Why 4^3 Codons?

Ricardo Ferreira

Departamento de Química Fundamental, UFPE, Cidade Universitária, 50670-901, Recife, Pernambuco, Brasil

Z. Naturforsch. **50 c,** 148–152 (1995); received August 8, 1994

Genetic Code, Codons, Oligoribotides, Amino Acids, Monochirality

An attempt is made to answer the question posed by the title of this paper. First we show that in primitive self-replicating oligoribotide systems, selection depended from the very start on the existence of 4 kinds of ribotides, forming 2 complementary pairs. Further selection required that the condensation reactions involving the two last positions of the 3′-end of a growing oligoribotide fragment and the first position of the 5′-end of another fragment were catalyzed by randomly synthesized peptides. This established a codon – amino acid concentration correlation and clinched triplet segments as the basis of the translation process. Finally, physical arguments are given to show that the monochirality of the ribotides arose from stereochemical reasons, as firstly described by Wald, but that of the amino acids is the result of natural selection acting during the peptide-assisted stage of oligoribotide growth.

Introduction

A modest, properly limited theory of the origins of the genetic code, should be able to give an answer to the question posed by the title of this paper on the basis of our chemical knowledge of the present translation scheme. A first consideration is that the vast span of chemical possibilities for self-replicating systems (Tjivikna *et al.,* 1990) should be narrowed to materials of our modern biota and from fossils. Thus, since the code embodies a linear correspondence between polyribotides and amino acids, we must limit ourselves to these classes of biomolecules.

Back in 1961 Crick (Crick *et al.,* 1961) established that the code is a triplet one, with each consecutive triplet of ribotides in one DNA (or RNA) strand corresponding to a single amino acid in the protein. In other words, the genetic code is not the simplest possible one, which would correspond to 2^2 codons, leading to only 4 kinds of amino acids. This is certainly too small a number to build foldable proteins, according to general construction principles established by Kauzmann (1964) and which are being verified by new experimental techniques (Englander, 1994).

Reprint requests to Dr. R. Ferreira.
Telefax: 55-081-271-8442.

Why not 2^3 Codons?

A triplet code made of only 2 bases could codify for $2^3 = 8$ kinds of amino acids. From what we understand about protein folding (Kauzmann, 1964; Englander, 1994) eight is a reasonable number, since it allows for one residue of each kind found in proteins: small and large hydrophilic, small and large hydrophobic, basic, acidic, and S–S bond forming, with one codon left for a terminator factor. Perhaps the only limitation of such code is the absence of redundancy (Brillouin, 1962). Crick (1968), however, concluded that such parsimony was the reason for the early introduction of a 4-base code, instead of a simpler 2-base one, and his proposal became largely accepted. Our answer to this question is of a different nature and based in the idea that the primitive RNA world required four bases from the very start.

Why 4 Bases?

Turning to 4-base codes, one should rule-out a primitive *binary* code in our biota, containing $4^2 = 16$ codons. If such code occurred in the past, with each consecutive doublet in the RNA strand corresponding to one amino acid, it could have no continuity with the modern triplet code, since all the information content then available would be lost as the reading frame shifted from doublet to triplet (Crick, 1968). One cannot, on the other hand, rule-out a primitive 4-base *triplet* code in

0939–5075/95/0100–0148 $ 06.00

which every third position of the codons was non-discriminating. Such a possibility is suggested, for example, by the fact that in the modern code eight of the twenty amino acids are entirely determined by the first two positions of their codons. It should be emphasized that in the absence of more data one can only make conjectures about the evolution of the code. Thus, since amino acid recognition depends of t-RNA molecules that carry the anticodons, it is possible that during such evolution, the number of codons was fixed in 64, but the number of anticodons could have increased. Such is the archetypal code proposed by Jukes (1983) containing 16 anticodons, each paired with 4 codons for a single amino acid. A similar proposal was made in 1988 (Ferreira, 1988) according to which the third non-discriminating position of some codons could at some stage begun to be distinguishable, for example, as one-ring (pyrimidine) and two-ring (purine) compounds.

