J Mol Evol (1984) 20:167–174

Journal of Molecular Evolution

Cluster Analysis of Genes in Codon Space

G.W. Rowe[1], V.L. Szabo[2], and L.E.H. Trainor[3]

[1] Department of Medicine and Membrane Biology Group, University of Toronto, Toronto, Ontario M5S 1A8, Canada
[2] Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada
[3] Departments of Physics and Medicine, University of Toronto, Toronto, Ontario M5S 1A8, Canada

Summary. We construct a "codon space" in which a given DNA sequence can be plotted as a function of its base composition in each of the three codon positions. We demonstrate that the base composition is very highly nonrandom, with sequences from more primitive organisms having the least random compositions. By using cluster analysis on the points plotted in codon space we show that there is a strong correlation between base composition and type of organism, with the most primitive organisms having the highest A or T content in the second and third codon positions. A smooth transition toward lower A + T and higher G + C content is observed in the second and third codon positions as the evolutionary complexity of the organism increases. Besides this general trend, more detailed structure can be observed in the clustering that will become clearer as the data base is increased.

Key words: Codon space — Cluster analysis — Evolutionary paths

I. Introduction

Since the first complete viral genome was fully sequenced, some 6 years ago, the development of several efficient methods for determining the base sequences of DNA and RNA has resulted in an ever-increasing rate of accumulation of sequence data. Between March 1982 and March 1983 the Los Alamos Library *Genbank* has expanded from some 4×10^5 to over 1.3×10^6 bases, from 1528 separate sequences. Several of the sequences represent complete genomes of mitochondria, viruses, and phages, with numerous smaller segments of DNA from vertebrates, invertebrates, bacteria, viruses, and phages also represented.

Contained in these sequences are the genes for a wide range of proteins, ribosomal RNAs, transfer RNAs, and noncoding regions (whose functions, in most cases, remain unknown). Such a vast and growing cache of information provides an ample data base for investigating the patterns of information storage and transmission in biological systems, and for studying the origin and evolution of such patterns from their nonbiological precursors. Studying the interactions of nucleic acids with each other and with other biological building blocks such as amino acids and lipids under presumed prebiotic conditions will, one hopes, lead to an understanding of how such complexity can arise.

Attempting to simulate the evolution of the detailed structure of any one DNA sequence will be futile until we have a more general understanding of the properties of sequences from various types of organisms, or coding for the genes of particular families of proteins. Defining intuitively meaningful measures of these general properties of DNA sequences will allow us to view the organization of biological informational systems more clearly.

Of the four main classes of DNA sequences (protein coding, rRNA coding, tRNA coding, and noncoding), we have chosen to study the protein-coding sequences, because they form the largest data base and are both structurally and functionally the best understood (we are currently extending our studies to the other forms of DNA sequences). Previous work in this area has revealed many regularities in the choice of degenerate codons for a given amino acid (Grantham et al. 1980, 1981; Shepherd 1981) and more persistent correlations in the choice of

Offprint requests to: Dr. Glenn W. Rowe, Dept. of Mathematics, University of Dundee, Dundee, Scotland DD1 4HN

base in each codon position (Rowe and Trainor 1983a). Attempts to relate the various codon biases to the source or function of the gene have shown that such relations do exist (Grantham 1980), but owing to the nature of the analysis, exactly what properties of the sequence are responsible is unclear.

We have discovered a measure of these sequence correlations that gives a much clearer picture of how they relate to the sources of the sequence. In the next section we define the measure, show how it can be used to define a nine-dimensional "codon space" in which DNA sequences can be plotted, and discuss some properties of the space. In Section III we apply cluster analysis to the set of points in the space to locate the regions of space belonging to various subsets of points. We show that the members of each subset correspond primarily to genes of a common source or function. Finally, Section IV summarizes our work and indicates the directions of our ongoing investigations.

II. Codon Space

We represent a DNA sequence by the number of each of the four bases in each of the three codon frames. We define the matrix M′ as

$$M' = \frac{1}{N}\begin{bmatrix} A_1 & T_1 & G_1 & C_1 \\ A_2 & T_2 & G_2 & C_2 \\ A_3 & T_3 & G_3 & C_3 \end{bmatrix} \tag{1}$$

where A_1 = number of As in codon position 1, etc., and N = length of sequence.

