J Mol Evol (2000) 50:258–263
DOI: 10.1007/s002399910030

JOURNAL OF MOLECULAR EVOLUTION

The Robust Statistical Bases of the Coevolution Theory of Genetic Code Origin

Massimo Di Giulio,[1] Mario Medugno[2]

[1] International Institute of Genetics and Biophysics, CNR, Via G. Marconi 10, 80125 Naples, Napoli, Italy
[2] Centro di Ricerche per il Calcolo Parallelo e i Supercalcolatori, CNR, c/o Dipartimento di Matematica ed Applicazioni, Complesso Universitario Monte S. Angelo, Via Cintia, 80126 Naples, Napoli, Italy

Received: 28 June 1999 / Accepted: 23 October 1999

Abstract. A paper (Amirnovin R, J Mol Evol 44:473–476, 1997) seems to undermine the validity of the coevolution theory of genetic code origin by shedding doubt on the connection between the biosynthetic relationships between amino acids and the organization of the genetic code, at a time when the literature on the topic takes this for granted. However, as a few papers cite this paper as evidence against the coevolution theory, and to cast aside all doubt on the subject, we have decided to reanalyze the statistical bases on which this theory is founded. We come to the following conclusions: (1) the methods used in the above referred paper contain certain mistakes, and (2) the statistical foundations on which the coevolution theory is based are extremely robust. We have done this by critically appraising Amirnovin's paper and suggesting an alternative method based on the generation of random codes which, along with the method reported in the literature, allows us to evaluate the significance, in the genetic code, of different sets of amino acid pairs in biosynthetic relationships. In particular, by using this method and after building up a certain set of amino acid pairs reflecting the expectations of the coevolution theory, we show that the presence of this set in the genetic code would be obtained, purely by chance, with a probability of 6×10^{-5}. This observation seems to provide particularly strong support to the coevolution theory.

Correspondence to: M. Di Giulio; *e-mail:* digiulio@iigb.na.cnr.it

Key words: Genetic code theories — Random code distributions — Coevolution — Biosynthetic relationships between amino acids — Hypergeometric distribution

Introduction

Up until now the literature has taken for granted the existence of a correlation between the biosynthetic relationships between amino acids and genetic code organization (Pelc 1965; Dillon 1973; Wong 1975; McClendon 1986; Miseta 1989; Taylor and Coates 1989; de Duve 1991; Di Giulio 1991; Morowitz 1992). Hence, Amirnovin's paper (1997) came as something of a surprise as it casts doubt over the existence of such a correlation. We have already criticized this work (Di Giulio 1999), and Amirnovin and Miller's reply (1999) to our letter is, in our opinion, somewhat inattentive.

However, as a few papers (Freeland and Hurst 1998a, b; Knight et al. 1999) cite Amirnovin's paper as evidence against the coevolution theory of genetic code origin, and to eliminate any doubt from this field, we have decided to reanalyze the statistical bases on which this theory is based.

Methods

On the basis of the experience accumulated in previous works on the subject (Di Giulio 1989a, b; Di Giulio et al. 1994; Di Giulio and

