VESTIGES OF EARLY MOLECULAR PROCESSES LEADING TO THE GENETIC CODE

RICARDO FERREIRA and ANDRÉ RICARDO DE O. CAVALCANTI
Departamento de Química Fundamental – UFPE, Recife, PE 50670–900, Brazil

(Received 6 February, 1996)

Abstract. We compare predictions from a proposed model for the origin of the genetic code (*J. Theor. Biol* (1993) **164**, 291–305) with existing information on the base content of codons and abundance of amino acid in different organisms. A comparison is also made between the three groups of amino acids suggested by the model and the two classes of aminoacetyl-tRNA synthetases. The observed agreements tend to suppot the model.

1. Introduction

Recently (Ferreira and Coutinho, 1993; Ferreira, 1995), we have proposed a model for the origin of the genetic code which signals a small but potentially important departure from the asymptotic situation of Crick's frozen accident (Crick, 1968). The model describes a mechanism for the early appearance in the prebiotic medium of a positive correlation between amino acid and codon-anticodon concentrations. In the case of the anticodons, such correlations were previously described by Lacey and Weber (1976), Weber and Lacey (1978), Lacey and Mullins (1983), Jungck (1978), and Lacey *et al*. (1992), who have shown that, with the exceptions of Ile, Trp and the GC anticodon of Arg, there is a strong correlation between properties such as polarity and hidrophobicity of a given amino acid and those of the corresponding anticodonic diribonucleotide, which suggests an attractive interaction.

It is possible to find, using solid models, that single amino acids are too small to interact specifically with triplets of ribonucleotides. Randomly synthesized peptides, on the other hand, are stereochemically appropriate and, if of small size, a non-negligible number of them will contain an excess of a given residue. For example, 1/200 of all possible pentapeptides contain 3, 4 or 5 residues of a given kind (Cavalcanti and Ferreira, 1995). We have proposed that the correlations observed for amino acids and anticodonic sequences will hold for peptides containing an excess of one given residue.

We assume (Ferreira and Coutinho, 1993) that crucial interactions occurred between oligoribonucleotides produced during the template-assisted self-replicating stage (Ferreira, 1987) and randomly synthesized peptides. If one searches for synergetic effects, that is, processes favoring the growth of both specific ribonucleotide sequences and the corresponding amino acid residues, one finds that the interactions must be those which, (i) increase the growth-rate of oligoribonucleotides with 3'-ends made of the first two bases of a given codon, caused by the catalytic

Origins of Life and Evolution of the Biosphere **27:** 397–403, 1997.

action of peptides in which the corresponding amino acid predominates, and, (ii) simultaneously stabilize these peptides by their differential adsorption to "companion oligoribonucleotides" with 3'-ends made of the complementary anticodons*. The catalytic action of peptide chains in the growth of oligoribonucleotides was experimentally reported by Jungck and Fox (1973), Brack and Barbier (1990) and Barbier *et al*. (1993).

To summarize, the establishment of an amino acid-codon-anticodon correlation depended on three factors:

1. Kinetic factors which produced a variable concentration of the codonic and anticodonic sequences in the oligoribonucleotides during the earliest self-replicating stage; C, G-rich sequences became more abundant (see Section 2 below).
2. The existence of randomly synthesized peptide chains with an excess of a given residue which interacted with specific oligoribonucleotides, leading to the described synergetic effects.
3. The simultaneous presence in solution of both codonic and anticodonic sequences forming groups of oligoribonucleotides growing at the same rate (companion oligoribonucleotides). For reasons given by Ferreira and Coutinho (1993) this factor splitted the amino acid residues in three groups; the most favored (a), the least favored (c), and the intermediate ones (b).

The model is testable in a popperian sense (Popper, 1962), for example, through molecular docking techniques of relevant peptides and oligoribonucleotides. The number of possible interactions is staggering and will require appropriate programs, similar to those being developed in Professor Onuchic's group at UCSD, La Jolla (Aquino *et al.*, 1995). In the present paper we simply confront some predictions of the proposed model with existing information.

