ISSN : 0974 - 7532 Volume 10 Issue 6

Research & Reviews in

BioSciences

Regular Paper

RRBS, 10(6), 2015 [231-234]

Informational polymers in the primitive Earth: Lempel-Ziv complexity and Entropy of tRNAs in anaerobic *Archea* compared to computer-generated random data

G.Bianciardi

Dpt. Medical biotechnology, University of Siena, Via delle Scotte 6, 53100 Siena, (ITALY)

E-mail: giorgio.bianciardi@unisi.it

ABSTRACT

In a possible scenario for the origin of life there was a time in which informational genetic polymers played primary roles. Transfer-RNA molecules may be considered possible "fossils" of that primitive era (billions years ago). In order to test the hypothesis that the nucleotide sequences of the primitive informational polymers might not be chosen randomly, we propose a comparison of computer-generated random sequences with tRNAs nucleotide sequences present in anaerobic *Archaea*. Random sequence data were obtained from the algorithm by Press and Teukolsky. Relative Lempel-Ziv complexity and Entropy of nucleotide sequences and of computer-generated random data represented as random walks were evaluated. Manhattan and Euclidean "fractal" dimensions (DM, DE) were also evaluated. Nonlinear parameters obtained from the *Archaea* are lower than the values of randomly generated sequences ($p<0.01$, $p<0.01$). DM and DE confirm the results ($p<0.01$, $p<0.01$). The observed deviation from pure randomness should be arisen from some constraints like the secondary structure of this biologic macromolecule and/or from a "frozen" stochastic transition when informational polymers originated. The nonlinear evaluation of nucletoide sequences expressed as random walk, here presented, provides an efficient assessment of the primary structure of nucleic acid sequences.

INTRODUCTION

A possible scenario for the origin of life needs an informational genetic polymer and an efficient prebiotic synthetic route to the component monomers. Nothing is known about the way(s) from which life born, and plausible pathways of prebiotic evolution remain obscure, however, in that context, t-RNA is considered the oldest known informational genetic polymer[1].

Billions years ago, according to the exon theory of genes[2] small RNAs translated into peptides of 15-20 amino acids: minigenes of pre-tRNAs codifying RNA hairpin structures. The dimerization of two equal RNA hairpin structures may have lead to the formation of the cruciform structure of the tRNA molecule: tRNA reflects the primordial genes of that era. Anaerobic prokaryotic cells lived during the earliest time (3.5 billion years ago): these microorganisms may retain ancestral signatures[3,4]. In order to test the hypothesis that the nucleotide sequences of the primitive informational polymers might not be chosen randomly, we performed a comparison of computer-generated random sequences with tRNAs nucleotide sequences present in anaerobic *Archaea*. Our method provides an efficient assessment of the primary structure of nucleic acid sequences.

MATERIALS AND METHODS

Based on the graphical approaches by Hamori & Ruskin[5] and Mizrahi & Ninio[6], we have performed nonlinear analysis of the nucleotide sequences of nonintronic tRNAs and of computer-generated random data described as random walks[7] by means of softwares developed by us in Visual Basic language. Briefly, orbit walks of nucleotide sequence data were obtained by letting the orbit walk a unit step in one of four directions (down, left, right, and up), depending upon the next base (A,C,G, T) in the sequence, and the distances from the origin calculated.

Samples

Thirty tRNAs (tRNA-gly, tRNA-val, tRNA-glu, tRNA-ala, t-RNA-arg) obtained from 6 anaerobic Archaea (*Methanocaldococcus jannaschii, Methanoregula boonei, Archaeoglobus fulgidus, Haloarcula marismortui, Hyperthermus butylicus, Pyrococcus furiosus*) were evaluated. Nucleotide sequence data of archaeal tRNA genes were obtained from the Institute for Genomic Research (TIGR)[8] and the GenBank library[9].

Random sequences (white noise) were obtained from the algorithm by Press and Teukolsky[10] and their orbit walks were obtained generating an uniformly and randomly distributed data points over the unit interval (0 to 1). The sequence was divided in equal intervals to which A,C,G,T letters were attributed, orbit walks and distances data were obtained as above. Twenty-five random sequences (lenght, n = 80) were evaluated.

Nonlinear analysis

Relative LZ complexity, LZ

Relative LZ complexity is a measure of the algorithmic complexity of a time series[11]. According to the Kaspar and Schuster algorithm, each data point is converted to a single binary digit according to whether the value is less than, or greater than, the median value of a set of data points.

