The Patterns of Base Sequences in the Nucleic Acids of Prokaryotes and Eukaryotes Reflect Features of their Abiotic Past

Frederico Pontes, Victor Rusu, Benício de Barros Neto, and Ricardo Ferreira*

Departamento de Química Fundamental, Universidade Federal de Pernambuco,
40060-901 Recife, Pernambuco, Brazil. E-mail: rferreira100@yahoo.com

* Author for correspondence and reprint requests

Z. Naturforsch. **63c**, 898–902 (2008); received March 10/May 28, 2008

The base sequences of the nucleic acids corresponding to ten proteins (aconitase, alcohol dehydrogenase, enolase, fumarase, isocitrate dehydrogenase, lactate dehydrogenase, phosphofructokinase, phosphoglycerate mutase, pyruvate kinase and succinate dehydrogenase) belonging to a total of 154 species, ranging from prokaryotes to vertebrates, were compared with the base sequences of oligoribotides whose growth rates were calculated by a chemical kinetics model. It was shown that oligoribotides grown according to the kinetics model have a fraction of repetitive bases larger than expected from random processes. The base sequences of nucleic acids of prokaryotes and eukaryotes retain, in decreasing proportions, this feature of their abiotic past. Chemically synthesized pentameric stretches with repetitive bases are slightly more abundant than those present in prokaryotes. Genetic drift and natural selection, operating as fundamental laws even for the most primitive living systems, reduced the original, chemically controlled, repetitive base frequency in prokaryotes, which was further reduced for eukaryotes.

Key words: Nucleic Acids, Oligoribotide, Abiotic Past

Introduction

According to the Continuity Hypothesis, biomolecules found in present-day biota must be directly connectable to the prebiotic, chemically synthesized molecules. Considering, for example, the oligoribotide chains, one might expect that the base sequences of nucleic acids of contemporary taxa, obtained from databases such as that at the NCBI (National Center for Biological Information, USA), should retain some features of their chemically synthesized ancestors. In the analysis presented here we shall assume the reasonable (though unproved) supposition that RNA-like molecules were the first biopolymers (Woese, 1967; Crick, 1968; Orgel, 1968). It is well known that the posterior discovery of ribozymes (Cech, 1982; Guerrier-Takada *et al.*, 1983) lent further credibility to this "RNA world" scenario.

To test the continuity idea, we should make a comparison between the nucleic acids base sequences present in living cells and those of oligoribotides produced *in vitro*. However, the actual base sequences of synthetic oligoribotides are not generally known (Von Kiedrowski, 1986). We must, therefore, use values for the oligoribotide growth rates and base sequences estimated from theoretical models.

The calculations to obtain these values were based on a previously proposed theoretical model for the growth of RNA-like oligomers (Ferreira and Coutinho, 1993), which will be succinctly described later. According to this model, oligomers grow by condensation reactions between two small random oligoribotides (dimers and trimers), following a Michaelis-Menten mechanism in which the catalytic active site is another triribotide bound to a mineral surface such as clay, as first suggested by Bernal (1951). We further proposed that the four bases, A, G, C, and U, could be grouped in two sets: a *strong* one (G, C), denoted by the letter *s*, and a *weak* one (A, U), denoted by *w*. This hypothesis, which has also been adopted by Miramontes *et al.* (1995), is in accordance with the fact that in the proposed chain-growth mechanism the four ribotides are assumed to interact either strongly (G---C) or weakly (A---U). This simplification allows us to considerably reduce the number of possible polymeric chain stretches. For the pentaribotides, for example, the number of chain stretches is diminished from 1024 (that is 4^5) to just 32 (2^5).

In this paper we statistically compare the base sequences of pentameric stretches in the calculated RNAs with those present in existing RNAs. Due to the continued divergence of pre-existing

0939–5075/2008/1100–0898 $ 06.00

Table I. Student's *t*-tests for the mean frequencies of occurrence, on individual proteins, of continuous pentameric sequences (*sssss*, *ssssw*, *wssss*, *wwwww*, *wwwws* and *swwww*) in bacteria (B) and eukaryotes (E).

