids in a precursor–product relationship originally used by Wong (1975), which they claim are not in such a relationship, and (2) defined a new set of codes in which to assay the statistical validity of the coevolution theory, reaching the conclusion that this theory cannot adequately explain the organization of the genetic code. Here I critically analyze the robustness of Ronneberg et al.'s conclusions.

Methods

The use of the hypergeometric distribution to establish whether or not a certain number of amino acid pairs in the genetic code is statistically significant was introduced by Wong (1975) and has also been used by Di Giulio (1991) and by Ronneberg et al. (2000). The probability that can be associated with a pair of amino acids in a precursor–product relationship is given by

$$P = \sum_{x}^{n} a!/(a-x)!x! \cdot b!/(b-n+x)!(n-x)! \cdot (a+b-n)!n!/(a+b)!$$

where a is the number of codons contiguous to the precursor codons; b is the number of noncontiguous codons, i.e., differing in more than one base from the codons specifying for the precursor amino acid; n is the number of codons codifying for the product amino acid; and x is the number of codons in the product that are contiguous to those in the precursor, i.e., differing in only one base from the latter. To complete the statistical test it must be remembered that the variable $-21\mathrm{n}P$ is distributed according to a χ^2 value with 2 degrees of freedom (df) (Fisher 1950).

The other method used to establish whether or not a certain number of amino acid pairs in the genetic code is statistically significant was introduced by Di Giulio and Medugno (2000). Briefly, in the set of amino acid permutation codes (Di Giulio 1989), i.e., codes that leave unchanged the arrangement of synonymous codon blocks as observed in the genetic code, 100 million random codes can, for instance, be generated (Di Giulio and Medugno 2000). The codon correlation score (CCS) (Amirnovin 1997; Di Giulio and Medugno 2000), a simple additive measurement (see Table 2, footnote a), is calculated for each of these codes using all the amino acid pairs whose statistical significance is to be determined (Di Giulio and Medugno 2000). In this way a frequency distribution is constructed for the CCS value in which we are interested, and the number of amino acid pairs in an actual precursor–product relationship can be used to determine the statistical significance for the set of amino acid pairs considered (Di Giulio and Medugno 2000).

Results and Discussion

The Amino Acid Pairs in a Precursor–Product Relationship Contested by Ronneberg et al.

The Thr–Met Pair

Ronneberg et al. (2000) claim that Thr cannot be the precursor of Met as assumed in the coevolution theory (Wong 1975) because, a considerable energy barrier, consisting of the conversion of Thr into homoserine (Fig. 1a), needs to be overcome to transform Thr into Met. This is certainly true. However, Wong (1975) acknowledged both that Met can be synthesized better by Asp than by Thr and that homoserine might have been a more primitive form of Thr (Thr and homoserine are isomers), but he considered Thr as a precursor of Met on the basis of the number of enzymatic steps. I believe that the latter choice was mistaken. Nevertheless, if we consider that homoserine is the first amino acid to be synthesized by Asp in this pathway (Fig. 1a), then on the basis of the coevolution theory postulates, we can believe, as Wong (1975) also recognized, that Asp might have conceded, as an intermediate stage, part of its codon domain to homoserine. Figure 2 shows this possible and, according to the coevolution theory, probable evolution. At the initial stage the codons AUN and ACN codified for Asp (Fig. 2). These codons were then conceded to homoserine, which later conceded the majority of them to Thr, conserving only one (or perhaps two) for itself. This codon was finally conceded to Met, but prior to this stage, Thr conceded part of its codon domain to IIe (Fig. 2). [I am impressed by the accuracy with which the coevolution theory (Wong 1975) manages to scan the evolution of this codon subdomain.]

This evolutionary pattern helps to clarify the further criticisms that Ronneberg et al. (2000) direct at the Thr–Met pair, which entail considering Met as the product of either Ser or Cys, as these amino acids intervene in the biosynthetic pathway of Met only after the formation of homoserine (Fig. 1a), when Asp had already conceded part of its codon domain to it (Fig. 2). According to the number of biosynthetic steps, Ser and Cys are, indeed, closer to Met (Fig. 1a), but as Wong (1975) suggested, neither Ser nor Cys are straightforward precursors of Met, i.e., most of the Met molecule is built with the atoms of the homoserine molecule. Therefore, from this viewpoint, it is more correct to consider homoserine, and not Ser or Cys, as the precursor of Met (see also following section).

Hence, although from an energy point of view it is incorrect to consider the Thr–Met pair as being in a precursor–product relationship, the probable (according to the coevolution theory) disappearance of homoserine from the genetic code places (from the formal standpoint) this pair as if in a real precursor–product relationship (Fig. 2).

However, in view of these uncertainties, Wong (1975) does not include the Thr–Met pair in the set of certain precursor–product amino acid pairs used to provide a statistical basis for the coevolution theory, while Ronneberg et al. (2000) include the Asp–Met pair in this set.