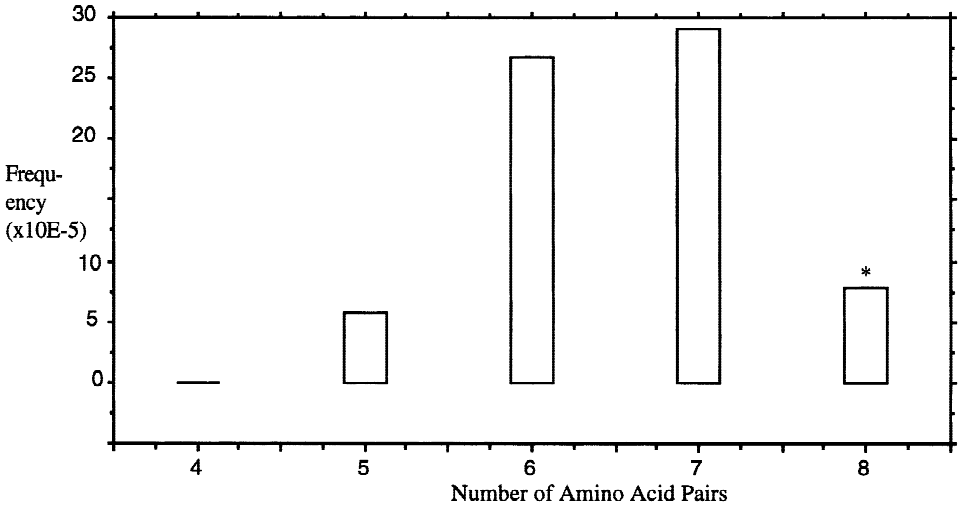

Fig. 1. The frequency distribution between the number of occurrences of random codes having a CCS value equal to or greater than 24 units and the number of amino acid pairs intervening in the determination of these CCS values, for the set of eight pairs in Amirnovin's (1997) Table 1. This distribution derives from a trial in which 10^8 random codes were generated. Only 69,294 codes (here represented) display a CCS value equal to or greater than 24 units, and only 7785 of these display CCS values determined by only eight pairs of amino acids. The *asterisk* denotes the class containing the eight pairs of amino acids encountered in the genetic code (Amirnovin 1997, Table 1).

Medugno 1998, 1999), we have written a program that makes it possible to calculate the probability that can be associated to a given set of pairs of amino acids in biosynthetic relationships.

In particular, the program generates random numbers and uses appropriate representations and manipulations of sets in order to build random codes. These codes are identical to the genetic code as far as the relative allocations of synonymous codons blocks are concerned but they differ from it in that the amino acids are permuted. In other words, a randomly built code is one of the 2.4×10^{18} (20 factorial) possible permutations of amino acids in the genetic code table (Di Giulio 1989a).[1]

For every random code, the program calculates the Codon Correlation Score (CCS) (Amirnovin 1997). That is, for every pair of amino acids *ij*, the program identifies the number of times that amino acid *i* is transformed into amino acid *j* on the basis of the genetic code structure and considering only single base changes (Di Giulio 1989b). (In our program these numbers are given in matrix form.) The sum of all these numbers for all the pairs of amino acids (biosynthetic relationships) considered makes up the CCS (Amirnovin 1997).

We use p to indicate the bijective correspondence mapping the ith position in the genetic code to the amino acid p_i. The inverse correspondence p^{-1} maps the ith amino acid to the position p_i^{-1}. The CCS is determined by examining the number of "correlations" between the codons of amino acids defined in the biosynthetic relationships as $b_r = (\pi(r), \sigma(r))$, $i = 1 \ldots N_{\text{rel}}$ considered in the analysis (Di Giulio and Medugno 1998). If A is the weight matrix of the genetic code (Di Giulio 1989b; Di Giulio and Medugno 1998, 1999), then the CCS turns out to be

$$\text{CCS} = \sum_{i=1}^{N_{\text{rel}}} a(p_{\pi(i)}^{-1}, p_{\sigma(i)}^{-1})$$

After calculating the CCS for a given random code, the program stores the value in a vector and increments by 1 unit the corresponding frequency at which that value was observed up to the previous iteration. Proceeding in this way, the program is able to build frequency distributions equivalent to those built by Amirnovin (1997).

The program also builds another frequency distribution. In this case only the CCS values equal to or greater than a certain threshold (preassigned in the infile) are considered useful to frequency distribution construction. Once it has been established that a certain CCS value can enter this distribution, the program simply increments by 1 unit the frequency value corresponding to the number of amino acid pairs actively intervening in the definition of that specific CCS value. Examples of this frequency distribution are given in Figs. 1, 2, and 3.

We checked that the program performed as required in two ways. The first of these entailed the printout of many random code configurations, which allowed a manual verification of their CCS values and the number of amino acid pairs intervening in their determination. The second used a CCS threshold of 0 and a number of amino acid pairs (biosynthetic relationships) equal to 1 so that we could check the program's adequacy by comparing the CCS frequency distribution to the theoretically expected values (in the discrete interval of 0–6 units of the CCS values, the value 5 is not expected) obtained on the basis of genetic code structure.