Enzyme	Mean B	Mean E	Difference	No. of B	No. of E	*p*
Aconitase	0.237	0.175	0.062	5	12	0.051
Alcohol dehydrogenase	0.208	0.170	0.038	6	18	0.102
Enolase	0.144	0.147	– 0.003	6	21	0.874
Fumarase	0.251	0.150	0.101	8	10	0.000003
Isocitrate dehydrogenase	0.206	0.187	0.019	8	19	0.191
Lactate dehydrogenase	0.253	0.184	0.069	12	22	0.00011
Phosphofructokinase	0.287	0.195	0.092	11	5	0.067
Phosphoglycerate mutase	0.234	0.150	0.084	5	11	0.00065
Pyruvate kinase	0.225	0.163	0.062	11	15	0.0183
Succinate dehydrogenase	0.204	0.173	0.031	7	12	0.368

taxa, resulting from mutations and natural selection, it is expected that RNAs from prokaryotes should bear a closer relation to the base distribution of calculated oligoribotides than RNAs from the presumably more recent eukaryotes. Since different proteins evolve at different rates (Kimura, 1968), it is natural that the comparisons should be made on the same protein, over different taxa. In this paper we make such comparisons for 10 different proteins, distributed over a total of 154 different species. However, since each protein of our set is represented by a relatively small sample, we also carry out a comparison on the whole set, to increase the number of degrees of freedom – and the degree of confidence – of the statistical results.

Materials and Methods

We started from the RNA counterparts of 10 representative proteins, all of them connected with fundamental metabolic functions, such as glycolysis and the Krebs cycle. The protein set represents 91 eukaryotes, 56 bacteria and 7 archaebacteria. The proteins were aconitase, alcohol dehydrogenase, enolase, fumarase, isocitrate dehydrogenase, lactate dehydrogenase, phosphofructokinase, phosphoglycerate mutase, pyruvate kinase and succinate dehydrogenase (Table I). The information on the corresponding base composition was taken from the NBCI database.

A computer program was written to print out each individual RNA composition in terms of pentameric stretches. Hexameric and heptameric sequences were also analyzed, but the results do not differ from the pentameric ones, except that for these larger stretches (and probably also for the still larger ones) the occurrence of repetitive domains decreases sharply.

The pentameric data table was converted to a set of relative frequencies of occurrence corresponding to each RNA, and then subjected to separate principal component analyses (PCAs), to search for possible patterns of association between specific stretches belonging to differing taxa.

In a PCA the original data matrix is projected onto a subspace defined by linear combinations of the original variables with maximal variance – that is, maximal information – and orthogonal to each other (Jolliffe, 2002). Each component is characterized by three mathematical entities: (a) the percent amount of explained variance/information; (b) a loadings vector, whose elements are the cosines of the angles the principal component (PC) axis forms with the original axes; (c) a scores vector, containing the coordinates locating the individual species in the PC axis. All calculations and graphs were made with the Statistica 6.1 computer program (StatSoft Inc., 2004).

We have considered the following as *repetitive pentameric stretches*: *5′p-sssss* and *5′p-wwwww*, *5′p-swwww*, *5′p-wssss*, *5′p-ssssw* and *5′p-wwwws.* If the five A, U, T, G, C were obtained by a complete random process, these six stretches should occur with a frequency of 6/32 = 0.1875. As shown in Tables I and II, they occur with higher frequen-

Table II. Student's *t*-test for the mean frequencies of occurrence, on the entire protein set, of continuous pentameric sequences (*sssss*, *ssssw*, *wssss*, *wwwww*, *wwwws* and *swwww*) in bacteria (B), eukaryotes (E) and archaea (A).

Taxus	Mean	Number	*p*
Bacteria (B)	0.230	79	$1 \cdot 10^{-13}$ (B vs. E)
Archaea (A)	0.230	16	$2 \cdot 10^{-6}$ (A vs. E)
Eukaryotes (E)	0.169	145	

cies than those expected for random processes in the case of bacteria and archaebacteria (0.230) but lower than expected in the case of eukaryotes (0.169).

The frequencies of the repetitive pentameric base sequences constitute the empirical data of our study. To test the Principle of Continuity between the Chemical and the Biological Evolution, the base sequences found in prokaryotes and eukaryotes should be compared with the corresponding frequencies in calculated oligoribotides, which are presumably similar to the primitive, chemically grown oligoribotides.