The evidence is now overwhelming that translation was preceded by a stage characterized by self-replicating RNA polymers (Kruger *et al.,* 1982; Guerrier-Takada *et al.,* 1983; Cech, 1986, 1987; Zaug and Cech, 1986; Doudna and Szostak, 1989; Orgel, 1987; Joyce, 1989). The chemical mechanism responsible for the replication of the chains is not known in detail, even because information on the precise physical conditions then obtaining is lacking. Because of the unfavorable free energy of condensation of ribotides in water, it is necessary to suppose that the RNA world started with activated dimers and trimers and that perhaps the medium was diluted water (De Meis, 1989). Simulation studies (Ferreira and Coutinho, 1993) have confirmed the result previously obtained by renormalization group theory (Tsallis and Ferreira, 1986): differential growth of oligoribotides arises from differences in the number and position of two interactions between the growing fragments: C–G (strong, s) and A–U (weak, w). Thus, the question of why four bases has a natural answer: selection occurs only if there are two pairs of interchain interactions, *i.e.,* two pairs of complementary bases.

Why a Ternary Code?

Our simulation studies (Ferreira and Coutinho, 1993) predict the growth of 2^n groups of oligoribotides, each containing 2^n oligomers (n is the size of the chains). Kinetic discrimination between the groups is achieved through different values of the pseudo-second order rate constant k_{cat}/K_M. The Michaelis constants (K_M) are proportional to the inverse of the product of the C,G and A,U interaction constants, normalized in the following way: if, in a growing oligoribotide, there are q(C,G) residues and m(A,U) ones, so that $m + q = n$, we take into account that the maximum number of residues per reactive sites is 3. Thus, we make $m' = 3m/(q + m)$ and $q' = 3q/(q + m)$, and write:

$$K_M \propto [K^{m'}{}_{AU}\, K^{q'}{}_{CG}]^{-1}. \quad (1)$$

Because $K_{CG} > K_{AU}$, C,G-rich segments were predominant over A,U-rich ones. This early imbalance may still be seen in the fact that the relation (C+G)/(A+U) is 1.21 in modern prokaryotes, but only 1.04 in exons of eukaryotes (Doolittle, 1989). As the size of the chains increased the kinetic discrimination between the groups decreased and self-replication became a less efficient mechanism for augmenting the information content. A similar behaviour has been found to occur in ribozyme kinetics (Herschlag, 1991). Further selective advantage will accrue for some members of a group if the corresponding k_{cat} increases. This would happen if the oligoribotides are mixed (through diffusion, convection, tide effects, *etc.*) with a solution of molecules capable to act as catalysts for the condensation reactions. Since the RNA world evolved to an RNA-protein one, such catalysts must have been related to the latter type of compounds. Single amino acids are not suitable catalysts; proteins were still in the future, waiting for a polyribotide-directed synthesis. The best candidates, then, are randomly synthesized peptides, perhaps the hydrolytic products of proteinoids (Fox and Dose, 1977).

Consider, for example, that the pentaribotide group which contains, among others, the pentamers 5′pUCAGG (I) and 5′pAGUCC (II), start interacting with an oligopeptide containing glycine residues in excess over the expected (average) composition ((I) and (II) belong to the same group because they have the same interaction pattern, 5′p-wswss). Such peptides would interact differently with the 3′-end of (I) and (II). Let us suppose that they will increase the rate of further condensation of (I), *viz.:*

$$5'pUCAGG + 5'pAUU \rightarrow H_2O + 5'pUCAGGAUU. \quad (2)$$

In the case of (II) the interaction with the glycine-rich peptide must be different, forming a reasonably stable complex at its 3′-end. This means that the adsorbed peptide would have its hydrolysis rate decreased by the presence in solution of pentaribotides with the 3′-end UCC, such as (II). Extrapolating from the known properties of enzymes such as RNA polymerases the transition state of the condensation reaction (2) involves three nucleotides, in this case pGGA, which has become a triplet present in the condensation product of (I).

Now, these processes did not established a permanent relation between the nature of the active sites of future enzymes and particular triplets in polyribotide chains. The constant shuffling along the evolutionary course saw to it. But an early *concentration correlation* would have been established between the codon GGA, the anticodon UCC, and glycine residues, and this reasoning can be extended to the other amino acids. If the correlations are to survive, diffusion had to be minimized, which calls for the existence of some sort of semipermeable membranes. In any case, it is seen that from the very beginning translation involved a ternary 4-base code.

Chance or Necessity?