2. The Early Predominance of C,G-rich Codons

A characteristic feature of our model is that both in the original renormalization group formalism (Ferreira and Tsallis, 1985) and in its chemical kinetics description (Ferreira, 1987), it predicts that four different monomers forming two complementary pairs were required from the very beginning of the template-assisted self-replicating stage. The requirement was equally operative in the earliest steps of the translation process, and it leads, without further *ad hoc* hypothesis to a total of $4^3 = 64$ codons (Ferreira, 1995).

In the initial self-replication stage the rate of growth of the oligoribonucleotides is proportional to $[K_{CG}^{p} K_{AU}^{q}]$, in which p and q are, respectively, the numbers of CG and AU contacts in the transition states of the relevant condensation reactions,

* For example, 5'p-GAUGG, 5'p-GAUCC, etc. are "companion pentaribonucleotides" in the sense that they have the same sequence of strong (C, G) and weak (A, U) interactors, and, according to our model, grow at the same rate.

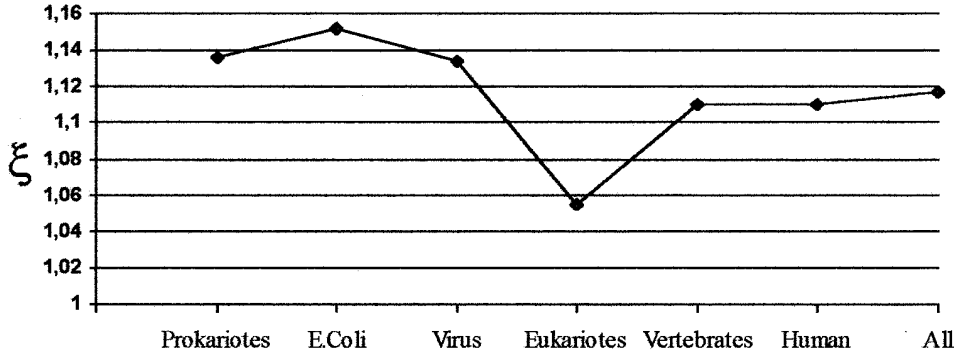

Figure 1. Values of ξ for various organisms.

and K_{CG} and K_{AU} are the corresponding interaction constants. Since $K_{CG} > K_{AU}$, sequences containing an excess of C, G bases over A, U bases grow faster.

This predominance should be reflected in the ratio $[C + G]/[A + U]$ of the earliest genomes, and we ask now if it is possible to discern such predominance in modern organisms.

To do so we define a variable, ξ, which represents the relative abundance of C, G over A, U in various species; this variable was calculated using Equation (1):

$$\xi = \frac{\sum_{i=1}^{20} F_i(C+G)_i}{\sum_{i=1}^{20} F_i(A+U)_i} \tag{1}$$

where F_i is the relative abundance of the amino acid i in the various organisms (Doolittle, 1989), and $C + G$ and $A + U$ are respectively the numbers of $C + G$ and $A + U$ in all codons of amino acid i. The values obtained are plotted in Figure 1.

According to our model, the index ξ should be a rough measure of the ages of the organism's lineage since in the self-replicating phase of oligoribonucleotides the kinetic advantage of C,G-rich over A,U-rich sequences would be dominant. Subsequently this advantage became irrelevant due to the emergence of enzymatic systems followed by the coming of aleatory mutations, which made ξ tend to 1.

One observes that ξ is 1.14 in prokaryotes, and only 1.05 in eukaryotes; although there is an increase in ξ for vertebrates and humans, it does not reach the prokaryote level. These results are supportive of a scenario in which an initial predominance of C,G-rich triplets changed through mutations and selective pressures to a more balanced composition.

3. The Relative Ages of the Two Classes of Aminoacetyl-tRNA Synthetases

The molecules that should better contain vestiges of the mechanism that led to the structuring of the genetic code are the tRNAs, since they establish the link between

the nucleic acid sequences and that of amino acids in proteins. In the light of the hypothesis of an RNA world (Westheimer, 1986, Joyce, 1989) tRNA (or tRNA like) molecules might have mediated the initial translation process. Phylogenetic analysis of the relationships between tRNAs have shown however that these moleculaes suffered a considerable divergence during evolution (Holmquist *et al*., 1973, Eigen *et al*., 1989).