White noise (a pure random signal, common in physical systems, that exhibits equal power across all the component frequencies of the signal), has an LZ value that is close to 1.0. Pink noise (flicker noise or 1/f noise), exhibits decreasing power as frequency increases, and is associated with a relatively low LZ value; it is common in biological systems (e.g. heart rate). A sine function with 10% superimposed Gaussian white noise yields an LZ value that is close to zero.

Entropy, K

The entropy index chosen here is a measure of the disorder in a data set and was calculated as the sum of the positive Lyapunov exponents[12].

Randomness is indicated by numerically high values of entropy. Ordered series like the sine function exhibit values that are close to 0.

In the present work, Relative Lempel-Ziv and Entropy indexes were evaluated over the distances of the obtained random walks, making use of Chaos Data Analyzer Pro v. 2.0 software[13].

“Fractal” dimensions of nucleotide sequences

To measure the scale properties of tRNAs and of random sequence data, the total Manhattan path lenght divided by the logarithm of the Manhattan distance of the endpoint from the origin (called “fractal” Manhattan dimension)

$$\mathbf{DM = \log[n(A)+n(G)+n(T)+n(C)]/\log[| n(G)-n(C) |+| n(T)-n(A) |]}$$

and the corresponding Euclidean “fractal” dimension:

$$\mathbf{DE = \log[n(A)+n(G)+n(T)+n(C)]/\log[SQR| n(G)-n(C) |^{2}+ SQR| n(T)-n(A)|^{2}]}$$

were calculated from the sequences, as proposed by Gates[14].

STATISTICAL TESTS

Non parametric Mann-Whitney test was used to ascertain the difference between the groups

RESULTS

Orbit walks of archaeal tRNA showed a different behaviour than the ones of computer-generated random sequences, the former appearing “less dispersed” that the latter (Figure 1). This different behaviour of the orbit walks between the two groups were quantitatively assessed by nonlinear analysis,

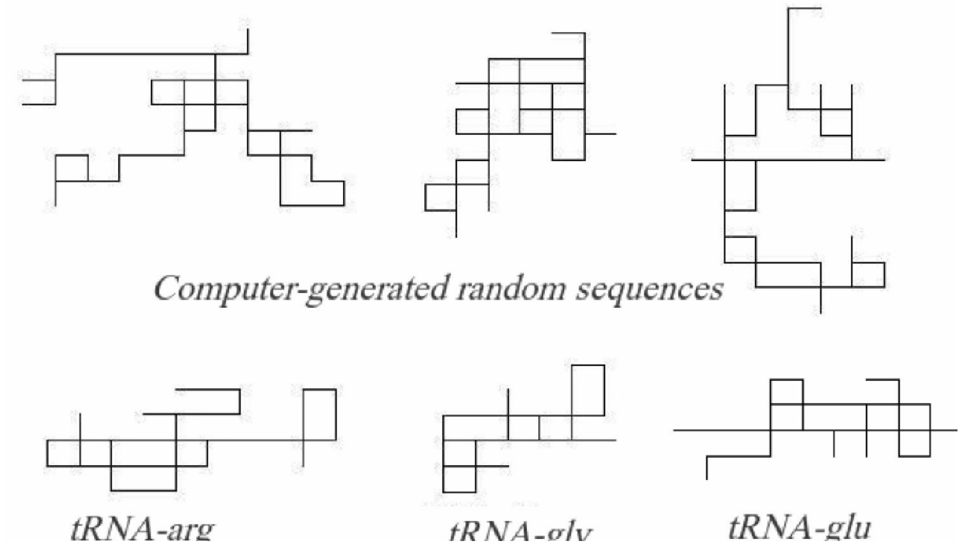

Figure 1 : Graphical representations of of computer-generated random data (top) and tRNAs, (bottom) as random walks. The sequence begins at the origin in the lower right corner of the graph. Transfer-RNAs random walks appears less "dispersed" than computer-generated random sequences.

TABLE 1 : Transfer-RNAs nucletotide sequences of anaerobic *Archaea* vs. computer-generated random data (mean values [SD]). Nonlinear indexes evaluated over the random walks, random data, n=25, (white noise), archaeal

LZ	**0.84 [0.05]**	**0.47 [0.04]**	****p<0.01**
K	0.51 [0.05]	0.35 [0.04]	** p < 0.01

Lempel-Ziv (LZ) and Entropy (K) of the distances calculated over the random walks of archeal tRNAs present significant lower values compared to the ones of random sequence data. Transfer-RNAs sequences are more ordered than random data.