In the PCA each enzyme (for a given species) is represented by a vector in a 32-dimensional space whose elements are the frequencies of occurrence of the 32 pentameric stretches, normalized to a total of 100 %. The sizes of the matrices were determined by the data available, and varied considerably from one enzyme to another. For example, while lactate dehydrogenase was represented by pentamers from thirty-nine species (5 archaebacteria, 12 eubacteria and 22 eukaryotes), the data for phosphofructokinase included only 11 bacteria and 5 eukaryotes.

Ten separate PCAs were performed, one for each enzyme. Despite the variability of the information available for the different enzymes, the results of all analyses were strikingly similar. The first principal component, which in this case reproduces almost 80 % of the total information, is clearly due to the separation between the *s*-rich pentamers, which are clustered with positive loadings, and the *w*-rich pentamers, which are clustered on the left-hand side of the plot, as in a mirror image of the first cluster (Fig. 1a). However, this is not associated with any differentiation between bacteria and eukaryotes, as can be seen from Fig. 1b, where the scores along PC1 clearly mix the two taxa.

The pattern along the second PC axis (16.4 % information) is remarkably different. All pentamers have negative loadings, indicating that species with higher proportions of either *s*-rich or *w*-rich pentamers (or both) will tend to have more negative scores on the second PCA (PC2). Moreover, this appears to be associated with some degree of separation between eukaryotes (whose scores on PC2 are, with a single exception, all positive) and bacteria, several of which have highly negative PC2 scores. This suggests that there might be a statistically significant difference between the mean values of frequencies of occurrence of continuous sequences in the two taxa. Since bacteria have more negative scores than eukaryotes, and the PC2 loadings are all negative, we expect the mean values for bacteria to be higher than the corresponding values for eukaryotes. This hypothesis

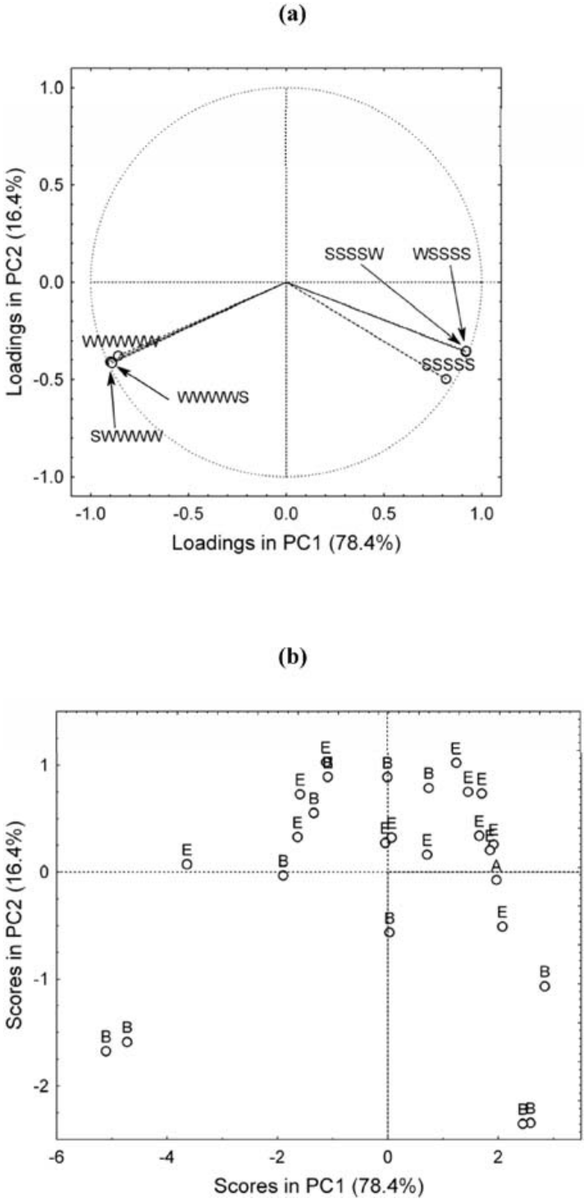


Fig. 1. Results of the principal component analysis of the pyruvate kinase data based on the relative frequencies of the pentamers *sssss*, *ssssw*, *wssss*, *wwwww*, *wwwws* and *swwww*. (a) Loadings plot. (b) Scores plot.

was tested with Student's *t*-tests for each enzyme, using the original frequency data. The results of these tests are shown in Table I.