The Glu–Arg Pair

Ronneberg et al. (2000) claim that the Glu–Arg pair of amino acids in a precursor–product relationship used

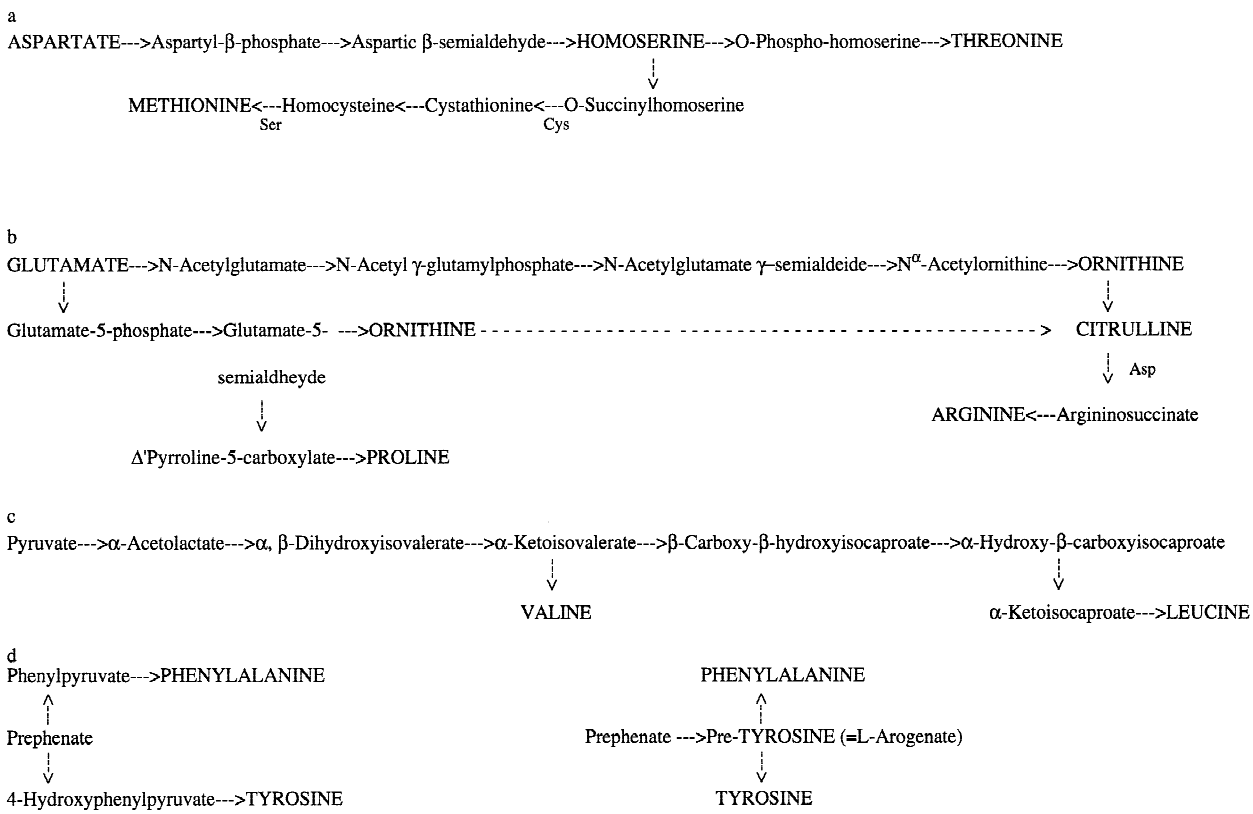

Fig. 1. A schematic representation of four biosynthetic pathways linking amino acids in a precursor–product relationship or in a strict biosynthetic relationship. In three reactions in which amino acids are involved (Ser, Cys, and Asp), these are also indicated. All the pathways were represented after consulting Greenberg (1969) and Voet and Voet (1990) except in **d,** where Jensen and Fisher (1987) was used.

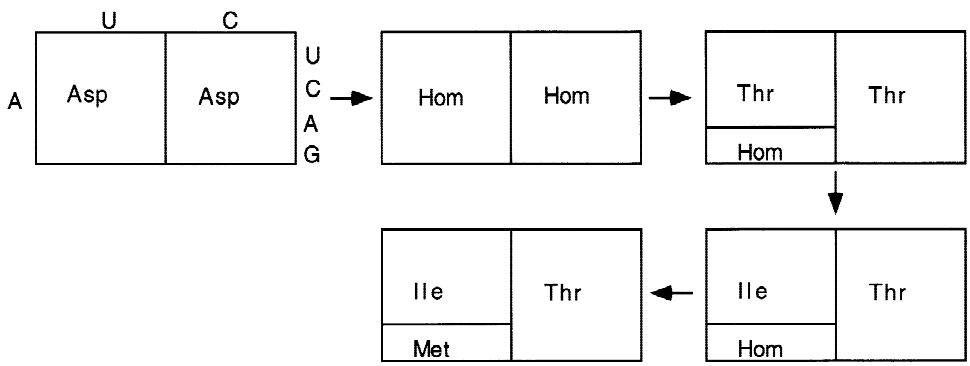

Fig. 2. The five stages of the evolution of the codons AUN and ACN could be hypothesized by means of the coevolution theory. In the third stage homoserine (Hom) could conserve two codons instead of one as indicated. Only in the final stage, with the entry of Met and as a secondary mechanism, might Ile have taken over this codon. See text for further comments.

