Eur. Phys. J. B **26**, 393–398 (2002)
DOI: 10.1140/epjb/e20020105

THE EUROPEAN PHYSICAL JOURNAL B
EDP Sciences

DNA nucleotides: A case study of evolution

S. Chattopadhyay[a], W.A. Kanner, and J. Chakrabarti

Department of Theoretical Physics, Indian Association for the Cultivation of Science, Calcutta 700 032, India

Received 3 January 2002

Abstract. The evolution in coding DNA sequences brings new flexibility and freedom to the codon words, even as the underlying nucleotides get significantly ordered. These curious contra-rules of gene organisation are observed from the distribution of words and the second moments of the nucleotide letters. We apply these statistical data to determine the relative positions of a few bacterial groups as per their divergence in the geological timescale.

PACS. 82.35.Pq Biopolymers, biopolymerization – 82.39.Fk Enzyme kinetics – 87.14.-g Biomolecules: types – 02.50.Ey Stochastic processes

Over the years the statistical approach to genes has become prominent. The hidden Markov models are used in the alignment routines of biological sequences. For the secondary structures of the sequences stochastic context-free and context-sensitive grammars are applied [1]. The recent discovery of the fractal inverse power-law correlations [2] in these biological chains have led to ideas that statistically these sequences have features of music and languages [3–5]. The purpose of this work is to track the statistical basis of the evolution in the coding DNA sequences (CDS).

The CDS of any gene does not have all the salient features that accompany change. The genes that are present in the whole range of organisms, from the lowest bacteria to the highest mammals, and therefore connected to fundamental life processes are normally considered to be best suited as evolutionary markers. With this in view we choose glyceraldehyde-3-phosphate dehydrogenase (GAPDH) CDS for its ubiquitous presence in all living beings. The enzyme it codes for catalyses one of the crucial energy-producing steps of glycolysis, the common pathway for both aerobic and anaerobic respiration.

Distribution of words is studied for languages. The frequency of words is plotted against the rank. Here the total number of occurrences of a particular word is termed its frequency. The word most frequent has rank = 1, the next most has rank = 2, and so on. For natural languages, the plot gives the Zipf [4] behaviour:

$$f_N = \frac{f_1}{N} \tag{1}$$

where N stands for the rank and f_1 and f_N are the frequencies of words of rank 1 and N respectively. The Zipf-type approach to the study of DNA has brought methods of statistical linguistics into DNA analysis [4]. The generalized Zipf distribution of n-tuples has provided hints that the DNA sequences may have some structural features common to languages. In this work we confine ourselves to the distribution of 3-tuples, the codons, in the CDS. The words, therefore, are non-overlapping and on the single reading frame.

The frequency-*vs.*-rank plot of the codon words show that these distributions, given the frequency of rank 1 and the length of the sequence, are almost completely defined through the universal exponential functional form [6]:

$$f_n = f_1 \mathrm{e}^{-\beta(N-1)}. \tag{2}$$

The parameter, called β, is determined by the ratio

$$\beta \approx \frac{f_1}{L} \tag{3}$$

β measures the frequency of rank 1 per unit length of the sequence. The exponential form (2) is to be compared to the usual Boltzmann distribution. The rank of the word is akin to energy; β is analogous to inverse temperature. The relationship (3) that β is frequency of rank 1 per unit length is supported well from data [6]. The analogy between word distributions and the classical Boltzmann concepts goes deeper. A decrease in β, from (3), implies frequency of rank 1 per unit length goes down. In that case the vocabulary clearly increases. More words are used, thereby more states are accessed. For the GAPDH CDS we find the evolution is driving it to higher temperatures; into more freedom for words, into more randomisation. β evolves monotonically.

