J Mol Evol (1988) 27:109–113

Journal of
Molecular Evolution

Informational Parameters and Randomness of Mitochondrial DNA

M.I. Granero-Porati and A. Porati

Department of Physics, GNBC-CNR, University of Parma, Parma, Italy

Summary. The informational content of genomes of nuclear and mitochondrial origin is examined. By using the parameters of Shannon's information theory the language of mitochondrial DNA is shown to be more similar to the language of bacterial DNA than to that of nuclear DNA in more evolutionarily advanced animals. Moreover, using the parameters of Kolmogorov's theory on randomness, genes of different organisms (*Neurospora crassa* and *Saccharomyces cerevisiae*) coding for the same protein (subunit 9 of ATPase) are shown to have, if both of mitochondrial origin, a similar degree of randomness, whereas genes coding for the same protein, both belonging to the same organisms, exhibit a quite different degree of randomness when one is of mitochondrial origin and the other of nuclear origin. These results are in favor of the symbiotic origin of mitochondria.

Key words: Mitochondrial DNA — Symbiotic theory — Endogenous theory — Information theory — S-entropy — Randomness — K-entropy

Introduction

Application of information theory to the analysis of the genetic message began shortly after the discovery of the genetic code (Gatlin 1968, 1972). Since 1977, when the complete sequence of ΦX174 DNA was decoded (Sanger et al. 1977), the library of DNA sequences for various organisms has been greatly enriched, and now we have at our disposal, for theoretical analysis and interpretation, a great deal of data regarding not only sequences of nuclear origin, but also of mitochondrial origin.

Offprint requests to: M.I. Granero-Porati

The determination of the most significant informational parameters of DNA sequences from various organisms can help to assign to them a degree of "linguistic complexity," and to relate it to an evolutionary meaning. In particular, the linguistic analysis of the genetic message of mitochondrial origin could provide some insight into the origin of these eukaryotic organelles. It is well known, in fact, that there are two main theories on this subject: according to the first one, the so-called "symbiotic theory," the mitochondrion was originally a free-living bacterium, while, for the second one, the "endogenous" or "nonsymbiotic" theory, all genes, including those of the mitochondria, arose within the organism.

In the present work we perform an informational analysis of some mitochondrial DNA sequences in order to compare them with sequences of nuclear origin, and to obtain some new indications about the symbiotic or nonsymbiotic origin of the mitochondrion itself.

The well-known concept of entropy, as introduced by Shannon in the context of information theory, recently provided some interesting results regarding the analysis of exact DNA sequences. In particular, the analysis of viral genomes showed that overlapping genes influence information content and are indications of the degree of dependence in base sequences (Granero-Porati et al. 1980; Rowe and Trainor 1983). Moreover, informational measures of chromosomal and extra-chromosomal coding sequences (Lipman and Maizel 1982) have been done, confirming the validity of the use of information content of DNA as a suitable evolutionary measure (Subba Rao et al. 1982).

The main difficulty in this kind of statistical analysis lies in the relative shortness of the sequences. We feel that, in the case of short sequences, this

obstacle may be overcome by the use of a more recent concept of entropy developed by Kolmogorov (1968) and applied by Ebeling and Jimenez-Montaño (1980) to biological molecular sequences.

Analysis of Mitochondrial Sequences with the Aid of S-entropy

In order to analyze polynucleotide sequences, let us briefly recall some definitions, as introduced by Gatlin (1972) in the context of Shannon's information theory:

$$D_1 \text{ (= divergence from equiprobability)} = H_{max} - H_1$$
$$D_2 \text{ (= divergence from independence)} = H_2^{ind} - H_2^{dep}$$
$$R \text{ (= redundancy)} = (D_1 + D_2)/2,$$

where

$$H_1 = -\sum_{i=1}^{n} p_i \log_2 p_i;$$

$$H_{max} = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} \left(\text{when } p_i = \frac{1}{n} \forall_i \right)$$

$$H_2^{ind} = -\sum_{i,j=1}^{n} p_i p_j \log_2 p_i p_j;$$

$$H_2^{dep} = -\sum_{i,j=1}^{n} p_i p_{ij} \log_2 p_i p_{ij}$$

where p_i is the relative frequency of the i^{th} base of the given sequence, and p_{ij} the conditional probability that the base x_j follows the base x_i.