by Wong (1975) must be replaced by the Asp–Arg pair because Asp intervenes in the conversion of citrulline (Cit) into argininosuccinate and is therefore closer to Arg (Fig. 1b). Although all the texts of biochemistry consider Arg as an amino acid from the Glu family (see, e.g., Voet and Voet 1990) Ronneberg et al.'s criticism is not unfounded but may be solved in the framework of the coevolution theory. As already suggested, ornithine (Orn) might have entered the genetic code before Arg (Jukes 1973; Wachtershauser 1988), and I could add that, on the basis of Orn's instability in proteins and in Orn-tRNA (Wachtershauser 1988), Cit might have been codified in the code before Arg. If this is true, then part of the Glu codon domain was conceded to Orn or Cit before Arg appeared. This therefore explains why Arg must have codons from the Glu domain, and not from that of Asp, because the latter intervenes subsequently to the appearance of Orn or Cit in the biosynthetic pathway leading to Arg (Fig. 1b), i.e., when there had already been a hypothetical concession of codons from Glu to Orn or Cit. This case is thus equivalent to that in the previous section.

In conclusion, although in both cases it can be claimed that the Ser–Met and Asp–Arg pairs are valid on the basis of the low number of enzymatic steps separating the presumed precursor from the product (Ronneberg et al. 2000) (Figs. 1a and b), this would be incorrect on the basis of the number of atoms in the molecule of the precursor ending up in that of the product. It is thus clear that the use of rigid definitions as adopted by Ronneberg et al. (2000), but not by Wong (1975), may be extremely dangerous because this could lead to the rejection of a theory that might well be substantially correct.

The Val–Leu Pair

Ronneberg et al. (2000) eliminate the Val–Leu pair from the set of precursor–product pairs used by Wong (1975) because they are not in a clear precursor–product relationship but are produced from alternative branches of a common intermediate (Ronneberg et al. 2000) (Fig. 1c). Wong (1975) also considers Val–Leu more as a pair of sibling amino acids than as being in a true precursor–product relationship.

The transformation of a-katoisovalerate into Val is a transamination reaction. In theory, the constant equilibrium for transamination reactions should be approximately equal to one (Greenberg 1969, p. 21), and hence, at equilibrium, for every molecule of a-katoisovalerate there will be a molecule of Val. This may imply that Val molecules can transform into a-katoisovalerate molecules, which, in turn, can change into Leu molecules. There is no energy barrier to overcome. Moreover, if all this took place on tRNA-like molecules, as envisaged by the coevolution theory (Wong 1975), then Val can clearly be considered a "true" precursor of Leu because in this way Val might have conceded a tRNA to Leu. In conclusion, Val can be considered a true precursor of Leu because there are no energy barriers to overcome in converting Val into a-katoisovalerate and then into Leu (Fig. 1c).

There is a general argument involving the mechanism on which the coevolution theory is founded, which leads to the conclusion that Ronneberg et al.'s elimination of the Val–Leu pair is mistaken. Although the coevolution theory (Wong 1975) is essentially based on the concept of amino acids in a precursor–product relationship, it is not incompatible with sibling pairs of type Val–Leu or Phe–Tyr (Figs. 1c and d) because there must necessarily be some amino acids whose precursors are non-amino acid precursors. Therefore, the same mechanism that was operative for the first amino acids entering the code might also have been operative for these pairs. For example, Wong (1975) considers the Asp–Glu pair as either a sibling pair or a precursor–product pair. Since, according to the coevolution theory, the latter two amino acids entered the genetic code early on (Wong 1975), the same mechanism that enabled their entry into the code might also have been operative for pairs of type Val–Leu or Phe–Tyr. Specifically, if at least the whole metabolism of amino acids took place on tRNA-like molecules, as the mechanism on which the coevolution theory is based seems to suggest (Danchin 1989; Di Giulio 1997a), then it is sufficient that two amino acids be in a strict biosynthetic relationship (although not necessarily in a clear precursor–product relationship) for this theory to be supported. This is because, if the biosynthetic pathway took place on a tRNA and if this tRNA can occupy codons in the genetic code by means of an amino acid, then all amino acids evolving in this pathway (although by hypothesis not in a clear product relationship with this amino acid) could occupy contiguous codons as they are synthesized on this tRNA and, thus, on a "sibling" tRNA. The important thing here is that we can provide clear examples of amino acids in a certain, unambiguous precursor–product relationship, and this is certainly the case (see, for instance, the pairs of types Asp–Asn and Glu–Gln).[1]

The Phe–Tyr Pair

Ronneberg et al. (2000) criticize the Phe–Tyr pair because they are not in a clear precursor–product relationship (Fig. 1d, left). However, they preserve this pair in the statistical analysis because Tyr is synthesized by Phe in a single step in its degradation pathway. Nevertheless, from the coevolution theory's point of view, there is a much more valid reason for considering the Phe–Tyr pair as useful for establishing the statistical bases of this theory in addition to the more general argument mentioned in the previous section.