At this point it is important to refer to the work of Som *et al.* [6] where it was shown that β increases with evolution for the genes that code for α-globin, β-globin, insulin and globulin. Though we believe the value

[a] e-mail: tpsc@mahendra.iacs.res.in

$$X = \frac{\textit{Second Moment of the Base Distribution in GAPDH CDS}}{\textit{Second Moment of the Base Distribution in the random sequence with identical strand bias}}$$

of β has evolutionary content, there are doubts regarding the potentiality of these genes as evolutionary markers. On the contrary we missed the hints shown by the gene for phosphoglycerate kinase (PGK) another glycolytic enzyme like GAPDH and therefore harboured by every living organism. The PGK and the GAPDH, being two extremely significant enzymes connected to one of the most fundamental metabolic processes, are well established as strong phylogenetic chronometers. Interestingly, in [6], the PGK showed a trend exactly similar to that shown by the GAPDH in the present study.

Underneath, however, there runs a curious counterflow. Suppose we look into the nucleotides that constitute the sequence, once again in windows of size 3 and in the same reading frame. First, we ask how much order there is in the sequence. To find out we study the second moments of the letters A, C, G and T. These second moments, by themselves, do not produce any pattern. The GAPDH CDS has about 1000 bases. For each organism the proportions of A, C, G and T in the GAPDH CDS are different. This strand-bias, interestingly, masks a remarkable underlying trend.

To get there the strand-bias has to be eliminated. The order in the sequence, we assume, is its deviation from the random. We define the quantity X, a measure of this deviation, as follows:

See equation above

X is thus strand normalised. X values of GAPDH change monotonically with evolution. The data tells us there is an increase in strand normalised persistence amongst the letters (in windows of size 3) with evolution in the CDS.

The evolution in the GAPDH CDS is then the result of these two contra trends: while words acquire greater uniformisation, the underlying letters have more order. The monotonic behaviours of β and X with evolution offer us insights on the relative periods of divergence of a few bacterial groups.

Methods

Word distributions

For the codons it is known [6] the exponentials give somewhat better fits over the usual power laws. The exponential form, equation (2), is characterized by the parameter β. The quantity has some universal features in that it is almost completely determined by f_1 and the length of the CDS. The relationship [6]

$$\beta \;=\; \frac{f_1 - 1}{L} \;+\; \frac{1}{2}\frac{(f_1-1)^2}{L^2} \tag{4}$$

is known to fit observations on diverse genes. For the bacterial GAPDH CDS the results of β are given in Table 2.

Moments

Consider the 4-dimensional walk model [7] such that A, C, G and T correspond to unit steps, in the positive direction, along $X_{\rm A}$, $X_{\rm C}$, $X_{\rm G}$ and $X_{\rm T}$ axes. After n-steps if the coordinate of the walker is ($n_{\rm A}$, $n_{\rm C}$, $n_{\rm G}$, $n_{\rm T}$), then, clearly,

$$n \;=\; n_{\rm A} + n_{\rm C} + n_{\rm G} + n_{\rm T} \tag{5}$$

and n_i ($i \equiv A$, C, G, T), is the number of nucleotide of type i in the sequence just walked.

If the sequence has n bases, and n_i is the number of base of type i, the strand bias of the sequence is the proportion of n_i in n, defined as

$$p_i \;=\; \frac{n_i}{n}\,. \tag{6}$$

The probability distribution for the single step in this 4-d walk is

$$P_1(x) \;=\; \sum_i p_i \delta(x_i - 1) \prod_{j\neq i} \delta(x_j) \tag{7}$$

where δ is the usual δ-function of Dirac, and x_i with $i =$ A, C, G, T label the axes of the 4-d walk space. The characteristic function of the step is the Fourier transform of equation (7),

$$P_1(k) \;=\; \sum_i p_i {\rm e}^{{\rm i}k_i}. \tag{8}$$

The characteristic function of l steps

$$P_l(k) \;=\; [P_1(k)]^l. \tag{9}$$

Since we are interested in codons we want to know the distribution of letters A, C, G and T after 3 steps. The window size, l, is 3. The moments of the distributions are obtained taking derivatives of $P_l(k)$ with respect to k. Thus for the random sequence (indicated by the subscript r) with the strand bias (6), we get the average values:

$$\left\langle n_i^2 \right\rangle_r \;=\; l[(l-1)p_i^2 + p_i] \tag{10}$$

$$\langle n_i n_j \rangle_r \;=\; l(l-1)(p_i.p_j) \qquad (i \neq j). \tag{11}$$

Readers familiar with multinomial distributions may identify (10) and (11) readily from (6).