The second group of such molecules are the aminoacetyl-tRNA synthetases. These enzymes are the sites where amino acids are fitted with RNA adapters, each designated enzyme recognizing and activating a given amino acid and binding it to the corresponding adapter. It was previously supposed that the synthetases were descended from a common ancestor (Wetzel, 1978), but more recent works of Eriani *et al*. (1990), Cusak *et al*. (1990), and Nagel and Doolittle (1991) have shown that there are two unrelated classes of these synthetases, each of them containing a set of 10 amino acids.

Figure 2 shows ξ for the two classes of aminoacetyl-tRMA synthetases of different species. These values were calculated making i in the summation of Equation (1), vary from 1 to 10 counting in this way all amino acids in each Class. The amino acids represented by Class II have consistently higher ξ than those amino acids of Class I tRNA synthetases. This is an indiciation that the tRNA Class II synthetases are older than those in Class I synthetases.

It is very improbable that the amino acids which are activated and bonded to the two Classes of tRNA synthetases appeared independently in two distinct time intervals, in the same way that most molecular evolutionists reject the idea of two classes of amino acids appearing in distinct environments (Nagel and Doolittle, 1995). Rather, it seems more likely that the modern aminoacetyl-tRNA synthetases must have been anticipated by simpler peptides capable of some discrimination with respect to anticodonic sequences. Compatible with this view is the proposal that the establishment of the coding process of the amino acids in Class II synthetases overlapped partially with the coding processes of Class I amino acids, and that the overlap increased gradually with time. This hypothesis is supported by a second feature of our model, which we describe presently.

As pointed out in Section 1 a consequence of the model is that the twenty amino acids can be divided-up in three groups, viz.:

a) Gly, Pro, Lys, Asn, Phe, (Leu)*
b) (Leu)*, Val, Ser, Thr, (Arg)*, Asp, Glu, Gln, Cys, Trp, His
c) Tyr, Ile, Met, Ala, (Arg)*

Group (a) is represented by amino acids with the first two positions of their codons occupied by the same kind of base, that is, AA, UU, GG and GC. In our scheme (Ferreira and Coutinho, 1993) these should be considered more primitive amino acids, since the mechanism that established the amino acid-codon concen-

* Leu and Arg, each with six codons, belongs to two groups each. In these case we ended up putting Leu in Group (b) and Arg in Group (c), both with four codons out of six in their respective group.