Finally, a number of trials allowed us to establish that 100 million random codes is a sufficient number of random codes to generate because it seems that the variability in this case is under control.

This program runs on PC and is available from the authors upon request.

Results and Discussion

Amirnovin's paper (1997) contains two types of mistake. The first type can be said to be absolute. In order to understand this type of error, we consider a hypothetical pair of amino acids in a precursor–product relationship.

[1] We note that in his Method 3, Amirnovin (1997) also permutes the meaning of termination codons (= Ter), as he treats Ter as if it were an amino acid. This is a mistake. If Ter is not permuted, it does not affect the probability calculation, whereas if it is permuted, it does, because changing the position of Ter is tantamount to changing the whole organization of the genetic code. This must not be allowed to happen, as we are trying to establish the probability that a given number of pairs of amino acids in biosynthetic relationships are allocated on a purely random basis in a certain way within the genetic code table, which must therefore be invariant in the relative arrangement of both the blocks of synonymous codons of amino acids and the three termination codons.

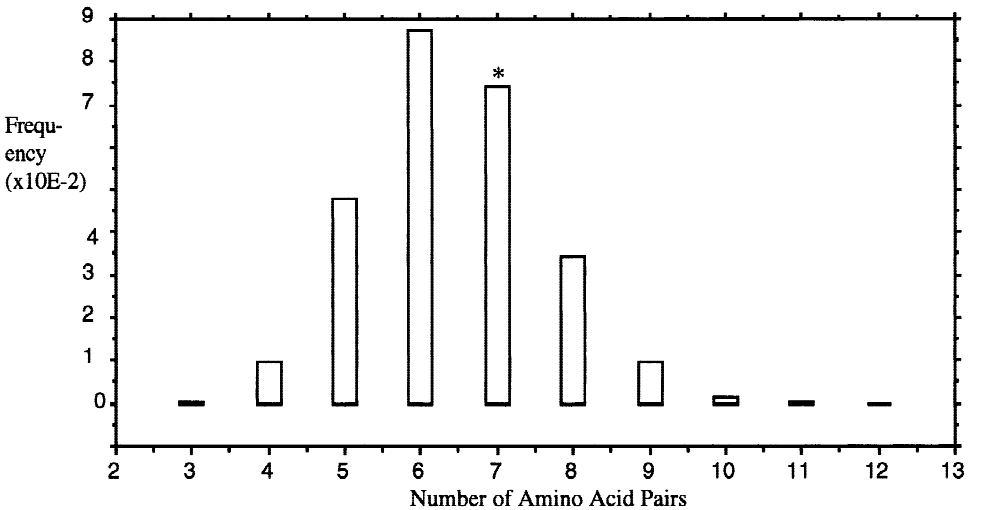

Fig. 2. The frequency distribution between the number of occurrences of random codes having a CCS value equal to or greater than 16 units and the number of amino acid pairs intervening in the determination of these CCS values, for the set of pairs in Amirnovin's (1997) Table 2. This distribution derives from a trial in which 10^8 random codes were generated; 26,535,635 codes (here represented) display a CCS value equal to or greater than 16 units, and only 12,009,767 of these display CCS values determined by a number of amino acid pairs that is equal to or greater than seven. The *asterisk* denotes the class containing the seven amino acid pairs that have a CCS value different from 0 and that are therefore encountered in the genetic code and are reported in Amirnovin's (1997) Table 2. In the trial we used 12 amino acid pairs (Amirnovin 1997, Table 2), i.e., the program looked for a specific CCS value using all 12 pairs assigned in the infile.