In a classic work Woese (1967) proposed that the code originated from stereochemical requirements connecting codons to amino acids. This idea as well as its anticodonic modification (Dunnill, 1966; Grantham, 1974; Lacey and Weber, 1976; Junck, 1978; Lacey and Mullins, 1983; Weber and Lacey, 1978), or the distance minimization proposal (Sonneborn, 1965; Woese *et al.,* 1966), has not been supported by experimental evidences (Salemme *et al.,* 1977; Wong, 1980) and has barely survived Crick's opposition, stated particularly in his hypothesis (Crick, 1968) that the present code is the result of a frozen accident. The fact remains, however, that only one genetic code (with one or two small variations) exists in our biota. The number of possible codes is enormous: 64 triplets coding for 20 amino acids and 2 terminators can give rise to $1.96197258 \ldots \times 10^{16}$ codes (Gomes, 1993). The hypothesis delineated in the previous section can be considered as a variation of the stereochemical models, since the main physical interaction occurs in the transition states of condensation reactions.

Direct evidence about the code goes back to only 40×10^6 years, whereas indications are that the code has remained essentially the same from the introduction of prokaryots, about 3.5×10^9 years ago. Thus the paradoxal conclusion that although it is possible to make testable models of the origin of the code, *i.e.,* of the translation scheme, it is very difficult to discuss the evolution of the code with some degree of confidence. In the absence of more information, through fossil records, speciation in particular niches, *etc.,* the evolution of the code is a problem at the limits of the art of the soluble (Medawar, 1967).

Dual Origin of the Monochirality in our Biota

The idea that the very early introduction of self-replicating RNA systems preceded the existence of proteins gains further support from stereochemical considerations. In 1957 Wald (1957, 1964) showed, using solid models, that polyribotides have very stringent stereochemical requirements: no helix is formed if C(3′) and C(4′) of the ribose moiety have opposite chiralities, and no base pairing is possible if, in addition, the C(1′)s are not all of the some chirality (β-glycosidic bonds). This agrees with Frank's proposal (Frank, 1953) that given an initial difference in the number of D and L isomers in the chain composition $\{(n_D - n_L)_{t=0} \neq 0\}$, a polymer will continue to grow by incorporating the predominant enantiomer out of a racemic solution, if the opposite enantiomer inhibits the chain growth.

Wald has also suggested that the monochirality of amino acids in proteins could have had a similar origin, and, in fact, folding of relatively large peptide chains in the tertiary structures of proteins requires strict monochirality. However, in the case of small peptide chains, the presumed ancestors of foldable proteins, it is possible to show that stable secondary structures can be built from L+D amino acids. For example, in a simulation using Program TOM we have shown that in the case of α-helices, peptide chains of considerable length (in our study, up to 11 residues) can be built with very small changes of the torsional angles Φ and Ψ by adding both L and D enantiomers (Lins *et al.,*

1994). In this case one enantiomer does not inhibit the binding into the chain of the opposite one, and the monochirality of amino acids could not have had the same origin as that of the ribotides.

In our proposed mechanism for the origin of a codon-amino acid correlation (Ferreira and Coutinho, 1993), small peptides of a given composition act as catalysts for the growth of certain oligoribotides. Since these latter molecules are monochiral, the selected peptides had to be built from monochiral residues in order to be efficient catalysts. Thus, the requirement of amino acid monochirality for the proper folding of proteins could be met by the previously selected peptide fragments and/or their hydrolytic products. It seems, therefore, that the observed monochirality of the proteinaceous amino acids is the result of Darwinian selection at the molecular level.

Acknowledgement

This work was supported in part by CNPq (Brazilian Government Agency).