Besides the pathway leading from prephenate, the common intermediate, to the synthesis of Phe and Tyr (Fig. 1d, left), there is another pathway that leads from prephenate to pretyrosine (pre-Tyr) and thus to Tyr and Phe (Fig. 1d, right) (Srenmark et al. 1974; Jensen and Fisher 1987). The latter pathway is considered the more ancient (Jensen and Stenmark 1975; Jensen and Fisher 1987; Wachtershauser 1988). We can therefore hypothesize, on the basis of the coevolution theory, that the codons that now codify for Phe and Tyr in the genetic code were assigned to pre-Tyr at an intermediate stage of its evolution. Only later were these four codons codifying for pre-Tyr conceded by this amino acid to Phe and Tyr. This is a further way of removing the difficulties that derive from considering Phe and Tyr as amino acids in a precursor–product relationship and suggests their use in statistical analysis, as they both become product amino acids of pre-Tyr and must therefore have similar codons.

[1] It is worth remembering that the latter point of view is used as the basis for a statistical test which provides a high statistical significance ($P = 6 \times 10^{-5}$) (Di Giulio and Medugno 2000) for the coevolution theory, although it has been only and mistakenly interpreted by Ronneberg et al. (2000) as favoring the observation that amino acids in a biosynthetic relationship tend to have codons with the same first base.

The Gln–His Pair

Ronneberg et al. (2000) criticize Wong's use of the Gln–His pair because Gln donates a single nitrogen atom to His, but at the same time, they consider Arg as a product of Asp even though the latter still donates a single nitrogen atom to Arg (Greenberg 1969, pp. 72–73). Clearly in considering two amino acids to be in a precursor–product relationship, there is both a quantitative and a qualitative aspect. We place great confidence in considering an amino acid as the precursor of another if the majority of the atoms in the former's molecule (the precursor's) end up in the latter's molecule (the product's). Nevertheless, the qualitative aspect is also extremely important. If we accept the general postulates of the coevolution theory, we must also accept the possibility that an amino acid may be the precursor of another even if it donates only a single atom. However, in the pair under investigation, it is unusual that in the complex biosynthetic pathway leading to His (Greenberg 1969, pp. 270–272) the only amino acid that seems to intervene specifically is Gln. All this seems to indicate that it is not arbitrary to consider Gln as a precursor of His.

The Statistical Significance of the Coevolution Theory in the Code Set Subject to the NNY Constraint

Ronneberg et al. (2000) define a set of codes different from those used by Wong (1975) which considers the NNY codons as if they were a single unit and not two, as Wong does, when they apply the hypergeometric distribution (see next section).

In view of the considerations made in the preceding sections, if we replace Asp–Met with Val–Leu in the eight main pairs on which Ronneberg et al. (2000) conduct their analysis, we get a significant χ^2 value ($\chi^2 = 30.49$, df $= 16$, $P = 0.016$; Table 1 the first eight pairs). Here the important thing is to include the Val–Leu pair rather than to exclude the Asp–Met pair, whose inclusion in the calculation does not result in major variations to the overall significance. Therefore, the nonsignificance ($P = 0.168$) reported by Ronneberg et al. (2000) is based solely on the exclusion of the Val–Leu pair, which, as we have seen above, is a pair that certainly favors the coevolution theory. In conclusion, the eight pairs of amino acids (Table 1) originally used by Wong (1975) maintain a certain significance in the set with the NNY constraint, although it is lower than that obtained in the set of codes not subject to this constraint (Wong 1975).

The next step for Ronneberg et al. (2000) was to include four other amino acid pairs in their analysis (see their Table 4). These pairs were introduced under the condition that the codons AAY and CAR codified in the evolving code for Asp and Glu, respectively, as envisaged by the coevolution theory (Wong 1975). I have performed an equivalent calculation, but in view of the considerations made in the previous sections, I substituted the Asp–Arg pair with Glu–Arg and I obtained a significant χ^2 value ($\chi^2 = 39.50$, df $= 24$, $P = 0.024$; Table 1, all pairs). Here, too, the significance depends more on the Val–Leu pair than on the substitution of the Asp–Arg pair with Glu–Arg, which causes only a very small variation in significance.