We performed a computer estimation of these parameters for some mitochondrial sequences, in order to compare our results with those already obtained by Gatlin (1972) relative to sequences of nuclear origin. It is to be noted that Gatlin's results were derived from nearest neighbor data, whereas ours were obtained from exact sequences. In particular, we examined the complete sequences of mitochondrial DNA in the following mammals: 1) *Mus musculus* (length 16,295 bases, Bibb et al. 1981), 2) *Homo sapiens* (length 16,569 bases, Anderson et al. 1981), and 3) *Bos taurus* (length 16,338 bases, Anderson et al. 1982).

The values of D_2 vs R for these mitochondrial DNAs (mtDNAs) and the values obtained by Gatlin relative to nuclear DNA extracted from organs of the same mammals are plotted in Fig. 1. From Fig. 1 one can easily see that in the (R, D_2) plane two clearly separated clusters of points are present: the first is formed of points belonging to mtDNA, the second to nuclear DNA. The cluster of mtDNA is characterized by a higher value of R and a lower value of D_2. Further, the values of D_2 for bacterial DNA (Gatlin 1972) are close to the values of mtDNA in mammals.

If we perform a similar analysis (i.e., D_2 vs R) for *single* mitochondrial genes, we notice (Table 1) that the values of D_2 and R are, in general, much higher than those found for complete mitochondrial sequences. In fact, sequences coding for tRNAs are also present in a complete mitochondrial sequence in which the nonspecific part, i.e., the one not subjected to specific rules, is intrinsically more "random" than the one of single genes. If we recall that D_2 means divergence from independence, and $R = (D_1 + D_2)/2$, it is clear that the insertion of random sequences has the effect of lowering the values of D_2 and R. It is interesting to note that we found the same results when some sequences of URF (unidentified reading frames, Table 2) were examined. Moreover, in some cases, the values of D_2 and R relative to the URF sequences are higher than those found for single genes. This result supports the hypothesis that the URF are coding sequences. In fact, it has been recently shown that some URF of human mtDNA encode components of the respiratory chain (Chomyn et al. 1985).

K-entropy Analysis

Sequence analysis based on the parameters of Shannon's information theory is greatly limited when dealing with short sequences ($\sim 10^2$ symbols), as occurs in many biological cases. In fact, in the case of short sequences, one cannot assign the "probabilities," p_i, to the observed frequencies, f_i, nor take as "conditional probabilities," p_{ij}, the observed doublet frequencies, f_{ij}.

A more recent approach to the problem of defining randomness (Chaitin 1966; Kolmogorov 1968), makes it possible to overcome these difficulties, because for short sequences, it is possible to assign a quantitatively defined "degree of randomness" and consequently a value of an entropy-like function as a measure of the randomness itself.

Following Kolmogorov and Chaitin: "A sequence of symbols $p = (A_{i1}, A_{i2}, \ldots, A_{in})$, where $A_{ik} \in (A_1, A_2, \ldots, A_n)$ is called 'random', if there does not exist a shorter program $q = (A_{j1}, A_{j2}, \ldots, A_{j\nu})$ (i.e. $\nu < n$), which uses the same alphabet, and is able to reconstruct the original sequence." This means that for a sequence of "true" random numbers, e.g., written in binary units, there does not exist a program, with a lower number of bits, able to reconstruct the given sequence.

On the other hand, a sequence with some "regularities" has a shorter representation than itself.