Only three of the four entries in each row of M′ are independent, since the sum of the entries in each row is N/3. We may therefore drop one column from M′ to obtain the matrix M with nine independent elements:

$$M = \frac{1}{N}\begin{bmatrix} A_1 & T_1 & G_1 \\ A_2 & T_2 & G_2 \\ A_3 & T_3 & G_3 \end{bmatrix} \equiv \begin{bmatrix} a_1 & t_1 & g_1 \\ a_2 & t_2 & g_2 \\ a_3 & t_3 & g_3 \end{bmatrix} \tag{2}$$

Note that the maximum value of any of the (lower-case) entries in M is ⅓, and that should one element be ⅓, all other elements in that row have values of 0.

The elements of M may be taken to span a nine-dimensional Cartesian space. This space is composed of three three-dimensional subspaces, each corresponding to one codon position, e.g., for codon frame 1, a sequence can be represented by a vector (a,t,g). For each codon frame, then, we can construct a three-dimensional plot showing the location of all the gene sequences.

To interpret the plots so obtained, we need to know the distribution of points in this space if a set of sequences is chosen at random. We can then compare the distribution of points from natural DNA with the random distribution in order to pick out anomalies. (Even if no anomalies in the distribution are detected, there may still be significant clustering of the natural sequences according to origin or function of gene product.) To this end let us calculate the density of states function, which determines the number of different sequences occupying each point in codon space. If we consider a three-dimensional subspace corresponding to a single codon frame and define n = N/3 and n_A = number of As in the codon frame, with n_G, n_T, and n_C defined similarly, then the number p of possible sequences at a given point is

$$p = \frac{n!}{n_A! \times n_T! \times n_G! \times n_C!} \tag{3}$$

If all of n, n_A, n_T, n_G, n_C are >10, we may use Stirling's approximation for n! ($n! \approx \sqrt{2\pi n}\, n^n e^{-n}$) to obtain

$$p \approx \frac{1}{(2\pi)^{3/2}}\left(\frac{n}{n_A \times n_T \times n_G \times n_C}\right)^{1/2} \cdot \frac{n^n}{{n_A}^{n_A} \times {n_T}^{n_T} \times {n_G}^{n_G} \times {n_C}^{n_C}} \tag{4}$$

Defining

$$\alpha = n_A/n; \quad \beta = n_T/n; \quad \gamma = n_G/n; \quad \delta = n_C/n$$

where $\alpha + \beta + \gamma + \delta = 1$, then

$$p \approx (2\pi n)^{-3/2}(\alpha^\alpha \beta^\beta \gamma^\gamma \delta^\delta)^{-n}(\alpha\beta\gamma\delta)^{-1/2} \tag{5}$$

As $n \to \infty$, $p \to \infty$, indicating that for a sequence of infinite length, an infinite number of different sequences is possible with given base composition.

A difficulty with Eq. (5) is that p is a function of n, the sequence length, so a direct comparison of the plot of a number of sequences of varying length with a single density of states function is not possible. However, we can define a probability density as a function of n:

$$\rho(\alpha,\beta,\gamma,n) = \frac{p(\alpha,\beta,\gamma,n)}{\int p(\alpha',\beta',\gamma',n)\, d\alpha'\, d\beta'\, d\gamma'} \tag{6}$$

where the limits of the integral are 0 to 1 for all three variables. The use of an integral in Eq. (6) despite the fact that α, β, and γ are discrete variables is no cause for concern, since the Monte Carlo method used to evaluate the integral (see below) is applicable to both sums and integrals. The meaning of Eq. (6) is as follows. Suppose we are given α_o, β_o, γ_o, and n. Then the probability of a sequence of length n, chosen at random, having a value of p less than or equal to $p(\alpha_o,\beta_o,\gamma_o,n)$ is:

$$P = \int_\Omega \rho(\alpha,\beta,\gamma,n)\, d\alpha\, d\beta\, d\gamma \tag{7}$$

where Ω is that volume of α–β–γ space such that $\rho(\alpha,\beta\gamma,n) \leq \rho(\alpha_o,\beta_o,\gamma_o,n)$. Thus, Eq. (7) gives the probability of finding a sequence that has either base compositions α_o, β_o, γ_o or a sequence that is less likely than this. The most probable sequence is one with $\alpha = \beta = \gamma$. The value of P for such a sequence is unity.