We are interested in codons, therefore, the window size l in equations (10, 11) is chosen to be 3. For the actual sequences we calculate $\langle n_i^2 \rangle_{\rm seq}$ and $\langle n_i n_j \rangle_{\rm seq}$. The number of codons in the CDS denotes the sample size. The averages of $\langle n_i^2 \rangle_{\rm seq}$ and $\langle n_i n_j \rangle_{\rm seq}$ are obtained over the sample size. As we plot these averages as the function of the sample size, the standard deviations in $\sum_i \langle n_i^2 \rangle_{\rm seq}$

Table 1. The average β and X values of GAPDH CDS for eukaryotic groups, along with the range of deviations in the respective groups.

Group	β	X
Vertebrates	0.05398 ($\pm$0.00414)	0.99698 ($\pm$0.004)
Invertebrates	0.07503 ($\pm$0.01067)	1.00235 ($\pm$0.00261)
Fungi	0.07742 ($\pm$0.00389)	1.00705 ($\pm$0.00175)

and $\sum_{i,j}\langle n_i n_j\rangle_{\rm seq}$ are determined. The deviations between $\langle n_i^2\rangle_r$ and $\langle n_i^2\rangle_{\rm seq}$ as well as $\langle n_i n_j\rangle_r$ and $\langle n_i n_j\rangle_{\rm seq}$ are found to be way above the standard deviations. The statistical significance is established to the level of 99.9%.

At this point, the quantities

$$X_{\rm D} \;=\; \frac{\langle n_i^2\rangle_{\rm seq}}{\langle n_i^2\rangle_r} \qquad (12)$$

(The symbol D stands for diagonal moments. Thus, $\langle n_i^2\rangle$ stands for average value of $n_{\rm A}^2$ or $n_{\rm C}^2$ or $n_{\rm G}^2$ or $n_{\rm T}^2$ in the window of size l=3.) and

$$X_{\rm OD} \;=\; \frac{\langle n_i n_j\rangle_{\rm seq}}{\langle n_i n_j\rangle_r} \qquad (i\neq j) \qquad (13)$$

(The symbol OD stands for off-diagonal moments. Thus $\langle n_i n_j\rangle$ stands for the average value of $n_{\rm A}n_{\rm C}$ or $n_{\rm A}n_{\rm G}$ or $n_{\rm A}n_T$ or $n_{\rm C}n_{\rm G}$ or $n_{\rm C}n_{\rm T}$ or $n_{\rm G}n_{\rm T}$.) measure the deviation of the diagonal and off-diagonal second moments of the sequence to those of the random sequence of identical strand bias respectively.

It is to be noted that the first moments are simple averages over numbers n_i. Fot the directed walk the quantities $\langle n_i\rangle_{\rm seq}$ are always equal to $\langle n_i\rangle_r$ irrespective of the nucleotide distribution. The difference in the distributions appear beginning with the second moments. Stated another way, for the fluctuations $\langle n^2\rangle - \langle n\rangle^2$ the second term, $\langle n\rangle^2$ is the same for the sequence as well as the random. We have, therefore, considered the second moments alone.

The over-all averaged index, X, is given by

$$X \;=\; \frac{\sum X_{\rm D} \;+\; \sum X_{\rm OD}}{10}\,. \qquad (14)$$

This X along with $X_{\rm D}$ and $X_{\rm OD}$ provide measures of the order in the sequences.

Observations and results

The X values for the GAPDH CDS for a wide variety of organisms are recorded in Tables 1, 2 and 3. The X value, we recall, is a measure of the deviation of the nucleotide organisation between the CDS from that of the random sequence of identical strand bias and length. Notice that these values are close to unity. It is, therefore, important to establish the statistical significance level of these data.