Table 1. The results of the application of the hypergeometric distribution to the indicated pairs in the code set obtained considering the NNY constraint (Ronneberg et al. 2000)[a]

Precursor → product	x	n	a	b	P	$-21n P$
Ser → Trp	1	1	21	20	0.512	1.34
Ser → Cys	1	1	21	20	0.512	1.34
Phe → Tyr	1	1	8	36	0.182	3.41
Thr → IIe	2	2	18	24	0.178	3.46
Gln → His	1	1	11	32	0.256	2.73
Glu → Gln	2	2	11	32	0.0609	5.60
Asp → Asn	1	1	8	36	0.182	3.41
Val → Leu	5	5	18	24	0.0101	9.20
						$\chi^2 = 30.49$
Asp → Lys	2	2	12	31	0.0731	5.23
Glu → Pro	2	3	16	25	0.334	2.19
Glu → Arg	2	5	16	25	0.713	0.68
Asp → Thr	1	3	12	31	0.636	0.91
						$\chi^2 = 39.50$

[a] These results are mostly the same as those reported in Ronneberg et al.'s Tables 3 and 4. In determining the a and b parameters, the three termination codons have not been included. The probability (P) is calculated using the hypergeometric distribution formula (see Methods). See text for further information.

Ronneberg et al. (2000) also include the latter four pairs because they believe Wong (1975) claimed that this addition of pairs would lower the probability value. Wong did not make such a claim. What he actually said was that, to obtain a lower probability, the Thr–Met pair and the pairs resulting from the biosynthetic relationships Ala–Ser–Gly and Glu–Asp–Ala (Wong 1975, p. 1910) had to be added, and this is indeed the case (Di Giulio 1991).

Finally, Ronneberg et al. (2000) present a test that gives $P = 0.62$, calculated under the condition that the AAY and CAR codons never codified for Asp and Glu. The comparable test performed on the data in Table 1 gives $P = 0.17$ ($\chi^2 = 30.49$, df $= 24$, $P = 0.17$). Clearly the latter test is performed under conditions that have little meaning for the coevolution theory. It is important to recognize that the AAY and CAR codons codified in an intermediate stage of genetic code evolution for Asp and Glu, respectively (Wong 1975), and these are not speculative assumptions, as Ronneberg et al. (2000) claim, because we have the molecular fossils of these ancient assignments. These are represented by the pathways Asp–tRNA^{Asn} → Asn-tRNA^{Asn} and Glu-tRNA^{Gln} → Gln-tRNA^{Gln} (Ibba et al. 1997), which exemplify the precursor–product amino acid transformations taking place on tRNA-like molecules hypothesized by Wong (1975) and involving AAY and CAR codons. Nowadays, contrary to the claims of Ronneberg et al.

(2000), these and other pathways are acknowledged as being molecular fossils by researchers from a wide range of different cultural backgrounds (Wong 1976, 1988; Wachtershauser 1988; Danchin 1989; de Duve 1991; Di Giulio 1997a, 1999; Tumbula et al. 2000) and are the most important proof in favor of the coevolution theory at our disposal (Di Giulio 1997b, 2000). In conclusion, it is more than reasonable to claim that the latter test cannot be used as evidence against the coevolution theory.[2]

Criticism of the Code Set Determined Using the NNY Constraint

The code set used by Wong (1975) to provide a statistical basis to the coevolution theory is the most general that can be defined. This set attributes each amino acid with the ability to occupy any codon in the genetic code without any restrictions whatever and, therefore, regardless of the evolutionary paths that actually took place. Ronneberg et al. (2000) criticize the use of this set because in the current translation apparatus there is no tRNA that can discriminate between codons terminating in U and those terminating in C. They therefore conclude that NNY codons behaved as if they were a single entity throughout genetic code evolution. Consequently, in applying the hypergeometric distribution, NNY codons must have a value of 1 unit, and not 2 as in the set considered by Wong (1975). (The reader is referred to Ronneberg et al.'s Fig. 2 for a representation of the sets discussed here.) They also claim that the translation apparatus seems to read the NNY codons as synonyms by necessity (Ronneberg et al. 2000), thus implying that evolution was unable to discriminate between codons terminating in U and those terminating in C. I believe that there was no such need in code evolution. If Ronneberg et al.'s claims were true, all the currently existing tRNAs should not be able to discriminate U from C in the third codon position. In other words, U and C in the third position should behave as if they were a single letter, i.e., there should be no tRNA able to distinguish between them. This is not the case. In the Bacteria domain there are tRNAs that can recognize the family boxes of four codons by means of the 5-hydroxyuridine in the first anticodon position, thus managing to read only the codons terminating in U, A, and G, and not those terminating in C (Soll and RajBhandary 1995, pp. 209, 213, 226), and therefore discriminating U from C. Hence it is not in the least automatic to consider the NNY constraint as Ronneberg et al. (2000) claim.