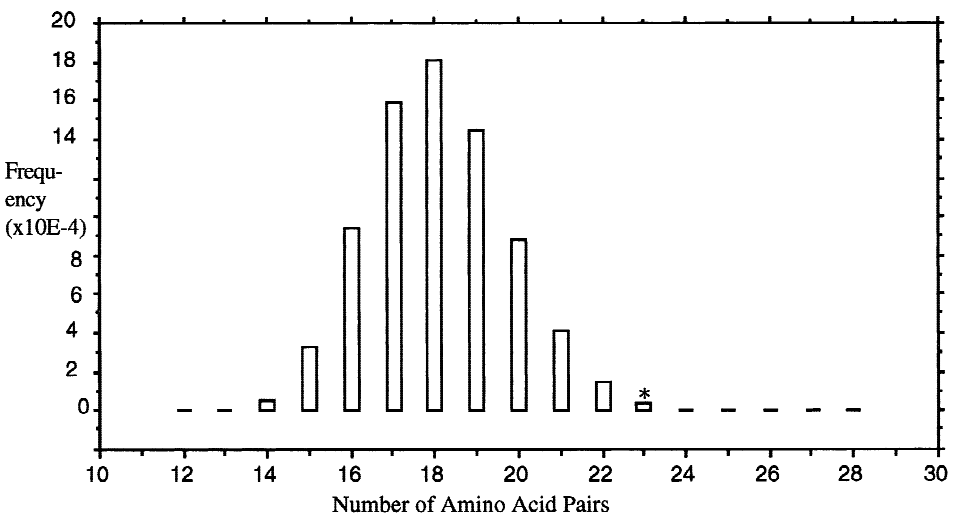

Fig. 3. The frequency distribution between the number of occurrences of random codes having a CCS value equal to or greater than 53 units and the number of amino acid pairs intervening in the determination of these CCS values, for the set of pairs in our Table 1. This distribution derives from a trial in which 10^8 random codes were generated. Only 767,232 codes (here represented) display a CCS value equal to or greater than 53 units, and only 5648 of these display CCS values determined by a number of pairs of amino acids that is equal to or greater than 23. The *asterisk* denotes the class containing the 23 pairs of amino acids encountered in the genetic code and that are specified in Table 1 (pairs with values different from 0). In the trial we used 31 amino acid pairs (Table 1), i.e., the program looked for a specific CCS value using all 31 pairs assigned in the infile.

We assume that this pair occupies the positions occupied in the genetic code by Ser and Arg. The value that Amirnovin's function (1997), i.e., the CCS, attributes to this pair is 6 units, which is given simply by the number of times that the codons of Arg (Ser) transform into those of Ser (Arg) on the basis of the genetic code structure and considering only single base changes. The value of 6 units is the maximum value that Amirnovin's function (1997) can attribute to a pair of amino acids on the basis of genetic code structure. It is to be expected then that this value should contribute to shifting the CCS value toward the right-hand tail of the distributions generated by Amirnovin (1997), and this contributes, in the code having such a pair, to making the code highly significant in probability terms. Moreover, Wong's (1975) method, which associates a probability (P) value to every pair of amino acids on the basis of hypergeometric distribution (Wong 1975; Di Giulio 1991), would associate a mean value of $P = 0.70$ to this pair (Ser–Arg, $a = 34$, $b = 24$, $n = 6$, $x = 4$, $P = 0.51$; Arg–Ser, $a = 28$, $b = 30$, n

$= 6$, $x = 2$, $P = 0.89$), which is not significant. Why, then, for the same pair of amino acids (i.e., for the same positions occupied by two amino acids in the code), do the two methods affect the probability values in such different ways? While Amirnovin's probability would certainly be lowered by a pair of amino acids occupying such a position in the code, the hypergeometric distribution (Wong 1975; Di Giulio 1991), on the other hand, would make such a pair nonsignificant.

We can clarify this with an example. Let us consider a hypothetical set of eight pairs of amino acids in biosynthetic relationships, such as those reported in Amirnovin's Table 1 (1997). If these precursor–product pairs occupied, in the randomly generated code, the positions occupied in the genetic code by Ser–Cys (for the Ser–Cys precursor–product pair), by Ser–Trp (for the Ser–Trp pair), by Pro–Leu (for the Glu–Gln pair), by Leu–Phe (for the Gln–His pair), by Ile–Met (for the Val–Leu pair), by Asp–Gly (for the Asp–Asn pair), by Glu–Ala (for the Thr–Ile pair), and by Thr–Arg (for the Phe–Tyr pair), then for these eight pairs we would obtain a CCS value of 24 units which, from Amirnovin's Fig. 1 (1997), turn out to have a highly significant probability. In contrast, these same eight pairs would be nonsignificant according to Wong's (1975) method based on the hypergeometric distribution ($\chi^2 = 18.69$, $n = 8$, df $= 16$, $0.20<P<0.30$).