Statistical significance

To test the level of statistical significance of the experimental values compared to the corresponding random ones, we study the fluctuations within $\sum_i\langle n_i^2\rangle_{\rm seq}$ and $\sum_{i,j}\langle n_i n_j\rangle_{\rm seq}$ for a particular CDS. We break up each CDS in different segments in such a way that for the smallest segment the sample size, *i.e.* the number of codons, is at least 300 so that the statistical analysis becomes meaningful.

The length of the GAPDH sequence, L, is about 1008 base pairs (bp) and varies to a small extent from species to species. For the test of significance, we make six segments of 900 bp, 921 bp, 942 bp, 960 bp, 981 bp and L so that each of them becomes multiple of three (the codon size). Since the sample size, N, *i.e.* the number of segments (here $N=6$), is small, we deal with t-distribution that can be used to attach confidence limits to our experimental values in the same way that the normal distribution can be used for large sample size ($N>30$).

From the six $\sum_i\langle n_i^2\rangle_{\rm seq}$ and six $\sum_{i,j}\langle n_i n_j\rangle_{\rm seq}$ values for the six segments of each GAPDH CDS, we get the mean values, *i.e.* $\langle\sum_i\langle n_i^2\rangle_{\rm seq}\rangle$ and $\langle\sum_{i,j}\langle n_i n_j\rangle_{\rm seq}\rangle$ (that we term $avg_{\rm D}$ and $avg_{\rm OD}$ respectively), and also the standard deviations for the CDS, $S_{\rm D}$ and $S_{\rm OD}$ respectively. Any statistically meaningful value for $\sum_i\langle n_i^2\rangle_{\rm seq}$, as per the t-distribution should lie within the following range:

$$avg_{\rm D} \;\pm\; tS_{\rm D}/\sqrt{N}. \qquad (15)$$

Similarly, for $\sum_{i,j}\langle n_i n_j\rangle_{\rm seq}$:

$$avg_{\rm OD} \;\pm\; tS_{\rm OD}/\sqrt{N}. \qquad (16)$$

The degrees of freedom $(N-1)$ in our case is 5. We therefore put $t=6.869$ to derive a 99.9 percent confidence interval. This suggests that the probability of any value of $\sum_i\langle n_i^2\rangle_{\rm seq}$ and $\sum_{i,j}\langle n_i n_j\rangle_{\rm seq}$ to lie outside the range of (15) and (16) respectively is just 0.1 per cent.

To show the results (Tab. 4), we choose 3 GAPDH CDS with X values closest to 1 compared to X for any other CDS (see Tab. 2). In each case, the results show that $\sum_i\langle n_i^2\rangle_r$ and $\sum_{i,j}\langle n_i n_j\rangle_r$ values are far outside the range of (15) and (16). Therefore the deviations between $\langle n_i^2\rangle_{\rm seq}$ and $\langle n_i^2\rangle_r$ and between $\langle n_i n_j\rangle_{\rm seq}$ and $\langle n_i n_j\rangle_r$ as well are statistically significant with a confidence limit of 99.9 per cent. Here we would like to add that we have also carried out the standard 3-sigma test with N, the sample size, greater than 30; the differences $\langle n_i^2\rangle_{\rm seq} - \langle n_i^2\rangle_r$ and $\langle n_i n_j\rangle_{\rm seq} - \langle n_i n_j\rangle_r$ are well above 3 times the standard deviation recorded from the samples.

Analysis of data

Having now established the level of statistical significance let us look at Table 1. The values of β and X of these eukaryotic groups show that fungi have the highest, followed by invertebrates. The values reach the minima for

Table 2. The β and the X values of the GAPDH CDS from the bacterial species that have been used in our study (source: GenBank and EMBL databases). The per cent values of A, C, G and T denote the strand bias in the GAPDH CDS for each species.