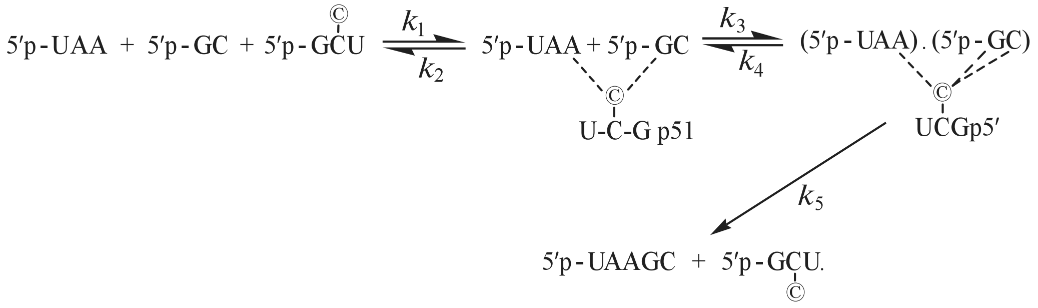
With the exception of enolase, for which a slightly negative difference favours eukaryotes, the means for bacteria are indeed higher. The statistical significance of the difference, however, varies from one enzyme to another, as shown by the *p*-values. The worst case is, of course, enolase, for which $p = 0.874$. In contrast, for pyruvate kinase the difference favouring bacteria is significant at the 98 % confidence level.

It is clear from Table I that many of the mean frequency comparisons are based on very few degrees of freedom, especially as regards the number of bacterial enzymes. To obtain a result based on a larger number of degrees of freedom, we have also run *t*-tests on the entire frequency data set, obtaining the numerical values given in Table II. The mean frequency observed for eukaryotes is significantly lower than those observed for either bacteria or archaea ($p < 10^{-13}$). The mean frequencies for these latter taxa, however, are indistinguishable.

A Kinetic Model for Oligoribotide Growth

To test the Principle of Continuity between the Chemical and the Biological Evolutions we should compare the frequencies of pentameric base stretches of actual oligoribotides, shown in Tables I and II, with the corresponding frequencies observed in modeled oligoribotides, which are expected to closely resemble those of chemically grown oligoribotides. One of the intrinsic difficulties of experiments on the oligomerization of ribotides is that the environmental conditions of the prebiotic phase encompass a very broad range of variables, such as temperature, pressure, acidity, ionic strength, presence of metallic cations, presence of hydrophilic solvents (such as ethylene glycol).

Faced with these (up to date) unsurmontable difficulties, we decided to obtain chemical data for frequency comparison using a kinetic model proposed by us (Ferreira and Coutinho, 1993), in which the growth of RNA-like oligomers takes place by condensation reactions between two small random fragments (di- and triribotides), following a Michaelis-Menten mechanism, catalyzed by a third triribotide which benefits from its binding to a mineral fragment. For each of the synthetic pathways for a given oligomer we write the corresponding Michaelis-Menten rate expression. Thus, for the growth of 5′p-UAAGC we write the following sequence of steps:



(1)

According to this mechanism, $K_{M1} = k_2/k_1 = (K_{AU})^{-1}$ and $K_{M2} = k_3/k_4 = (K_{GC})^{-2}$. Therefore, the formation of 5′p-UAAGC follows the rate equation

$$v = k_5\, K_{AU}\, K^2_{GC} [5'\text{p-UAA}]\, [5'\text{p-GC}].$$

We make the assumption that k_5, the rate constant for the (almost) irreversible step of the Michaelis-Menten mechanism, is independent of the specific nature of the ribotides involved, since the different bases A, U, G and C are approx. 5 Å away from the O(3′) and O(5′) ribose atoms.

The two equilibrium constants, K_{AU} and K_{GC}, are the formation constants of the complementary pairs in the corresponding DNA-DNA interactions. The inverses of these constants are the Michaelis-Menten constants for reaction (1). The base pair guanine-cytosine is called strong (*s*), whereas the adenine-uracyl pair is called weak (*w*). As stated before, this is the basis for reducing the $4^5 = 1024$ pentameric sequences down to just $2^5 = 32$ sequences.