Clearly, by characterizing the pairs of amino acids by means of a single variable, Amirnovin's (1997) method is unable to pay due consideration to the whole genetic code structure, which is indeed fundamental if we are to establish whether or not two of its positions occupied by amino acids in a precursor–product relationship significantly "overlap," while the hypergeometric distribution can achieve this (Wong 1975). The latter distribution is a function of (1) the number of codons in the genetic code that are contiguous to those of a certain precursor amino acid, (2) the number of codons that are not contiguous to this precursor, (3) the total number of codons codifying for a given product amino acid, and (4) the number of these product codons that are contiguous to those of the precursor. These properties of the hypergeometric function are the very ones that are expected from a function that has to establish whether or not a certain pair of amino acids in precursor–product relationship possesses a significant probability on the basis of the genetic code structure. Amirnovin's CCS does not do this. Consequently, Amirnovin's (1997) method does not give a correct evaluation of the probability to associate to a certain number of amino acid pairs because it only takes into account a single variable characterizing the amino acid pairs in the genetic code, whereas it should also consider the number of amino acid pairs that actually take part in determining a specific CCS value. This introduces the second type of mistake present in Amirnovin's paper (1997).

To calculate the probability to associate to a certain number of amino acids, Amirnovin (1997) generates random codes that allow him to build a frequency distribution between the number of occurrences of random codes and the respective CCS values. Then, according to the position that the CCS value of the considered set of pairs occupies in this distribution, he establishes the probability (Amirnovin 1997). However, the probability that he calculates is only the probability of obtaining a certain CCS value, whereas we are not really interested in this value but, rather, in a conditioned probability. That is, we have to calculate the probability of observing a certain CCS value on the condition that the CCS value is produced only by those that have a number of amino acid pairs at least equal to that of the pairs that are effectively specified in the genetic code and whose significance has to be established, whereas we have to exclude the random codes that have the same rare CCS value but which are actively produced only by a number of amino acid pairs lower than the number of pairs effectively specified in the genetic code and whose significance has to be established. Amirnovin's method does not do this. By taking into account these two characteristics of random codes, one being the CCS value and the other given by the number of amino acid pairs that must actually contribute to defining the CCS value, we manage to introduce into this probability calculation the idea that the rarity of a certain random code is based both on a high CCS value and, above all, on a high number of amino acid pairs contributing to that CCS value. It is clear that, for the same CCS values, a random code to which a higher number of amino acid pairs contributes is less likely to be obtained, because of the sole effect of chance, than a random code possessing the same CCS but to which a lower number of amino acid pairs contributes. This is true in the region to the far right of the distribution, i.e., in the region in which we are most interested.

In short, the probability (or, rather, the estimate of the probability, i.e., a frequency) that has to be calculated is simply given by the ratio between (1) the number of all random codes having a value equal to or greater than a certain CCS value, on the condition that these CCS values are determined by a number of amino acid pairs equal to or greater than the number of amino acid pairs effectively specified in the genetic code of the set being analyzed (this set may also contain a higher number of amino acid pairs than the ones effectively specified in the genetic code), and (2) all the random codes generated in that trial. [However, the probability calculated by Amirnovin (1997) considers only the CCS value and treats in the same way, i.e., considers as belonging to the same logical category, random codes that might differ by several units in the number of amino acid pairs that actively intervene in determining that CCS value.]

In this way, Amirnovin's probability calculation is

corrected and improved and seems to run parallel to Wong's (1975) method.

Therefore, by using this new method we have recalculated the probability for the pairs in Amirnovin's Tables 1 and 2, to obtain respective values of $P = 8 \times 10^{-5}$ (Fig. 1 and its legend) and $P = 0.12$ (Fig. 2 and its legend).

