AN EVOLUTIONARY MODEL OF NUCLEIC ACID SEQUENCE
--The Model of Inflation-Mutation-Selection--

Liaofu Luo *
Department of Physics, University of
Toronto, Toronto M5S]A7, Canada

Received 13 December 1990, 5 April 1991

ABSTRACT: The statistical characteristics of the coding part of nucleic acid sequences are reviewed. They are:], a short range of statistical correlation of bases; 2, a dependence of informational parameters on evolution; 3, specific features of subsequence parameters; 4, the frequency of codon usage with respect to evolution; 5, a frozen ratio of repeated fragments; 6, a biased GC content in repeated fragments; 7, the independence of length on evolution. To explain these characteristics it is supposed that in the early epoch of molecular evolution there exists an inflation stage. That is, in a relatively short time the sequence is enlarged through repeating and linkage of small fragments. Only after a definite length is reached could a new stage of evolution begin, this being dominated by base mutation and selection. Single base mutation is stochastic and random while natural selection on the molecular level makes the genetic language exhibit higher reliability and a greater ability to resist disturbance. The joint action of the two explains the dependence of informational parameters on evolution.

* * *

]. AN OUTLINE OF RESULTS FROM THE STATISTICAL ANALYSIS OF NUCLEOTIDE SEQUENCE DATA

A great deal of information exists in nucleotide sequences. Despite the explosive increase in sequence data, people know little about their theoretical implications. It is a 'heaven book' indeed. Recently we have analyzed nucleotide sequences (coding regions, for the most part) statistically in order to find the rules behind them which dominate molecular evolution. The main results are as follows.

]) The statistical correlation of bases for most coding regions of nucleotide sequences is only for a short distance. Rowe and Trainor (]983a) demonstrated that for viral DNA only strong triplet correlation exists. We define the correlation length to be such that for distances longer than s between bases the base correlations have a random structure . For 234] coding regions we discover that 6.7% of them have s = 0, 59% have] ≤ s ≤ 2, 9.7% have 3-periodical long range correlations and 24% have nonperiodical long range correlations. Furthermore it can be shown that the s = 0 component (random sequences) changes with evolution, from 30% to 20& for phages, chloroplasts and mitochondria,]0% for bacteria and viruses to about]% for vertebrates.(For details see Luo and Li]99])

* * *

* Permanent address Physics Department, Inner Mongolia University, Huhehote, China.

Evolutionary Theory 10: 75-81 (December, 1991)
The editors thank two referees for help in evaluating this paper.

2) The statistical dependence of informational parameters (IPs) on molecular evolution. Since the statistical correlation for most sequences is of short range, to give an approximate description one may use a few IPs, namely, the first-order informational redundancy, D_1, describing the divergence from equiprobability of base composition and the second-order informational redundancy, D_2, describing the divergence from independence of neighbouring bases or their combinations. In fact Gatlin first tried to investigate nucleotide sequences using D_1 and D_2 (Gatlin 1972). However, only a few data were published at that time, so her results are far from conclusive. Recently we have calculated D_1, D_2, $X=D_2/(D_1+D_2)$, and F (G+C content) for each sequence and averaged them for each category. We discovered meaningful statistical correlations between these parameters and molecular evolution, namely, $\langle D_2 \rangle$ varying from category to category, i.e., from 0.03 for mitochondria, 0.05-0.06 for phages and bacteria, to 0.11 for mammals; and $\langle x \rangle$ varying from 0.2 to 0.8 correspondingly. The latter is a comparatively good statistical quantity with a smaller deviation within each category. $\langle F \rangle$ also changes with evolution, from 0.28 for mitochondria to 0.53 for mammals on the average (Luo et al. 1988).

3) Specific feature of subsequence parameters. According to the position of a base in the codon we break the data of coding regions into three constituent subsequences and calculate their IPs. For example, for subsequence 1, which comprises all bases that occur in the first codon position, $D_1^{(1)}$ describes the base composition of the first codon position in the sequence, $D_2^{(1)}$ describes the correlation of a base in the first codon position with its nearest neighbouring base to the left in the sequence, and $X^{(1)}=D_2^{(1)}/(D_1^{(1)}+D_2^{(1)})$, etc. We find that only for the first sequence does there exist a good statistical correlation between $\langle D_2^{(1)} \rangle$ or $\langle X^{(1)} \rangle$ with molecular evolution. On the other hand, $D_1^{(3)}$ of the third subsequence is larger than $D_1^{(1)}$ and $D_1^{(2)}$ (by a factor of 1.5-2.5 for most sequences), which is related to the nonrandom usage of synonymous codons. In connection with this, $\langle F^{(3)} \rangle$ changes with evolution remarkably(Luo and Shengli 1990).