TABLE 2 : Transfer-RNAs nucletotide sequences of anaerobic *Archaea* vs. computer-generated random data (mean values ± SD). Fractal dimensions, random data, n=25, (white noise) archaeal tRNA, n=30

DM	**3.7 [0.3]**	**2.72 [0.02]**	**** p < 0.01**
DE	4.5 [0.4]	2.9 [0.4]	** p < 0.01

Manhattan (DM) and Euclidean (DE) fractal dimensions of the distances calculated over the random walks of archeal tRNAs present significant lower values compared to the ones of random sequence data. Transfer-RNAs sequences are more ordered than random data.

where Lempel-Ziv and Entropy indexes appeared statistically lower in the archaeal tRNA compared to the random data (p<0.01, p<0.01, TABLE 1). Likewise, "Fractal" Dimensions resulted lower than random data (p<0.01, p<0.01, TABLE 2).

DISCUSSION

We have introduced the nonlinear analysis of sequences described as random walks, a method that appears capable of highlighting primary structural features of the nucleic acids. We have applied it to the study of tRNAs carrying older-considered amino acids[15] presents in genomes of ancient microorganism (anaerobic *Archaea*, e.g., *Pyrococcus f.*[3], here tested). We found that this ancient informational polymer presents significant lower values of Lempel-Ziv and Entropy indexes, measure of algorithmic complexity and disorders, respectively, than the ones of random data (white noise). Interestingly, Manhattan and Euclidean "fractal" dimensions, also, resulted lower than the ones of random data (white noise), confirming the results obtained by the nonlinear analysis of the distances calculated over the random walks. Both the results reveal a significant shift of the tRNA nucleotide sequences from pure randomness, i.e.: a more ordered structure, than a "pure" random sequence.

The observed deviation from pure randomness may be arisen from some constraints like the secondary structure of this biologic macromolecule, or from the frozen stochastic transition from which life had its origin. We may recall, Gayle and Freeland[16] that showed that the 20 amino acids present in the

Last Universal Common Ancestor, cell that lived perhaps 3 or 4 billion years ago, were not chosen randomly and O.Weiss et al.[17] that evidenced a significant small reduction of the Shannon entropy (-1%) in protein sequences compared to random polypeptides. Together with our results, these data indicate that evolution earlier chose nonrandom "alphabets".

CONCLUSION

We have introduced the nonlinear analysis of nucleic acid sequences described as random walks that appears to provide an efficient assessment of the primary structure of nucleic acid sequences. Applying them to the study of "ancient" tRNAs, a fundamental molecule for the study of the origin and evolution of life, their nucleotide sequence showed a significant shift from randomness, results that may be of relevance in evolutionary studies

REFERENCES

[1] M.Eigen, B.F.Lindemann, M.Tietze *et al*; Science, **244**, 673 **(1989)**.

[2] M.Di Giulio; J.Theor.Biol, **191**, 191 **(1999)**.

[3] J.L.Howland; "The surprising archaea", Oxford University Press, New York, **(2000)**.

[4] F.D.Ciccarelli, T.Doerks, C.von Mering *et al*; Science, **311**, 1293 **(2006)**.

[5] E.Hamori, J.Ruskin; J.Biol.Chem., **258**, 1318 **(1983)**.

[6] E.Mizrahi, J.Ninio; Biochimie, **67**, 445 **(1985)**.

[7] W.Feller; "An introduction to Probability Theory and its Applications", 3rd Edition, Wiley Series in Probability and Mathematical Statistics, Wiley, **(1968)**.

[8] http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi,

[9] http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank/

[10] W.H.Press, S.A.Teukolsky; Computers in Physics, **6**, 522 **(1992)**.

[11] F.Kaspar, H.G.Schuster; Physical Review A, **36**, 842 **(1987)**.

[12] P.Grassberger, I.Procaccia; Physical Review A, **28**, 2591 **(1983)**.

[13] J.C.Sprott, G.Rowlands; "Chaos data analyzer", Physics Academic Software, New York, **(1995)**.

[14] M.A.Gates; J.Theor.Biol., **119**, 319 **(1986)**.

[15] A.Gutiérrez-Preciado, H.Romero and M.Peimbert; Nature Education, **3(9)**, 29 **(2010)**.

[16] K.P.Gayle, S.J.Freeland; Astrobiology, **11**, 235 **(2011)**.

[17] O.Weis, M.A.Jimenez-Montano, H.Herzel; J.Theor.Biol., **206**, 379 **(2000)**.