The $P = 8 \times 10^{-5}$ refers to the amino acid pairs used by Wong (1975) to support the coevolution theory, while the $P = 0.12$ refers to the amino acid pairs used by Amirnovin (1997, Table 2) to claim that the coevolution theory cannot be sustained by comparing the biosynthetic pathways of amino acids to the organization of the genetic code (Amirnovin 1997; Amirnovin and Miller 1999). We have already discussed the inadequacies of this set of pairs in representing the biosynthetic relationships between amino acids (Di Giulio 1999). Here we wish to add only that the value of $P = 0.12$, for these 12 pairs, is different from Amirnovin's (1997), which is equal to 0.34, and therefore, our value has a significance level of around 10%. More significantly, we have noticed that if, in these amino acid pairs, we substitute the pair Thr–Met for Asp–Met [this can be justified, as Thr is the closest amino acid to Met in terms of biosynthetic steps (Wong 1975)], we obtain a significant value of $P = 0.044$ (data not shown). Therefore, these 12 amino acid pairs, which, we repeat, do not adequately represent the expectations of the coevolution theory (Di Giulio 1999), also display a certain significance.

However, is there a set of amino acid pairs that, if suitably chosen, can objectively test the coevolution theory or, more generally, the relationship between the biosynthetic pathways of amino acids and the organization of the genetic code? We believe that there is.

The coevolution theory of the origin of genetic code structuring suggests that there was an evolutionary stage during genetic code origin in which only precursor amino acids were codified (Wong 1975). At this evolutionary stage, the theory claims that the codon domains of precursor amino acids had already been defined. Just as the product amino acids evolved from these, part or all of the codon domain of precursors was ceded to the products (Wong 1975). Consequently, the theory predicts that most of the codons of the product amino acids of a certain precursor should be contiguous in the genetic code, i.e., they should differ only in a single base. Indeed, if a precursor amino acid, e.g., Asp [which characterizes an entire biosynthetic family of amino acids that are derived from it (Wong 1975; Miseta 1989; Taylor and Coates 1989)], had been attributed with a non-contiguous codon domain, then the theory, even though true by hypothesis, would be falsified, as the contiguities between the codons of the product amino acids relative to a specific precursor and those between the precursors and the products would, by definition, not be observed.

Therefore we used the biosynthetic relationships reported in Taylor and Coates' Fig. 1 (1989) and we defined the families of amino acids in biosynthetic relationships as the ones obtained by considering an entire group of amino acids that biosynthetically derive from an amino acid precursor or non-amino acid precursor. In this way we established a set formed of all the possible combinations of amino acid pairs obtainable from all the families. This set (reported in Table 1) does not contain the amino acid His, which, according to Taylor and Coates (1989), is metabolically isolated and, in our opinion, represents a set on which the coevolution theory can be objectively tested because it is derived from a rigorous definition of the amino acid families in a biosynthetic relationship. Moreover, as this set is based on all the possible combinations of amino acid pairs of a biosynthetic family, the analysis need not use only the amino acids in a precursor–product relationship, which is a fundamental concept to the coevolution theory but which here interferes with the analysis.

Table 1. All the combinations of amino acid pairs relative to the five biosynthetic families of amino acids defined according to a single amino acid precursor or non-amino acid precursor (Taylor and Coates 1989, Fig. 1)[a]

Serine family
Ser–Gly = 2, Ser–Cys = 4, Ser–Trp = 1, Gly–Cys = 2, Gly–Trp = 1, Cys–Trp = 2
Phosphoenolpyruvate family
Phe–Tyr = 2
Pyruvate family
Ala–Val = 4, Ala–Leu = 0, Val–Leu = 6
Aspartate family
Asp–Asn = 2, Asp–Thr = 0, Asp–Ile = 0, Asp–Met = 0, Asp–Lys = 0, Asn–Thr = 2, Asn–Ile = 2, Asn–Met = 0, Asn–Lys = 4, Thr–Ile = 3, Thr–Met = 1, Thr–Lys = 2, Ile–Met = 3, Ile–Lys = 1, Met–Lys = 1
Glutamate family
Glu–Gln = 2, Glu–Arg = 0, Glu–Pro = 0, Gln–Arg = 2, Gln–Pro = 2, Arg–Pro = 4

[a] The numbers indicate the number of times that the amino acids in the pair are interchanged on the basis of the genetic code structure, and the sum of all these numbers is the CCS for this set. See text for further information.