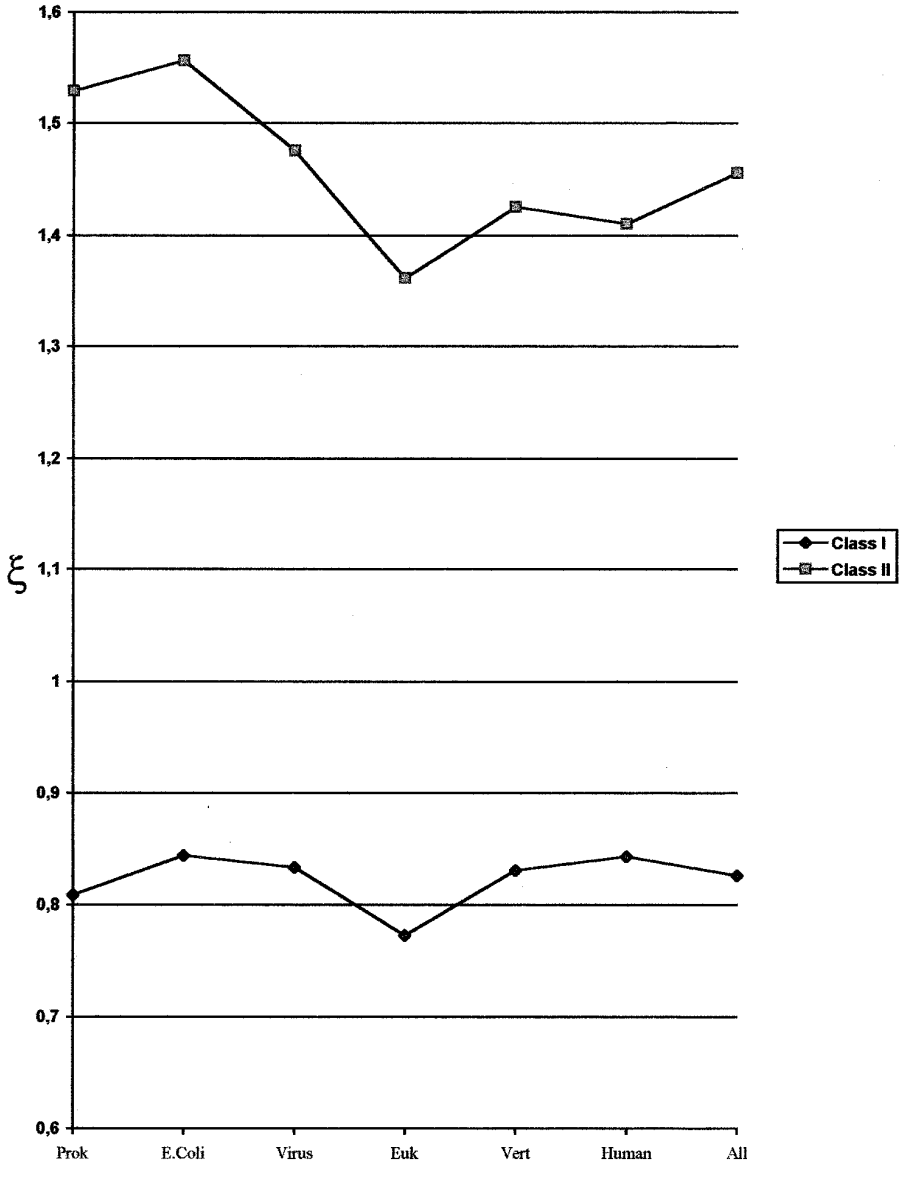

Figure 2. Values of ξ for amino acids belonging to Classes I and II of Aminoacetyl-tRNA synthetases.

tration correlation could have started with very small oligoribonucleotides, as soon as peptides with excess of the corresponding amino acids were available. Amino acids in Group (b) are those with codons starting with AG, AC, GA, CA, GU, UG, UC and CU; in these cases only oligoribonucleotides of moderate size, requiring

Table I

Aminoacids of the two classes of aminoacyl-tRNA synthetases

Class I	Class II
Tyr[c]	Gly[a]
Met[c]	Phe[a]
Ile[c]	Lys[a]
Val[b]	Asn[a]
Leu[b]	Pro[a]
Cys[b]	His[b]
Arg[c]	Asp[b]
Trp[b]	Ser[b]
Glu[b]	Thr[b]
Gln[b]	Ala[c]

Superscripts a, b and c refer to the three groups of amino acids in Ferreira and Coutinho, 1993.

a longer time to grow, could establish the synergetic correlation with the appropriate peptides. Group (c) is made of amino acids for which the synergetic effects involving both codons and anticodons could not be operative, and the concentration correlation was established by a less efficient mechanism.

The amino acids of the two classes of synthetases are depicted in Table I, which also shows the three groups of amino acids. The resulting distribution with all five Group (a) amino acids in Class II, and four of the five Group (c) residues in Class I synthetases, has a probability to have resulted from pure chance that is smaller than 0.004. The case of alanine, the only Group (c) amino acid in Class II tRNA synthetases, is anomalous to certain extent, and can be understood only if alanine was a very abundant amino acid in the prebiotic environment.

Phylogenetic analysis of the tRNA synthetases themselves (Nagel and Doolittle, 1995, Hartman, 1995) suggests that Class II synthetases are indeed more primitive than Class I enzymes. It seems that our mechanism is in accord with these proposals.