Organism	Accession No.	β	X	%A	%C	%G	%T
1. *Bacillus/Clostridium* gr.							
Bacillus megaterium	M87647	0.07663	1.01191	33	20	20	27
Bacillus subtilis	X13011	0.07431	1.00928	31	22	22	25
Clostridium pasteurianum	X72219	0.07838	1.0049	36	15	20	29
Lactobacillus delbrueckii	AJ000339	0.08528	1.0186	27	25	22	26
Lactococcus lactis	L36907	0.06039	1.0025	31	17	23	29
2. Proteobacteria							
Pseudomonas aeruginosa	M74256	0.08161	1.00398	18	36	31	15
Escherichia coli	X02662	0.08366	1.00471	26	26	24	24
Brucella abortus	AF095338	0.05718	1.00639	22	32	26	20
Zymomonas mobilis	M18802	0.07721	1.00485	22	29	25	24
Rhodobacter sphaeroides	M68914	0.06546	1.00609	18	33	33	16
Xanthobacter flavus	U33064	0.06847	1.00148	19	36	30	15
3. Cyanobacteria							
Anabaena variabilis	L07497	0.04029	0.99277	27	23	24	26
Synechococcus PCC 7942	X91236	0.05028	1.0002	24	28	26	22
Synechocystis PCC 6803	X83564	0.06042	0.99122	26	26	26	22

vertebrates. We conclude β and X decrease with evolution of the GAPDH CDS. The data further suggest that fungi and invertebrates came around the same time and evolved in parallel for a length of time. Vertebrates came later in evolution. Here it might be worth mentioning that fossil records suggest both fungi and invertebrates originated during the Cambrian, Ordovician and Silurian periods; vertebrates came somewhat later, during the Silurian and Devonian periods [8].

Let us now look at 14 bacterial species from three groups: cyanobacteria, proteobacteria (that includes vast majority of gram-negative bacteria), and the *Bacillus/Clostridium* group, a type of gram-positive bacteria. Table 2 summarises β and X values of these samples along with the strand bias. These bacterial groups arose during the Precambrian period of geological time-scale, but there are several schools of thought regarding their specific times of divergence within this period.

We approach the bacterial GAPDH CDS with two differing statistical measures, β and X. Interestingly, both give us almost identical trends (Fig. 1). *Lactobacillus delbrueckii*, a member of the *Bacillus/Clostridium* group, has the highest β and X values (Tab. 2). There is then a large measure of overlap between the *Bacillus/Clostridium* group and the proteobacteria (Fig. 1). The extent of overlap of β values is somewhat more than that of X. The cyanobacterial samples have the minimum values of β and X. There is no overlap between the cyanobacterial values of β and X with the *Bacillus/Clostridium* group. The overlap between the proteobacteria and the cyanobacteria is small. Only one proteobacterial sample, *Brucella abortus* has greater β value than the cyanobacterial member, *Synechocystis* sp. (strain PCC 6803).

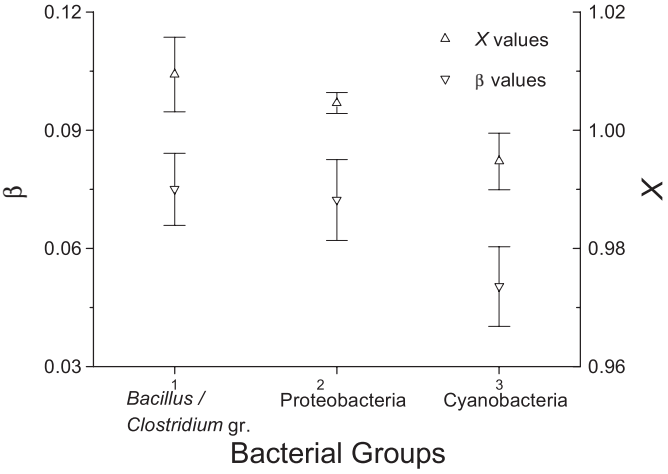

Fig. 1. The average β (denoted by $\triangledown$) and X (denoted by $\triangle$) values for the GAPDH CDS from three bacterial groups (see Tab. 3). The error bars indicate the standard deviation from the average values.