The probability obtained for this set of amino acids (Table 1) is $P = 6 \times 10^{-5}$ (Fig. 3 and its legend), which is thus highly significant. Wong's (1975) method (Di Giulio 1991) applied to this set of pairs (Table 1) also gives a highly significant value of $P = 1.2 \times 10^{-5}$ ($\chi^2 = 98.03$, $n = 23$, df = 46). [The amino acids in each of the 23 amino acid pairs taking part in the determination of the χ^2 value were alternately considered here both as precursor amino acids and as product amino acids. The resulting two values of the $-2\ln P$ quantity (Wong 1975; Di Giulio 1991) were used to obtain a mean that contributed to the χ^2 aggregate value.]

While the probability in the method based on the generation of random codes is directly affected by the 8 pairs of amino acids with a CCS value of 0 (Table 1), as it is

estimated under the condition that all 31 amino acid pairs (Table 1) may be selected to contribute to the properties needed in the random codes (Fig. 3 and its legend), this does not happen in the probability calculated using Wong's method, in which these 8 pairs are simply not considered. Therefore, although in this case the two probabilities are of the same order of magnitude, they are not in actual fact comparable. Even if we can imagine some ways of introducing the influence of the eight pairs of amino acids with a CCS value of 0 (Table 1) into the probability calculation, we thought it appropriate only to refer the following result because it imposes an upper limit on the value of this probability. If we consider that each of the eight pairs contributes to the χ^2 value with 0 units, i.e., we assign $P = 1$, and hence a value of $-2\ln P = 0$ (Wong 1975; Di Giulio 1991), to these pairs, we still obtain a highly significant probability ($\chi^2 = 98.03$ $n = 31$, df $= 62$, $0.001 < P < 0.01$).

In conclusion, it seems that in this case, the probability estimated by means of random code generation is more reliable than the one obtained using Wong's (1975) method, as it also takes into account the eight amino acid pairs that have a CCS value of 0, which the latter method does not. However, we must repeat that some of the random codes present in the right-hand tail of these distributions, e.g., in Fig. 3, might have a non-significant probability according to Wong's (1975) method, but as has been shown above, this is not the case for the genetic code. Although the two methods use a logic that, while different, shares certain features, they nevertheless manage to show the intimate relationship between the biosynthetic pathways of amino acids and the organization of the genetic code.

Conclusions

The data available in the literature (Pelc 1965; Dillon 1973; Wong 1975; McClendon 1986; Miseta 1989; Taylor and Coates 1989; de Duve 1991; Di Giulio 1991; Morowitz 1992) and the observations referred above substantiate the intimate connection between the biosynthetic relationships between amino acids and the organization of the genetic code. Many of these data can be used to corroborate the coevolution theory of the origin of genetic code organization (Wong 1975). These same data could also support a more comprehensive theory of the coevolution theory, but no one has so far been able to put forward a better theory than the one formulated by Wong 25 years ago.

References

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

Amirnovin R, Miller SL (1999) Response. J Mol Evol 48:254–255

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 (1989a) 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 (1989b) Some aspects of the organization and evolution of the genetic code. J Mol Evol 29:191–201

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

Di Giulio M, Medugno M (1998) The historical factor: The biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code. J Mol Evol 46:615–621

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, Capobianco MR, Medugno M (1994) On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. J Theor Biol 186:43–51

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

Freeland SJ, Hurst LD (1998a) The genetic code is one in a million. J Mol Evol 47:238–248

Freeland SJ, Hurst LD (1998b) Load minimization of the genetic code: History does not explain the pattern. Proc R Soc Lond B 265:2111–2119

Knight RD, Freeland SJ, Landweber LF (1999) Selection, history and chemistry: The three faces of the genetic code. Trends Biochem Sci 24:241–247

McClendon JH (1986) The relationship between the origins of the biosynthetic paths to the amino acids and their coding. Origins 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

Pelc SR (1965) Correlation between coding-triplets and amino-acids. Nature 207:597–599

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

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