4) The frequency of codon usage with respect to evolution. From the statistics of 10^6 codons in GenBank we obtained the frequencies of codon usage for each codon multiplet and each category of species. The usage of synonymous codons is far from random and exhibits a definite relation with evolution. An important rule is that the C/U ratio and G/A ratio of the third bases in a degenerate multiplet change with evolution. For example, the C/U ratio in Phe, Cys, Tyr, Asp, His and Asn changes from 0.4/0.6(i.e., 0.7) for chloroplasts and phages to 0.6/0.4(i.e., 1.5) for eukaryotes. In degenerate quartlets a similar relation holds too(Hu and Luo, 1989). The diversity in GC content at the third position of codons in vertebrates had been reported in earlier work(Bernardi and Bernardi 1986, Aota and Ikemura 1986).

5) Frozen ratio of repeated segments. We have investigated the repeated segments in coding regions with length $N<4000$. To minimize stochastic effects only repeated segments with length $l \geq 6$ were taken into account. Evidently, for a sequence of $N<4^6=4096$ the repitition of a segment with $l \geq 6$ could not be due to random recurrence. Define the reducing rate $R=(N-L)/N$, where L is the length of the reduced sequence, in which each repeated segment has been replaced by a single letter. Define the repeating rate $S=K/N$, where K is the total number of repeated segments in the sequence. The definitions are the same as Ebeling's(Ebeling and Jimenez-Montano, 1980). By calculation of 257 sequences

of different species we obtain the averaged value of R and S

$$\langle R\rangle \sim 0.20\text{-}0.24 \qquad \langle S\rangle \sim 0.084\text{-}0.090 \tag{1}$$

which takes nearly the same value for different categories of species and thus shows independence of evolution. We call Equation (1) the frozen ratio of repeated segments(Zhou and Luo, 1990).

6) The structure of repeated segment - large D_1' and biased GC content. Define the first-order informational redundancy D_1' and GC content F' of an artificial sequence which is obtained by joining all repeated segments. We discover that for most sequences, D_1' in repeated segments is larger than D_1 in the corresponding real sequence by a factor 2 to 3. The total length of repeated segments for a coding sequence is about half N, as seen from (1). Of course, the fluctuation due to short length would make D_1' larger. However, detailed analysis shows that it could not explain such a large D_1' as 2 to 3 times D_1. In fact the largeness of D_1' is closely related to the biased GC content in repeated segments. One finds that if $F>0.5$ then the corresponding $F'>F$ and if $F<0.5$ then $F'<F$, the change of which (namely, $|F'-F|/F$) is several percent to 10 percent or more(Zhou and Luo, 1991).

7) Independence of length on evolution. The lengths of most coding sequences are above 1000(except perhaps mitochondria, which have comparatively small values of length). Furthermore we could not find any meaningful statistical association between length and evolutionary level.

2. THE INFLATION-MUTATION-SELECTION MODEL OF SEQUENCE EVOLUTION

The historical formation and stabilization of a nucleic acid sequence is a very complicated process. There are too many accidental factors. So a reasonable method must be a statistical treatment which considers a large number of sequences, divides them into categories and finds the statistical relationships between them. Following this approach we have found a series of statistical characteristics of coding sequences which are summarized in section 1. Evidently these statistical characteristics have deep theoretical implications which are especially important for understanding of the evolution of nucleotide sequence. In order to explain 1), 2), and 4) we have proposed a simplified model of molecular evolution(Luo et al. 1990). The basic assumptions are

1) Consider an ensemble of sequences(strings). Each string consists of two letters A and B. The base A may undergo a mutation to base B with a specific probability in one time step and base B to base A with the same probability. The assumption represents a muchanism of neutral mutation and random drift in molecular evolution.

2) Each pair of bases is called a codon, which is the unit of selection. There are four types of codons, namyly AA, AB, BA and BB, the numbers of which in one string are denoted by n_1, n_2, n_3 and n_4 respectively. Because of a selection rule the fitnesses of four codons are different from each other. For example, the selection rule may be specified as follows. If the mutations in one step lead to

$$\delta S \geq 1 \quad \text{(2a)} \qquad \text{or} \qquad \delta S \leq -1 \quad \text{(2b)}$$

for a particular string then the string will reproduce a new identical one (case 2a) or be eliminated from the ensemble (case 2b) with a probability β. Selection rule (2a) corresponds to selective advantage or advantageous mutations and rule (2b) to selective elimination or lethal mutations. Let the fitness function

$$S = n_2 - n_3 - \tfrac{1}{2}n_4 \tag{3}$$

From the above-mentioned assumptions we deduce a set of evolution equations for the probability p_i ($i=1,\ldots,4$) of the ith codon in the ensemble. Then we obtain the solution of the codon probability p_i

and the evolution of entropy, Markovian entropy and IPs. For example, when time is large enough we have

$$p_1 = \frac{1}{4} + \frac{\beta}{32} \qquad p_2 = \frac{1}{4} + \frac{7\beta}{32}$$
$$p_3 = \frac{1}{4} - \frac{5\beta}{32} \qquad p_4 = \frac{1}{4} - \frac{3\beta}{32} \tag{4}$$