Apart from these considerations, if we assume that Ronneberg et al.'s choice regarding the NNY constraint is correct, we must also consider, for example, that for much or even most of the genetic code's evolution the amino acids codified by four codons were decodified by a single anticodon and thus by a single tRNA (Osawa and Jukes 1988). This is equivalent to saying that in the origin of the genetic code, all synonymous codon blocks were decodified by a single anticodon. This seems to be justified also by the implausibility of the alternative hypothesis, which envisages, as an ancestral condition under which the genetic code evolved, that more than one tRNA with different anticodons decodified the same synonymous codon block, as now happens. These considerations thus lead us to believe that the set of codes on which to define the statistical bases of the coevolution theory is the one represented by the amino acid permutation codes, i.e., the one that leaves the synonymous codon blocks unchanged with respect to the genetic code and makes possible amino acid permutation on these blocks (Di Giulio 1989). This is because the latter set is the one that was asserted and must therefore conserve many of the characteristics of the evolving code, including the possibility that the synonymous codon blocks were decodified by a single tRNA for much of the genetic code's evolution. Therefore, they should be considered as a single unit when the hypergeometric distribution is applied.

However, the use of the hypergeometric distribution on the set of codes in which every synonymous codon block in the genetic code is a single unit does not seem to be pertinent, as every single probability thus calculated for every pair of amino acids is not significant. For example, the probability associated with the precursor–product pair Val–Leu is equal to 0.16 ($x = 2$, $n = 2$, $a = 9$, $b = 13$), while in the sets considered by Wong (1975) and by Ronneberg et al. (2000), the probability of this pair is one of the lowest that can be calculated. This problem also arises, although only partly, in the set considered by Ronneberg et al. (2000), and it may be one of the reasons for the lower statistical significance of the coevolution theory. Although the aggregate of the probabilities might still be significant despite these adverse conditions, it is clear that to establish the statistical significance of a certain number of amino acid pairs in this set, it is more convenient to use a method based on the generation of random codes on the synonymous codon blocks (Di Giulio and Medugno 2000).[3]

[2] In Ronneberg et al.'s paper there are some inaccuracies in the sections regarding these statistical tests. (1) In their Table 3 the *a* and *b* parameter values of the Gln–His pair are 13 and 33, respectively, and not 11 and 35 (values in brackets) as reported. (2) The legend to their Table 4 reports $\chi^2 = 28.70 + 8.42$ (28.39 + 8.53), which finds no confirmation in the calculations performed and cannot be justified. (3) The χ^2 value of 28.70 is actually equal to 30.30, and even if the values in their Tables 3 and 4 are used, a χ^2 value of 29.62 is obtained, which is different from the 28.70 reported.

[3] This also makes it possible to remove certain problems deriving from

Table 2. The results of a simulation consisting of the generation of 100 million random codes in the amino acid permutation code set (Di Giulio and Medugno 2000)[a]

Precursor–product pair set	CCS	Probability
1. Ser-Trp = 1, Ser-Cys = 4, Phe-Tyr = 2, Thr-Ile = 3, Gln-His = 4, Glu-Gln = 2, Asp-Asn = 2, Asp-Met = 0	18	3.9×10^{-3}
2. Ser-Trp = 1, Ser-Cys = 4, Phe-Tyr = 2, Thr-Ile = 3, Gln-His = 4, Glu-Gln = 2, Asp-Asn = 2, Asp-Met = 0, Asp-Lys = 0, Glu-Pro = 0, Asp-Arg = 0, Asp-Thr = 0	18	0.10
3. Ser-Trp = 1, Ser-Cys = 4, Phe-Tyr = 2, Thr-Ile = 3, Gln-His = 4, Glu-Gln = 2, Asp-Asn = 2, Asp-Met = 0, Asp-Lys = 0, Glu-Pro = 0, Asp-Arg = 0, Asp-Thr = 0, Val-Leu = 6	24	0.038

[a] The probability represents the frequency at which the particular set of amino acid pairs considered is encountered in 10^8 random codes. For every amino acid pair a number is given to represent the number of times that the two amino acids interchange on the basis of the genetic code structure and considering only the single base changes as equiprobable. The sum of these numbers extended to all the pairs in the set forms the codon correlation score (CCS) (Amirnovin 1997; Di Giulio and Medugno 2000). See text for further information.

In conclusion, each of the sets discussed has characteristics making it suitable to represent (1) an abstract situation, i.e., a very general one—and in this lies the strength of Wong's code set, which does not allow any restriction by the evolutionary paths that the genetic code actually took, thus evaluating the probability in totally generalized terms; (2) a situation, that of the NNY constraint (Ronneberg et al. 2000), which truly, but only partially, includes the evolutionary paths followed by the codons of the genetic code; (3) a situation, that of the amino acid permutation codes (Di Giulio 1989), which seems to include most of the evolutionary paths followed by the codons of the genetic code and which could therefore be considered the most correct from this viewpoint. In the next section, therefore, I complete the analysis by presenting the results relative to the amino acid permutation code set.