The averages of β or X has the maximum value in the *Bacillus/Clostridium* group, followed by the proteobacteria, while the cyanobacteria samples have the lowest values. In line with our observations on the eukaryotes, we propose (Fig. 1) that the *Bacillus/Clostridium* group originated some time before the proteobacterial species, but later both groups evolved in parallel. The cyanobacterial samples are of recent origin compared to these groups.

Table 3. The average β and X values of the GAPDH CDS for the three bacterial groups, along with the range of deviations in the respective groups.

Group	β	X
Bacillus/Clostridium	0.075 ($\pm$0.00913)	1.00944 ($\pm$0.0063)
Proteobacteria	0.07227 ($\pm$0.01029)	1.00458 ($\pm$0.00177)
Cyanobacteria	0.05033 ($\pm$0.01007)	0.99473 ($\pm$0.0048)

Table 4. The segment-wise analysis of three GAPDH CDS to study the level of statistical significance for the experimental values. The random values are found to lie well outside the 99.9 per cent confidence interval of t-distribution for the experimental values.

Acc. No.:	X91236				
S_l	S_n	$\left\langle \sum_i \langle n_i^2 \rangle_{\rm seq} \right\rangle$	$\langle \sum_i \langle n_i^2 \rangle_r \rangle$	$\left\langle \sum_{i,j} \langle n_i n_j \rangle_{\rm seq} \right\rangle$	$\left\langle \sum_{i,j} \langle n_i n_j \rangle_r \right\rangle$
900	300	4.44	4.51845	2.28	2.24078
921	307	4.42671	4.51705	2.28665	2.24147
942	314	4.42675	4.51475	2.28662	2.24262
960	320	4.425	4.51436	2.2875	2.24282
981	327	4.41284	4.51468	2.29358	2.24266
1020	340	4.41177	4.51559	2.29412	2.2422
		$avg_D \pm t.S_D/\sqrt{N}$	=4.42384$\pm$0.02931	$avg_{OD} \pm t.S_{OD}/\sqrt{N}$	=2.28808$\pm$0.01464
900	300	4.48	4.6919	2.26	2.15405
921	307	4.4658	4.68485	2.2671	2.15758
942	314	4.4586	4.67938	2.2707	2.16031
960	320	4.4625	4.6755	2.26875	2.16225
981	327	4.44343	4.6743	2.27829	2.16285
1008	336	4.43452	4.67417	2.28274	2.16292
		$avg_D \pm t.S_D/\sqrt{N}$	=4.45747$\pm$0.04568	$avg_{OD} \pm t.S_{OD}/\sqrt{N}$	=2.27126$\pm$0.02286
900	300	4.66	4.56215	2.17	2.21893
921	307	4.66775	4.56304	2.16612	2.21848
942	314	4.66879	4.56357	2.16561	2.21822
960	320	4.65625	4.56737	2.17188	2.21632
981	327	4.64526	4.5669	2.17737	2.21655
1014	338	4.68639	4.57029	2.15681	2.21486
		$avg_D \pm t.S_D/\sqrt{N}$	=4.66407$\pm$0.03898	$avg_{OD} \pm t.S_{OD}/\sqrt{N}$	=2.16797$\pm$0.01949

The trends in β and X give us identical patterns that segregate the bacterial species into groups. Amusingly, the results are largely in agreement with what is accepted so far regarding the phylogenetic relationships among these three groups [9]. Our study of the GAPDH CDS, its word distributions, and the moments gives us the measures to propose relative positions of the bacterial groups in the phylogenetic tree.

As we look at the standard deviations of average β and X values for different eukaryotic (Tab. 1) and bacterial (Tab. 3) groups, we find they are almost similar. This reestablishes the fact that the bacterial community is as diverse as Eukarya in structure, function, habit and habitat, and unanimously regarded as a separate domain. Cyanobacteria, proteobacteria and *Bacillus / Clostridium* group which we deal with represent three broad categories of the bacterial domain; therefore there is no reason to expect any smaller deviations in their average β and X values compared to those for the three eukaryotic groups.