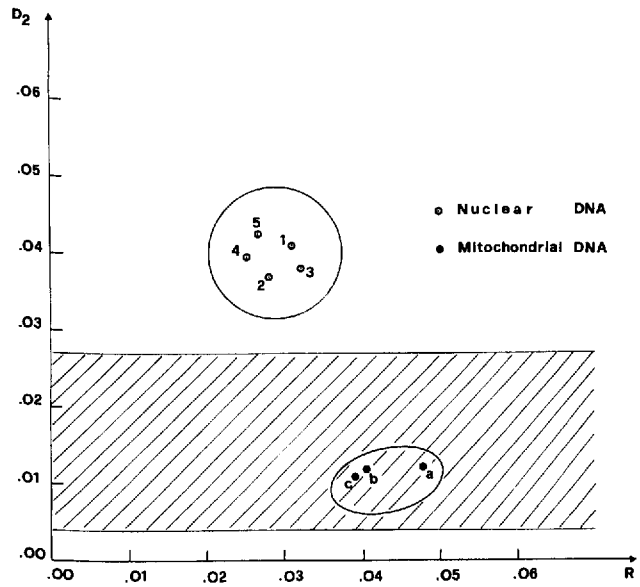

Fig. 1. The points representing nuclear and mitochondrial DNA are grouped in two clearly separated clusters. The dashed region contains the values of D_2 for bacteria.

	D_2	R
Mitochondrial DNA		
a) *Mus musculus*	0.0120	0.0482
b) *Homo sapiens*	0.0122	0.0406
c) *Bos taurus*	0.0110	0.0388
Nuclear DNA[a]		
1) Mouse liver	0.0409	0.0309
2) Mouse thymus	0.0367	0.0281
3) Human spleen	0.0380	0.0318
4) Bovine liver	0.0394	0.0251
5) Bovine thymus	0.0424	0.0266
Bacterial DNA[a]		
1) *Micrococcus lysodeikticus*	0.0132	0.0709
2) *Mycobacterium phlei*	0.0273	0.0582
3) *Haemophilus influenzae*	0.0225	0.0315
4) *Bacillus subtilis*	0.0200	0.0149
5) *Aeromonas aerogenes*	0.0196	0.0145
6) *Escherichia coli* B_b	0.0200	0.0109
7) *Escherichia coli* B_a	0.0159	0.0080

[a] Data from Gatlin (1972)

Table 1. Informational parameters of mitochondrial genes

Gene	D_2	R
Cytochrome oxidase subunit I		
Homo sapiens	0.0345	0.0383
Mus musculus	0.0629	0.0594
Bos taurus	0.0547	0.0524
Cytochrome oxidase subunit II		
Homo sapiens	0.0541	0.0608
Mus musculus	0.1122	0.1100
Bos taurus	0.1079	0.1040
Cytochrome oxidase subunit III		
Homo sapiens	0.0653	0.0650
Mus musculus	0.0707	0.0746
Bos taurus	0.0714	0.0673
Cytochrome B		
Homo sapiens	0.0799	0.0925
Mus musculus	0.0629	0.0811
Bos taurus	0.0783	0.0830
ATPase 6		
Homo sapiens	0.0290	0.0690
Mus musculus	0.0349	0.0778
Bos taurus	0.0354	0.0682

On the basis of the intuitive idea that the absence of regularity is a symptom of randomness, Kolmogorov gave the following quantitative definition of "entropy of a sequence of symbols": "The entropy K(p) of a sequence of symbols $p = (A_{i1}, A_{i2}, \ldots, A_{i\mu})$ is the minimum length of the computer program capable of generating p." Kolmogorov also demonstrated that K(p) exists, but that a "rule" for finding K(p) does not exist.

The concept of K-entropy has been recently applied to the analysis of molecular sequences (Ebeling and Jimenez-Montaño 1980). In the present work we use the parameters defined in the paper of Ebeling and Jimenez-Montaño, to which the reader is referred for more details.

Given a sequence of μ symbols (a "word"):

$$p = (A_{i1}, A_{i2}, \ldots, A_{i\mu})$$

the "complexity of the production rule" is defined as:

$$K(\sigma \rightarrow q) = l(q)$$

where l(q) is the length of the rule, and the "entropy" of the word is defined as:

$$K(p) = \sum_{\sigma} K(\sigma \rightarrow q) + l_f$$

where l_f is the length of the final word. A measure of the "randomness" of the word is the ratio $K(p)/K_{max}$, where K_{max} is the initial length, μ, of the word itself.