The Statistical Significance of the Coevolution Theory in the Amino Acid Permutation Code Set

Using the method proposed by Di Giulio and Medugno (2000), which makes it possible to generate a high number of random codes that, like the genetic code, leave the allocations relative to the synonymous codon blocks unchanged (Di Giulio 1989) but allows the amino acids to occupy any of the 20 positions, I subjected the sets of pairs of amino acids reported in Table 2 to a statistical significance test. However, in estimating the probability I considered as a rarer event (in a set of precursor–product pairs) the number of pairs participating in a given CCS value (Table 2), which turns out to be higher than the number of pairs in a real precursor–product relationship for that set, regardless of the CCS value associated with that number of pairs. This choice leads, in any case, to an overestimate of the probability associated with a given set of amino acid pairs and is, therefore, more conservative than the one calculated using the method reported by Di Giulio and Medugno (2000). The difference between the two probabilities is nevertheless minimal.

For the set of pairs reported in Ronneberg et al.'s Table 3, for which they find $P = 0.168$, I found $P = 3.9 \times 10^{-3}$ (Table 2, set 1). Therefore, changing the code set is associated with a large variation in statistical significance.

Clearly for the set obtainable from the pairs in Ronneberg et al.'s Tables 3 and 4, and assuming Wong's postulates to be true (i.e., the AAY and CAR codons codified for Asp and Glu, respectively, in the evolving code), the probability cannot be easily calculated unless we consider the appearance times of the different amino acids in the genetic code, as envisaged by the coevolution theory (Di Giulio and Medugno 1999). This, however, would require a separate study in its own right. Nevertheless, we can perform some extremely useful checks.

If we consider the amino acid pairs in Ronneberg et al.'s Tables 3 and 4 but assuming Wong's postulates to be untrue, we obtain $P = 0.10$ (Table 2, set 2), versus $P = 0.62$ in Ronneberg et al.'s paper. This probability ($P = 0.10$), which is only marginally significant, is obtained under conditions that have little meaning for the coevolution theory, and this clearly indicates that the predictions of this theory are nevertheless deeply rooted in the organization of the genetic code otherwise, with no

the use of the hypergeometric distribution, such as the dependence between the probability values associated with single amino acid pairs and used in the Fischer test (1950), which requires these values to be independent. However, contrary to Ronneberg et al.'s claims, it is unclear why, randomizing all the possible orderings, the inaccuracy deriving from the probability dependence should make the low probability values turn out to be underestimates, whereas if, as seems to be the case, the biosynthetic pathways of amino acids are linked to the organization of the genetic code, then, for instance, several amino acids biosynthetically linked to the same precursor should restrict the possibility of another product being contiguous to this precursor since the *a* parameter value of the hypergeometric distribution decreases, and should therefore result in an overestimate, and not an underestimate, of the probability. The effect of the dependence of probabilities is somewhat difficult to predict, as it is a function of the individual amino acids appearance times in the genetic code, although the mean effect could lead to an overall probability approximately equivalent to the one in which the individual probabilities are truly independent.

fewer than 5 of 12 amino acid pairs which are not contiguous in the genetic code (Table 2, set 2), we should have obtained a much higher probability, which is not the case. To confirm this we need only insert the Val–Leu pair into this set to recover statistical significance ($P = 0.038$, Table 2, set 3).

Obviously, the set that can be derived from the first eight pairs in Table 1 has an extremely high level of significance ($P = 8 \times 10^{-5}$) (Di Giulio and Medugno 2000).

Overall these observations seem to indicate that the coevolution theory is sufficiently supported if we analyze its statistical significance in the amino acid permutation code set (see also footnote 1), even if we consider the pairs favored by Ronneberg et al. (2000).

Conclusion

Ronneberg et al. (2000) suggest abandoning the coevolution theory on the basis of two criticisms. They claim that (1) various amino acid pairs in a precursor–product relationship used by Wong (1975) are not actually such and (2) by changing the code set in which to investigate the statistical robustness of the coevolution theory, the latter would be discredited. Here I have shown that neither of these criticisms is justified. (1) I have made more than reasonable observations and arguments which lead to the conclusion that all the precursor–product amino acid pairs used by Wong (1975) are such or, at least, can be interpreted as such. If we use these pairs to check the statistical significance of the coevolution theory in the code set proposed by Ronneberg et al. (2000), we find that it is statistically significant. (2) Furthermore, the code set suggested by Ronneberg et al. (2000) is open to the criticism that if we accept the NNY constraint, we must also consider that the other synonymous codon blocks were, for most of the origin of the genetic code, decodified by a single anticodon. Therefore, all the synonymous codon blocks should be considered as a single unit when applying the hypergeometric distribution, with the consequence that the code set subject to the NNY constraint is substituted by the amino acid permutation code set in which even the precursor–product pairs favored by Ronneberg et al. (2000) are significantly in favor of the coevolution theory.