4. Conclusions

The results presented here are compatible with the following scenario. In the first self-replicating stage, C,G-rich ribonucleotide sequences predominated over A,U-rich ones. A codon-anticodon-amino acid correlation was then established through a synergetic mechanism of the kind suggested by Ferreira and Coutinho (1993). Eventually a transcription-translation apparatus became the keystone of genetical change. Beginning with the second stage, the advantage of being C,G-rich fell gradually and the index ξ tended asymptotically to a value of 1.

Acknowledgements

This work was supported in part by CNPq (Brazilian Government Agency). RF wishes to thank Prof. Onuchic for his hospitality at UCSD.

References

Aquino, A. J. A., Beroza, P., Beratan, D. and Onuchic, J. N.: 1995, *Chem. Phys.* **197**, 303–31.
Barbier, B., Visscher, J. and Schwartz, A. W.: 1993, 10th Intl. Congr. Orig. Life (Intl. Soc. Stud. Orig. Life), Barcelona, p. 62.
Brack, A. and Barbier, B.: 1990, *Origins Life Evol. Biosphere* **20**, 139–144.
Cavalcanti, A. R. O. and Ferreira, R.: 1995, *An. Acad. Ci. Brasil* **67**, 401–402.
Crick, F. H. C.: 1968, *J. Molec. Biol.* **38**, 367–379.
Cusak, S., Berthet-Colominas, C., Hartlein, M., Nassar, N. and Leberman, R.: 1990, *Nature* **347**, 249–253.
Doolittle, R. F.: 1989, 'Prediction of Protein Structure', G. D. Fasman (ed.), Plenum Press New York, pp. 599–623.
Eigen, M., Winkler-Ostwatisch, R., Dress, A. and Haeseler, A.: 1989, *Science* **244**, 673–679.
Eriani, G., Delarue, M., Poch, O., Gangloff, J. and Moras, D.: 1990, *Nature* **347**, 203–206.
Ferreira, R. and Tsallis, C.: 1985, *J. Theor. Biol.* **117**, 303–313.
Ferreira, R.: 1987, *J. Theor. Biol.* **128**, 289–295.
Ferreira, R. and Coutinho, K. R.: 1993, *J. Theor. Biol.* **164**, 291–305.
Ferreira, R.: 1995, *Zeit. f. Naturf.* **50c**, 148–152.
Hartman, H.: 1995, *J. Mol. Evol.* **40**, 541–544.
Holmquist, R., Jukes, T. H. and Pangburn, S.: 1973, *J. Mol. Biol.* **78**, 91–116.
Joyce, G. F.: 1989, *Nature* **338**, 217–224.
Jungck, J. R. and Fox, S. W.: 1973, *Naturwiss.* **60**, 425–427.
Jungck, J. R.: 1978, *J. Mol. Evol.* **11**, 211–224.
Lacey, J. C. and Weber, A. L.: 1976, 'Protein Structure and Evolution', Fox, L. *et al*. (eds.), New York: M. Dekker, pp. 213–222.
Lacey, J. C. and Mullins, D. W.: 1983, *Origins Life Evol. Biosphere* **13**, 3–42.
Lacey, J. C., Wickramasinghe, N. S. M. D. and Sabatini, R. S.: 1992, *Experienta* **48(4)**, 379–383.
Nagel, G. M. and Doolittle, R. F.: 1991, *Proc. Natl. Acad. Sci.* (U.S.A.) **88**, 8121–8125.
Nagel, G. M. and Doolittle, R. F.: 1995, *J. Mol. Evol.* **40**, 487–498.
Popper, K.: 1962, *Conjectures and Reflections*, Basic Books, New York.
Weber, A. L. and Lacey, J. C.: 1978, *J. Mol. Evol.* **11**, 199–210.
Westheimer, F. H.: 1986, *Nature* **319**, 534–535.
Wetzel, R.: 1978, *Origins Life Evol. Biosphere* **9**, 39–50.