To clarify, let us give a simple example: given the sequence of binary symbols p = (1 0 1 0 1 1 0 1 0 1 1 1 0 1 0) ($\mu = K_{max} = 15$), we pose $\sigma_1 \rightarrow 1\,0$, and we obtain the sequence $p_1 = \sigma_1\,\sigma_1\,1\,\sigma_1\,\sigma_1\,1\,1\,\sigma_1\,\sigma_1$ with $K(\sigma_1 \rightarrow 1\,0) = 2$. Going further, we pose $\sigma_2 \rightarrow \sigma_1\,\sigma_1$, [$K(\sigma_2 \rightarrow \sigma_1\,\sigma_1) = 2$], and we obtain $p_2 = \sigma_2\,1\,\sigma_2\,1\,1\,\sigma_2$. At this point the procedure ends, and we have $K(p) = (2 + 2) + 6 = 10$ and $K(p)/K_{max} \cong 0.67$.

Obviously, this procedure is not unique. If we take the following production rules: $\xi_1 \rightarrow 1\,0\,1$, and, successively, $\xi_2 \rightarrow \xi_1\,0$, we obtain, as final sequence: $\xi_2\,1\,\xi_2\,1\,1\,\xi_2$, with K(p) = 11, and $K(p)/K_{max} \cong 0.73$.

Table 2. Informational parameters of URF sequences

Gene	D_2	R
URF 1		
Homo sapiens	0.0260	0.0628
Mus musculus	0.0252	0.0566
Bos taurus	0.0732	0.0821
URF 2		
Homo sapiens	0.0618	0.0965
Mus musculus	0.0546	0.1081
Bos taurus	0.0260	0.0818
URF 3		
Homo sapiens	0.1712	0.1592
Mus musculus	0.1535	0.1569
Bos taurus	0.1881	0.1601
URF 4		
Homo sapiens	0.0329	0.0805
Mus musculus	0.0418	0.0829
Bos taurus	0.0617	0.0806
URF 4L		
Homo sapiens	0.1981	0.1611
Mus musculus	0.1686	0.1551
Bos taurus	0.2311	0.1930
URF 5		
Homo sapiens	0.0462	0.0806
Mus musculus	0.0434	0.0820
Bos taurus	0.0551	0.0838
URF 6		
Homo sapiens	0.1363	0.1824
Mus musculus	0.1322	0.2006
Bos taurus	0.1517	0.1821
URF A6L		
Homo sapiens	0.1833	0.2207
Mus musculus	0.1731	0.2170
Bos taurus	0.2360	0.2328

One can, of course, try with other production rules, for example with $\eta_1 \rightarrow 1\,0\,1\,0$, obtaining a K(p) = 10, or with $\eta_2 \rightarrow 1\,0\,1\,0\,1$, a K(p) = 12. At this point, we can say that the *true* value of K(p) is "probably" the lowest value we found, namely K(p) = 10, and that the relative randomness of our word p is $10/15 \cong 0.67$.

Because the final value of K(p) depends on the production rules used, it is clear that it is possible to use a great number of different algorithms, and then choose the best value (i.e., the minimum value) of K(p) only with a fast computer. We can only say that the *true* value $K_t(p)$ is less than or equal to the lowest value $K_f(p)$ found: obviously the probability that $K_t(p) = K_f(p)$ increases as the number of trials increases.

With this kind of procedure we examined three relatively short (medium length, 240 bases) nucleotide sequences, i.e.: α) subunit 9 of nuclear ATPase gene from the ascomycete *Neurospora crassa*, β) subunit 9 of mitochondrial ATPase gene from *Neurospora crassa*, and γ) subunit 9 of mitochondrial ATPase gene from the yeast *Saccharomyces cerevisiae* (data from van den Boogart et al. 1982).

These three sequences are genes coding for the same protein: our interest was focused on the evaluation of $K(p)/K_{max}$ for these genes in order to see if the two mitochondrial genes from different organisms were more "similar" to each other, in the sense of linguistic complexity, than two genes from the same organism (*Neurospora crassa*), but on two different genomes (nuclear and mitochondrial).