In conclusion, the arguments and statistical analysis reported in the present paper, along with the observations available in the literature (Dillon 1973; Wong 1975, 1976, 1988; McCledon 1986; Miseta 1989; Taylor and Coates 1989; Danchin 1989; Di Giulio 1991, 1996, 1997a, 1999, 2000; de Duve 1991; Morowitz 1992; Tumbula et al. 2000), make Ronneberg et al.'s analysis of minor scientific value. This is, in a certain sense, paradoxical, as, for instance, their definition of the amino acids in a precursor–product relationship seems to be scientifically correct. But if it is applied rigidly and without the due elasticity, it leads us to refute the coevolution theory, which is arguably the best theory at our disposal to explain the organization of the genetic code. Therefore, in more general terms, Ronneberg et al.'s paper is, in my opinion, an example of an extreme use of scientific method. In the field of evolutionary biology, this method can be difficult to use because it sometimes requires a certain elasticity that may not be easy to incorporate into the analysis, with the risk that, if it is not incorporated, it will result simply in blind empiricism.

References

Amirnovin R (1997) An analysis of the metabolic theory of the origin of the genetic code. J Mol Evol 44:473–476

Danchin A (1989) Homeotopic transformation and the origin of translation. Prog Biophys Mol Biol 54:81–86

de Duve C (1991) Blueprint for a cell: The nature and origin of life. Neil Patterson, Carolina Biological Supply Company, Burlington, NC, pp 175–181

Di Giulio M (1989) The extension reached by the minimization of polarity distances during the evolution of the genetic code. J Mol Evol 29:288–293

Di Giulio M (1991) On the relationships between the genetic code coevolution hypothesis and the physicochemical hypothesis. Z Naturforsch 46C:305–312

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

Di Giulio M (1997a) On the origin of the genetic code. J Theor Biol 187:573–581

Di Giulio M (1997b) The origin of the genetic code. Trends Biochem Sci 22:49

Di Giulio M (1999) The coevolution theory of the origin of the genetic code. J Mol Evol 48:253–254

Di Giulio M (2000) The RNA world, the genetic code and the tRNA molecule. Trends Genet 16:17–18

Di Giulio M, Medugno M (1999) Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. J Mol Evol 49:1–10

Di Giulio M, Medugno M (2000) The robust statistical bases of the coevolution theory of the genetic code. J Mol Evol 50:258–263

Dillon LS (1973) The origins of the genetic code. Bot Rev 39:301–345

Fisher RA (1950) Statistical methods for research workers, 11th ed. Oliver and Boyd, Edinburgh and London, p 99

Greenberg DM (1969) Metabolic pathways. Amino acids and tetrapyrroles, Vol III, 3rd ed. Academic Press, New York

Ibba M, Curnow AW, Soll D (1997) Aminoacyl-tRNA synthesis: Divergent routes to a common goal. Trends Biochem Sci 22:39–42

Jensen RA, Fisher R (1987) The postprephenate biochemical pathways to phenylalanine and tyrosine: An overview. Methods Enzymol 142:472–478

Jensen RA, Stenmark SL (1975) The ancient origin of a second microbial pathway for L-tyrosine biosynthesis in prokaryotes. J Mol Evol 4:249–259

Jukes TH (1973) Arginine as an evolutionary intruder into protein synthesis. Biochem Biophys Res Commun 53:709–714

McClendon JH (1986) The relationship between the origins of the biosynthetic paths to the amino acids and their coding. Orig Life 16:260–270

Miseta A (1989) The role of protein associated amino acid precursor molecules in the organization of genetic codons. Physiol Chem Phys Med NMR 21:237–242

Morowitz HJ (1992) Beginnings of cellular life: Metabolism recapitulates biogenesis. Yale University, Vail-Ballou Press, Binghamton, NY, pp 160–171

Osawa S, Jukes TH (1988) Evolution of the genetic code as affected by anticodon content. Trends Genet 4:191–198

Ronneberg TA, Landweber LF, Freeland SL (2000) Testing a biosynthetic theory of the genetic code: Fact or artifact? Proc Natl Acad Sci USA 97:13690–13695

Soll D, RajBhandary UL (1995) tRNA structure, biosynthesis, and function. ASM Press, Washington, DC

Srenmark SL, Pierson DL, Jensen RA, Glover GI (1974) Blue-green bacteria synthesise L-tyrosine by the pretyrosine pathway. Nature 247:290–292

Taylor FJR, Coates D (1989) The code within the codons. BioSystems 22:177–187

Tumbula DL, Becker HD, Chang WZ, Soll D (2000) Domain-specific recruitment of amide amino acids for protein synthesis. Nature 407:106–110

Voet D, Voet JG (1990) Biochemistry. John Wiley & Sons, New York

Wachtershauser G (1988) Before enzymes and templates: theory of surfage metabolism. Microbiol Rev 52:452–484

Wong JT (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912

Wong JT (1976) The evolution of a universal genetic code. Proc Natl Acad Sci USA 73:2336–2340

Wong JT (1988) Evolution of the genetic code. Microbiol Sci 5:174–182