Simultaneously we obtain the time evolution of D_1 and D_2. D_1 decreases monotonically and approaches $\beta^2/128$ and, if the initial string is supposed to be an independent sequence, D_2 will increase from 0 to $\beta^2/512$. The former is due to entropy production from random mutation. The latter is due to selection which makes codons AB occur more frequently than others and leads to the deviation from independence of neighbouring bases. As a result, $X(X=D_2/D_1+D_2)$ increases in the course of evolution. Of course, if a different fitness function S is used instead of (3), then one obtains another selective direction corresponding to the new S. However, the general picture of the variation of D_1, D_2 and X will remain unchanged. In fact, the single base mutation under selection causes the entropy growth - a wider distribution of base frequency or vocabulary constitution of genetic language, on one side, and increases the correlation between neighbouring bases - the clarification of grammatical rules, on the other side. The two factors are united in one parameter X, which describes the reliability of the genetic language and the ability to resist disturbance. Generally speaking, base correlation can be divided into short-range and long-range. In spite of the high level of organization of long-range correlation, single-base mutation could change only the short-range correlation effectively. On the other hand selection pressure acts in the dimension of one codon. So the evolution of the genetic language is manifested mainly in the growth of shor-range correlation. The point has been exemplified in the above-mentioned simple model.

If the codons BA and BB are supposed to belong to a degenerate multiplet, then BB being more fit than BA according to selection rule (3) implies a selection pressure which leads to non-random usage of synnonymous codons, namely the growth of the ratio BB/BA in the course of evolution(see Equation (4)). Consequently the model could explain the statistical characteristic 4 as well. In fact, the nonrandom usage of codons and its relation to evolution is a direct evidence of the codon being a unit of selection.

For real nucleic acids the sequence is written in four letters and the codon is composed of three bases. The generalization of our model to the real case is straightforward. Of course, the key point is to formulate a reasonable selection rule. From the viewpoint of molecular biology the function of a protein is determined by its conformation. The secondary structure of protein is mainly determined by dipeptide correlation(Luo and Dong 1988) in addition to single-peptide frequency (Chou and Fasman 1978). So one may assume a selection pressure acting on one codon as well as two neighbouring codons. The latter results in increasing the statistical correlation of bases of adjacent codons, which would be able to explain the statistical characteristic 3 of nucleic acid sequence because $D_2^{(1)}$ and $X^{(1)}$ describe the correlation of neighbouring codons exactly.

The selection pressure reflects the influence of the requirement of function on the structure of sequence. Different environmental conditions lead to different selection rules and different selection rules lead to different evolutionary trajectories of sequences. So in this model different environmental conditions lead to a diversity of species from selection and they develop in different evolutionary tra-

jectories and rates and arrive at different levels at the present stage.

So far we have succeeded in explaining the statistical characteristics of coding sequences by use of the proposed model. But it is difficult to understand the characteristics of repeated fragments. Since the observed repeating is not due to statistical recurrence and its biased GC content implies that these fragments could hardly be formed in evolution we suppose that, these repeated fragments are remnants of prebiotic evolution of the sequence statistically. There may have existed a prebiotic epoch in which the sequence was enlarged rapidly through repeating and linkage of bases and small fragments. The process could be understood from thermodynamic considerations as Rowe and Trainor had pointed out(Rowe and Trainor]983b). The primitive DNA fragment first formed in this way is strongly codon-biased in general. However, the codon content depends on energy, which in turn depends sensitively on the environment. To ensure the superiority of a definite bias in a long segment one should assume a homogeneous condition for the sequence growth, which requires the process to proceed rapidly and sharply. In fact,'life had to arise very quickly', 'all the prebiotic processes we know about are fast'(Waldrop]990). Perhaps our discussion may provide indirect evidence on this point. As for why these repeated fragments, resisting molecular noise, have remained a frozen ratio in the long years of evolution, it is a difficult problem. For the repetitious or non-specific DNA in noncoding regions, it has been suggested that it 'exists for no other purpose than its own perpetuation'(Grant]98]). Of course, the reason for repeating in coding regions may be different from noncoding regions. A possible answer is as follows. The repeated fragments with length $\geqslant 6$ in DNA correspond to the repeating of 2 or 3 residues in the amino acid sequence. Since the dipeptide(tripeptide) correlation is a main factor in determining the secondary structure of protein, so the observed repetition in sequence may be necessary for sustaining some secondary structures even though the fragment itself may not participate in any active site of the protein. In the prebiotic epoch, free from the constraint of function, the mutation proceeds randomly and frequently. Accompanying the enlargement of sequence the frequent mutations break a lot of repeated fragments and make R and S descend gradually. After a comparatively short time the length N of sequence and the amount of information possibly stored (roughly proportional to log N) have mostly accumulated. In the meantime, the reducing rate R and repeating rate S have decreased by a large amount. As a definite length of sequence and a definite value of R and S is reached, then primitive life occurs and a new stage of evolution begins. We call the above-mentioned prebiotic epoch the inflation stage of sequence evolution. Evidently, if the inflation of sequences proceeds rapidly and ends sharply enough then it is not difficult to understand why the 'frozen' phenomenon of R and S has happened.