At the level of the nucleotide letters A, C, G and T, the order is measured by the quantities X, $X_{\rm D}$ and $X_{\rm OD}$. As we look into the diagonal averages $X_{\rm D}$, (12), we find it increases with evolution. For the window of size 3, this growing diagonal moment implies a rising persistent correlation. In consequence, the off-diagonal averages $X_{\rm OD}$, (13), go down, decreasing antipersistence. Looked at from the letters, the sequences become less uniform and deviate more from the random sequence of identical strand bias. The order, or the information, in the arrangement of letters shows a rising trend with evolution.

Codon usage bias is well studied [10]. This bias leads to the peak at $f = \frac{1}{3}$ in the power spectrum of the CDS. The period 3 oscillation in the mutual information function also is ascribed to codon usage. These have been used

extensively in routines that separate coding from non-coding sequences [11]. In our case the X measures persistence/antipersistence within codons. There are triplets (AAA, CCC) that have same nucleotide in all the 3 positions; in some other codons (AAC, GTG etc.), out of 3 nucleotides, two are identical; while the rest (AGT, GCA etc.) have different nucleotides in all 3 positions. Our findings suggest that as new codons are added, a bias develops that increases persistence within codons with evolution. It is known that while the younger groups of organisms bear rich set of vocabulary in the coding sequences compared to the older groups in geological time scales, both synonymous and non-synonymous codons are used non-randomly [10]. Our results, we believe, would add information to this extensively growing literature on codon usage bias [10,11].

Does any CDS that is an evolutionary marker evolve in ways similar to the GAPDH? We have worked with the CDS of other glycolytic enzymes, such as phosphoglycerate kinase, and found they behave similarly. Other evolutionary markers such as the ribulose-1,5-bisphosphate carboxylase/oxygenase enzyme large segment (rbcL) show similar behaviour. We use these data for biological subclassification. The CDS for ribosomal RNA is another class of sequence that is being investigated. It does not code for protein, but for RNA, and has periods other than 3. The 3 period does exist, but is not predominant.

Sequence modeling has recently become important. The fractal correlations in the sequences led to the expansion-modification system [12]. Later came the insertion models [13]. Here the differences in the CDS and non-coding sequences were observed and the non-coding sequences modeled. The unifying models of copying-mistake-maps [14] modeled both the coding and the non-coding parts. In these models the statistical features of the non-coding sequences have received emphasis. The evolutionary features of the GAPDH CDS isolates the statistical aspects that underlie evolution in coding sequences. The statistics of the word distributions and the subtle cross current of the second moments, we hope, will lead further in these efforts.

S.C. thanks Professor Anjali Mookerjee for many discussions. W.A.K. is supported by the John Fulbright foundation in the laboratory of J.C.