With the aid of the computer, we used an algorithm able to discriminate, step by step, a monotonically decreasing value of $K(p)/K_{max}$, starting from the initial value 1. When the value of $K(p)/K_{max}$ did not decrease further, the computer stopped. The values found for $K(p)/K_{max}$ were *Neurospora crassa*, mitochondrial = 0.62; *Neurospora crassa*, nuclear = 0.56; and *Saccharomyces cerevisiae*, mitochondrial = 0.60. We recall that all these genes probably code for the same protein (subunit 9 of ATPase).

Our linguistic analysis shows that there is a good similarity (in the sense of K-entropy, of course) between the two mitochondrial genes. These results, although obtained in the case of three genes only, are in agreement with the ones derived with the aid of Shannon entropy. When, in the future, many similar data (i.e., sequences of mitochondrial and nuclear origin coding for the same sequences) will be available, it will be possible to cluster the data and gain more meaningful insight on the problem of the origin of mitochondria.

Moreover, the number of steps by which the minimum value of K(p) is reached by the computer is exactly the same for the two mitochondrial genes (14 steps), and very different from the number of steps necessary for the nuclear one (18 steps). These results are also in agreement with the test of homology of the amino acid sequences of the proteins encoded by these genes (van den Boogart et al. 1982).

Conclusion

The linguistic complexity analysis of polynucleotide sequences, with the aid of the parameters of Shannon entropy, seems to indicate that the language of the mitochondrial genetic message is more similar to the language of the bacterial genome than to the language of the nucleus in more evolutionarily advanced animals. Moreover, within the same organism (*Neurospora crassa*), the two genes coding for the same specific protein, one of mitochondrial origin and the other of nuclear origin, are very different in the sense of Kolmogorov entropy, whereas the gene of mitochondrial origin is very similar to the

one, again of mitochondrial origin, belonging to another organism (*Saccharomyces cerevisiae*). These results appear to support the theory of the exogenous symbiotic origin of mitochondria.

Acknowledgments. We thank Prof. C. Saccone for the DNA sequences supplied from Banca Dati Sequenze Acidi Nucleici in Bari.

References

Anderson S, Bankier AT, Barrel BG, de Bruijn MHL, Coulson AR, Drovin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–464

Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG (1982) Complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683–717

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA (1981) Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167–180

Chaitin G (1966) On the length of programs for computing finite binary sequences. J Assoc Comput Mach 13:547–569

Chomyn A, Mariottini P, Cleeter MWJ, Ragan CI, Matsuno-Yagi A, Hatefi Y, Doolittle RF, Attardi G (1985) Six unidentified reading frames of human mitochondrial DNA encode components of the respiratory-chain NADH dehydrogenase. Nature 314:592–597

Ebeling W, Jimenez-Montaño MA (1980) On grammar, complexity and information measures of biological macromolecules. Math Biosci 52:53–71

Gatlin LL (1968) The information content of DNA. II. J Theor Biol 18:181–194

Gatlin LL (1972) Information theory and the living system. Columbia University Press, New York

Granero-Porati MI, Porati A, Zani L (1980) Informational parameters of an exact DNA base sequence. J. Theor Biol 86: 401–403

Kolmogorov A (1968) Logical basis for information theory and probability theory. IEEE Trans Information Theory IT-14: 662–664

Lipman DJ, Maizel J (1982) Comparative analysis of nucleic acid sequences by their general constraints. Nucleic Acids Res 10:2723–2739

Rowe GW, Trainor LEH (1983) On the informational content of viral DNA. J Theor Biol 107:151–170

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchinson III CA, Slocombe PM, Smith M (1977) Nucleotidic sequence of bacteriophage ΦX174 DNA. Nature 265: 687–695

Subba Rao J, Geevan CP, Subba Rao G (1982) Significance of the information content of DNA in mutations and evolution. J Theor Biol 96:571–577

van den Boogart P, Samallo J, Agsteribbe E (1982) Similar genes for mitochondrial ATPase subunit in the nuclear and mitochondrial genomes of *Neurospora crassa*. Nature 298: 187–189

Received June 3, 1986/Revised July 27, 1987/Accepted August 19, 1987