Brillouin L. (1962), Science and Information Theory. Academic Press, New York, 2nd Edition.
Cech T. R. (1986), A model for the RNA-catalyzed replication of RNA. Proc. Natl. Acad. Sci. U.S.A. **83,** 4360–4363.
Cech T. R. (1987), The chemistry of self-splicing RNA and RNA enzymes. Science **236,** 1532–1539.
Crick F. H. C., Barnett L., Brenner S. and Watts-Tobin R. J. (1961), General nature of the genetic code for proteins. Nature **192,** 1227–1232.
Crick F. H. C. (1968), The origin of the genetic code. J. Mol. Biol. **38,** 367–379.
De Meis L. (1989), Role of water in the energy of hydrolysis of phosphate compounds. Biochim. Biophys. Acta **973,** 333–349.
Doolittle R. F. (1989), Prediction of Protein Structure (G. D. Fasman, ed.). Plenum Press, New York, pp. 599–623.
Doudna J. A. and Szostak J. W. (1989), RNA-catalyzed synthesis of complementary strand RNA. Nature **339,** 519–522.
Dunnill R. (1966), Triplet nucleotide-amino acid pairing. Nature **210,** 1267–1268.
Englander S. (1994), In pursuit of protein folding. Science **262,** 848–849.
Ferreira R. (1988), A conjecture on the early evolution of the genetic code. Leite Lopes Festschrift. World Scientific Co., Singapore, pp. 241–252.
Ferreira R. and Coutinho K. R. (1993), Simulation studies of self-replicating oligoribotides, with a proposal for the transition to a peptide-assisted stage. J. Theor. Biol. **164,** 291–305.
Fox S. W. and Dose K. (1977), Molecular Evolution and the Origin of Life. Dekker, New York.
Frank F. C. (1953), On spontaneous asymmetric synthesis. Biochim. Biophys. Acta **11,** 459–463.
Guerrier-Takada C., Gardiner K., Marsh T., Pace N. and Altman S. (1983), The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell **35,** 849–857.
Gomes M. A. F. (1993), private communication.
Grantham R. (1974), Amino acid difference formula to help explain protein evolution. Science **185,** 862–864.
Herschlag D. (1991), Implications of ribozyme kinetics for targeting the cleavage of specific RNA molecules *in vivo:* more isn't always latter. Proc. Natl. Acad. Sci. U.S.A. **88,** 6921–6925.
Joyce G. F. (1989), RNA evolution and the origins of life. Nature. **338,** 217–224.
Jukes T. H. (1983), Evolution of the amino acid code: inferences from mitochondrial codes. J. Mol. Evol. **19,** 219–225.
Junck J. R. (1978), The genetic code as a periodic table. J. Mol. Evol. **11,** 211–224.
Kauzmann W. (1964), Tridimensional structures of proteins. Biophys. J. **4,** 43–53.
Kruger K., Grabowski P. J., Zaug A. J., Sands J., Gottschling D. E. and Cech T. R. (1982), Self-splicing RNA: autoexcision and autocatalyzation of the ribosomal RNA-intervening sequence of *Tetrahymena.* Cell **31,** 147–157.
Lacey J. C., Jr. and Weber A. L. (1976), The origin of the genetic code: an amino acid-anticodon relationship. In: Protein Structure and Evolution (J. L. Fox, Z. Deyl and A. Blazy, eds.). Dekker, New York, pp. 213–222.
Lacey J. C., Jr. and Mullins D. W., Jr. (1983), Experiments related to the origin of the genetic code and the process of protein synthesis. Origins of Life **13,** 3–42.
Lins R. D., Soares T. A., Garratt R. C. and Ferreira R. (1994), On the stability of α-helices built from L + D amino acids. An. Acad. Bras. Ci., to appear.
Medawar P. B. (1967), The Art of the Soluble. Methuen, London, p. 7.
Orgel L. E. (1987), Evolution of the genetic apparatus: a review. Cold Spring Harbor Symp. Quant. Biol. **52,** 9–16.
Salemme F. R., Miller M. D. and Jordan S. R. (1977), Structural convergence during protein evolution. Proc. Natl. Acad. Sci. U.S.A. **74,** 2820–2824.

Sonneborn T. M. (1965), Evolving Genes and Proteins (V. Bryson and H. J. Vogel, eds.). Academic Press, New York, pp. 322–397.
Tjivikna T., Ballester P. and Rebek J., Jr. (1990), A self-replicating system. J. Am. Chem. Soc. **112,** 1249–1250.
Tsallis C. and Ferreira R. (1986), A critical phenomenon approach to biogenesis. Physica **140-A,** 348–360.
Wald G. (1957), The origin of optical activity. Ann. N.Y. Acad. Sci. **69,** 352–368.
Weber A. L. and Lacey J. C., Jr. (1978), Genetic code correlations: amino acid and their anticodonic nucleotides. J. Mol. Evol. **11,** 199–210.
Woese C. R., Dugre D. H., Dugre S. A., Kondo M. and Saxinger W. C. (1966), The molecular basis of the genetic code. Proc. Natl. Acad. Sci. U.S.A. **55,** 966–974.
Woese C. R. (1967), The Genetic Code. Harper & Row, New York, pp. 179–195.
Wong J. T. (1980), Role of minimization of chemical distances between amino acids in the evolution of the genetic code. Proc. Natl. Acad. Sci. U.S.A. **77,** 1083–1086.
Zaug A. J. and Cech T. R. (1986), The intervening sequence RNA of *Tetrahymena* is an enzyme. Science **231,** 470–475.