The justification for this differentiation is, first of all, that the GC pair in DNA has *three* hydrogen bonds, whereas the AT pair in DNA is formed by two hydrogen bonds. There are also some calculations (Yanson *et al.*, 1979) showing that the G-C bond is twice as strong as the A-T bond. In this paper the ratio K_{GC}/K_{AU} was varied between 3/2 and 2/1.

The observation that repetitive pentameric stretches such as 5′p-GCCGC and 5′p-AUUAU in prokaryotes occur with a frequency larger than the value expected from random processes (0.1875) can only be explained, according to the Principle of Continuity between the Chemical and the Biological Evolutions, if the concentration of A + U in the original environment is *larger* than the concentration of the complementary base pair, G + C. The free energies of formation of (G + C) and (A + U) are undoubtedly different, the A + U pair being the more stable one of the two. The physico-chemical conditions prevailing on the Earth about $3 \cdot 10^9$ years are of course beyond our knowledge, but we have found that, taking the ratio K_{GC}/K_{AU} as either 3/2 or 2/1, and the concentration ratio [A + U]/[G + C] as high as 2.8, the repetitive pentamers *5′p-sssss, 5′p-wssss, 5′p-ssssw, 5′p-wwwww, 5′p-swwww and 5′p-wwwws* grow with frequency values of 0.248 and 0.242, respectively, which is in good agreement with the results of our statistical analysis. Therefore, we have assumed in the present work that the ratio [A + U]/[G + C] is equal to three.

Conclusions

It has been shown (Ferris *et al.*, 1996) that it is possible to synthesize relatively large oligoribotides using mineral surfaces as catalysts. By this technique of successive feeding, for example, it was possible to synthesize oligomers up to 55 monomers long. The base sequences of these oligoribotides, however, could not be ascertained.

One fundamental difficulty faced by proposals addressing the origin of the oligoribotides is the uncertainty concerning the prevailing physical properties and chemical composition under prebiotic conditions. We have tried to circumvent this problem by using a previously published theoretical model to calculate the growth rates of the differing ribotide chains. The basic experimental data of the present-day biota point to the fact that pentameric stretches of nucleic acids contain repetitive bases (either *s* or *w*) with a frequency that is larger than that expected for random processes in the case of prokaryotes, and that this frequency decreases to smaller values than expected in the case of eukaryotes. This difference has never been pointed out before, probably because the concept of weak/strong bases has not been recognized, even though it has also been advanced by Miramontes *et al.* (1995).

Acknowledgements

This work was partially supported by research and scholarship grants from the Brazilian Agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Bernal J. D. (1951), The Physical Basis of Life. Routledge & Kegan Paul, London.

Cech T. R. (1982), A model for the RNA-catalyzed replication of RNA. Proc. Natl. Acad. Sci. USA **83**, 4360–4363.

Crick F. H. C. (1968), Origin of the genetic code. J. Mol. Biol. **38**, 367–379.

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.

Ferris J. P., Hill A. R., Liu R. H., and Orgel L. E. (1996), Synthesis of long prebiotic oligomers on mineral surfaces. Nature **381**, 59–61.

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.

Jolliffe I. T. (2002), Principal Component Analysis, 2nd ed. Springer, New York.

Kimura M. (1968), Evolutionary rate at the molecular level. Nature **217**, 624–626.

Miramontes P., Medrano L., Cerpa C., Cedergren R., Ferbeyre G., and Cocho G. (1995), Structural and thermodynamic properties of DNA uncover different evolutionary histories. J. Mol. Evol. **40**, 698–704.

Orgel L. E. (1968), Evolution of the genetic apparatus. J. Mol. Biol. **38**, 381–393.

StatSoft Inc. (2004), Statistica 6. Tulsa, OK, USA.

Von Kiedrowski G. (1986), A self-replicating hexadeoxynucleotide. Angew. Chem. Int. Ed. Engl. **25**, 932–935.

Woese C. R. (1967), The Genetic Code. The Molecular Basis for Genetic Expression. Harper and Row, New York.

Yanson I. K., Teplisky A. B., and Sukhodub L. F. (1979), Experimental studies of molecular interactions between nitrogen bases of nucleic acids. Biopolymers **18**, 1149–1170.