We have proposed a theoretical model to explain the observed statistical characteristics in coding regions of nucleic acid sequences. Of course the evolution of coding regions may be closely related to the evolution of different kinds of non-coding regions which play an important role in regulating gene expression. The latter is beyond the scope of this article. In this note we only analyse the coding region as an independent entity and neglect its connection with other genetic elements. We expect that a more satisfactory evolutionary model of sequences will be established on the bases of complete investigation of the whole genome along the same line as this article.

Finally, we point out that according to modern cosmology, to over-

come a series of difficulties such as why the monopole has not been found, why the Universe is so homogeneous, why the mass of the Universe is so near its critical value, etc., an inflation theory has been proposed which suggests an inflation stage in the very early epoch of the Universe's evolution. Likewise, to explain the statistical characteristics of nucleic acids it seems desirable to assume an inflation stage in the early evolution of sequence, too. The great similarity between life and the Universe reminds us of the idea of ancient Chinese philosophers: 'the unity of Universe and Man'

ACKOWLEDGEMENT The author would like to thank his students and colleagues Y.M. Zhou, L. Tsai, Y. Z. Dong, W. J. Lee, H. Li, Shengli and S. Z. Hu of Inner Mongolia University, who have done a great deal of statistical studies on nucleic acid sequences. The work is completed under the Canada-China scholarly exchange program. The hospitality of Profs. M. B. Walker and L. E. H. Trainor of the University of Toronto is gratefully acknowledged. The author is specially indebted to Prof. Trainor for his enlightening discussion of collaborative work with Dr. G. W.Rowe, which gave great impetus to the present investigation.

REFERENCES

Aota, S., and Ikemura, T. (1986) Diversity in G+C content at 3rd position of condons in vertebrate genes and its cause. Nucleic Acid Res. 14: 6345 -55.

Bernard, G., and Bernard,G.(1986) Compositional constraints and genome evolution. J. Mol. Evol. 24:1 -11.

Chou, P.Y., and Fasman, G.D. (1988) Emprical predictions of protein conformations. Ann. Rev. Biochem. 47:251 -276.

Ebeling, W., and Jimenez-Montano, M.A. (1980) On grammars, complexity and information measures of biological macromolecules. Math. Biosci. 52: 53 -71.

Gatlin, L. (1972) Information Theory and the Living System. Columbia Univ. Press.

Grant, B. (1981) The safe-neighbored hypothesis of junk DNA. J. Theor. Biol. 90:149 -50.

Hu, S.Z., and Luo, L.F. (1989) The frequencies of codon usage and its relation to evolution. Acta Sci. Nat. Univ. Intramongolicae 20:483 -90.

Luo, L.F., and Dong, Y.Z. (1988) Statistical analysis of peptide correlation and prediction of protein conformation. Chinese Biochem. J. 4:174 -83.

Luo, L.F., Tsai, L., and Zhou, Y.M. (1988) Informational parameters of nucleic acid and molecular evolution. J. Theor. Biol. 130:351 -61.

Luo, L.F., Tsai, L., and Lee,W.J. (1990) Model of evolution of molecular sequences. Phys. Rev. A41:5441-50.

Luo, L.F., and Li, H. (1991) The statistical correlation of nucleotides in protein-coding DNA sequences. Bull. Math. Biol. 53.

Luo, L.f., and Shengli (1990) The information-theoretic investigation of heterogeneous DNA sequences. Acta Sci. Nat. Univ. Intramongolicae 21:229 -34.

Rowe, G.W., and Trainor, L.E.H. (1983a) On the informational content of viral DNA. J.Theor.Biol. 101:151 -70.

Rowe, G.W., and Trainor, L.E.H. (1983b) A thermodynamic theory of codon bias in viral genes. J. Theor.Biol. 101:171 -203.

Waldrop, M. (1990) Goodbye to the warm little pond? Science 250:1078-80.

Zhou, Y.M., and Luo, L.F. (1990) Statistical analysis of repeated segments in nucleic acid sequences. Acta Sci. Nat. Univ. Intramongolicae 21:83-89.

Zhou, Y.M., and Luo, L.F. (1991) Repeated segments in coding regions of nucleotide sequences. Acta Sci. Nat. Univ. Intramongolicae. 22.