References

1. D.B. Searls, Bioinformatics **13**, 333 (1997).
2. R.F. Voss, Phys. Rev. Lett. **68**, 3805 (1992); W. Li, K. Kaneko, Europhys. Lett. **17**, 655 (1992); C.-K. Peng, S.V. Budyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Nature **356**, 168 (1992); See also S. Nee, Nature **357**, 450 (1992); V.V. Prabhu, J.M. Claverie, Nature **359**, 782 (1992); S. Karlin, V. Brendel, Science **259**, 677 (1993); D. Larhammer, C.A. Chatzidimitriou-Dreissman, Nucleic Acids Res. **21**, 5167 (1993); A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995); R. Roman-Roldan, P.B. Galvan, J.L. Oliver, Pattern Recogn. **29**, 1187 (1996); H. Herzel, E.N. Trifonov, O. Weiss, I. Große, Physica A **249**, 449 (1998); For earlier work, see C. Fuchs, Gene **10**, 371 (1980); R. Nussinov, J. Biol. Chem. **256**, 8488 (1981). For a recent review, see W. Li, Comput. Chem. **21**, 257 (1997).
3. V. Brendel, J.S. Beckmann, E.N. Trifonov, J. Biomol. Struct. Dyn. **4**, 11 (1986); R. Nussinov, J. Theor. Biol. **125**, 219 (1987).
4. P.A. Pevzner, M.Y. Borodvsky, A.A. Mironov, J. Biomol. Struct. Dyn. **6**, 1013 (1989); R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994); B.E. Blaisdell, A.M. Cambell, S. Karlin, Proc. Natl. Acad. Sci. USA **93**, 5854 (1996); C. Martindale, A. K. Konopka, Comput. Chem. **20**, 35 (1996); P. Chaudhuri, S. Das, Curr. Sc. **17**, 1161 (2001).
5. G.S. Attard, A.C. Hurworth, J.P. Jack, Europhys. Lett. **36**, 391 (1996).
6. A. Som, S. Chattopadhyay, J. Chakrabarti, D. Bandyopadhyay, Phys. Rev. E **63**, 051908 (2001).
7. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959); E.W. Montroll, M.F. Shlesinger, *Nonequilibrium Phenomena II From Stochastics to Hydrodynamics*, edited by J.L. Lebowitz, E.W. Montroll (North-Holland, Amsterdam, 1984).
8. M. Thain, M. Hickman, *The Penguin Dictionary of Biology* (Penguin Books, London, 1994), p. 244; P. Stein, B. Rowe, *Physical Anthropology* (McGraw-Hill, Berkshire, UK, 1995); S.C. Stearns, R.F. Hoekstra, *Evolution: An Introduction* (Oxford Univ. Press, New York, 2000).
9. G.E. Fox, *et al.*, Science **209**, 457 (1980). For alternate views regarding the cyanobacterial origin, see P.J. Keeling, W.F. Doolittle, Proc. Natl. Acad. Sci. USA **94**, 1270 (1997); R.S. Gupta, FEMS Microbiol. Rev. **24**, 367 (2000).
10. D.C. Shield, P.M. Sharp, D.G. Higgins, Mol. Biol. Evol. **5**, 704 (1988); G. D'Onofrio, *et al.*, J. Mol. Evol. **32**, 504 (1991); P.M. Sharp, G. Matassi, Curr. Opin. Genet. Dev. **4**, 851 (1994); N. Sueoka, J. Mol. Evol. **40**, 318 (1995); E.N. Trifonov, T. Bettecken, Gene **205**, 1 (1997); J.M. Comeron, M. Aguade, J. Mol. Evol. **47**, 268 (1998); H. Akashi, A. Eyre-Walker, Curr. Opin. Genet. Dev. **8**, 688 (1998); H. Chiapello, F. Lisacek, M. Caboche, A. Henaut, Gene **209**, GC1 (1998); L. Duret, D. Mouchiroud, Proc. Natl. Acad. Sci. USA **96**, 4482 (1999); M. Kreitman, J. Comeron, Curr. Opin. Genet. Dev. **9**, 637 (1999); S. Kanaya, Y. Yamada, Y. Kudo, T. Ikemura, Gene **238**, 143 (1999); N. Sueoka, J. Mol. Evol. **49**, 49 (1999); B. Lafay *et al.*, Nucl. Acids Res. **27**, 1642 (1999); E.N. Trifonov, Gene Ther. Mol. Biol. **4**, 313 (1999); C. Gautier, Curr. Opin. Genet. Dev. **10**, 656 (2000); A.O. Mooers, E.C. Holmes, Trends Ecol. Evol. **15**, 365 (2000); H. Akashi, Curr. Opin. Genet. Dev. **11**, 660 (2001); R.D. Knight, S.J. Freeland, L.F. Landweber, Genome Biol. **2**, 0010.1 (2001).
11. J. Fickett, C.-S. Tung, Nucl. Acids Res. **20**, 6441 (1992); D.J. States, W. Gish, J. Comput. Biol. **1**, 39 (1994); J. Fickett, Comput. Chem. **20**, 103 (1996); J. Fickett, Trends Genet. **12**, 316 (1996).
12. W. Li, Europhys. Lett. **10**, 395 (1989); W. Li, Phys. Rev. A **43**, 5240 (1991).
13. S.V. Buldyrev, *et al.*, Phys. Rev. E **47**, 4514 (1993).
14. P. Allegrini, *et al.*, Phys. Rev. E **57**, 4558 (1998).