The evaluation of Eq. (7) can be achieved via the Monte Carlo method described by Gillespie (1975). The algorithm is as follows:

1. Generate a random point x inside Ω.
2. Calculate $y = f(x)$; keep running totals of y and y^2.
3. Perform steps 1 and 2 N_t times, then

$$A_1 = \sum y/N_t; \quad A_2 = \sum y^2/N_t$$

Then

$$\int_\Omega f(x)\,dx \approx |\Omega|\left[A_1\left(\frac{A_2 - A_1^2}{N_t}\right)^{1/2}\right] \qquad (7a)$$

We have a complication here in that the volume Ω is unknown. However, if we generate random points $x \equiv [\alpha,\beta,\gamma]$ and count the fraction of these points that give $P(x,n) < P(x_o,n)$ this fraction is the fraction of the total volume that is included in the integral. Thus the Monte Carlo method will give us the integral *and* the volume without any need to know the shape or precise location of the volume.

The total volume available from which to choose x is defined by $\alpha + \beta + \gamma < 1$ and is the tetrahedron shown in Fig. 1.

The integral (7a) was evaluated for all three codon frames for the 332 genetic sequences in the Los Alamos Library longer than 150 base pairs (this restriction was imposed to ensure statistical significance of the sequences included in the analysis). One million (10^6) points were used in the integral evaluation. This number gives an accuracy of about $\pm 1\%$ for sequences 250 base pairs long, decreasing to $\pm 10\%$ accuracy for a sequence length of 5000 base pairs. This accuracy is more than sufficient to illustrate our point, as we shall now see. The results are shown in Fig. 2. It can be seen that for all three frames, the vast majority of sequences are found in a region of codon space where the probability of finding a sequence chosen at random is less than 0.1. Many sequences have an almost vanishingly small probability ($< 10^{-10}$) of being found in a random choice.

We see, therefore, that the base composition of each codon frame is highly nonrandom. Two questions arose naturally at this point: (1) Is there some connection between the location of a gene in codon space and the origin and/or function of the gene? (2) How and why did such nonrandomness arise? The first question will be addressed in the next section, while the second question will be tackled in Section IV.

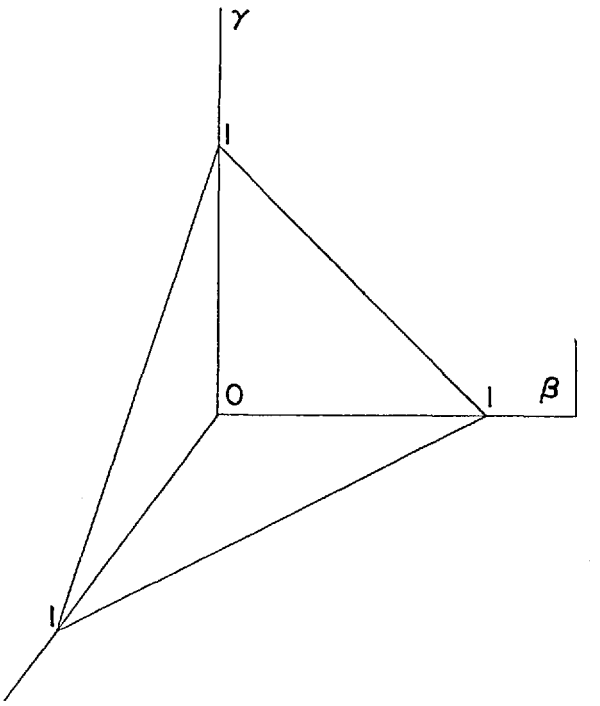


Fig. 1. The tetrahedral volume over which the integrals in the text are performed

III. Cluster Analysis of Genetic Sequences

The 332 genetic sequences used were plotted in the nine-dimensional codon space according to coordinates defined by Eq. (2). To search for structure in the mass of points so generated, we used cluster analysis to organize the points into 16 groups. The nonhierarchical clustering procedure is described in detail by Anderberg (1973), so only a summary is presented here.

The objective is to group the data into a number N (chosen by the investigator) of nonoverlapping clusters according to the following algorithm.

1. Choose N initial points to serve as centroids for the clusters.
2. Assign each datum to the cluster with the nearest centroid.
3. Calculate N new centroids as the average positions of the data in the N clusters.
4. Repeat steps 1 to 3 until no data are reassigned in step 2.

The distance measure (metric) we used to determine the distance between two points in codon space is the so-called city-block metric, defined as

$$d_{km} = \sum_{i=1}^{\nu} |x_{k_i} - x_{m_i}| \qquad (8)$$

where d_{km} is the distance between points x_k and x_m (with coordinates x_{m_i}) in a ν-dimensional space. In

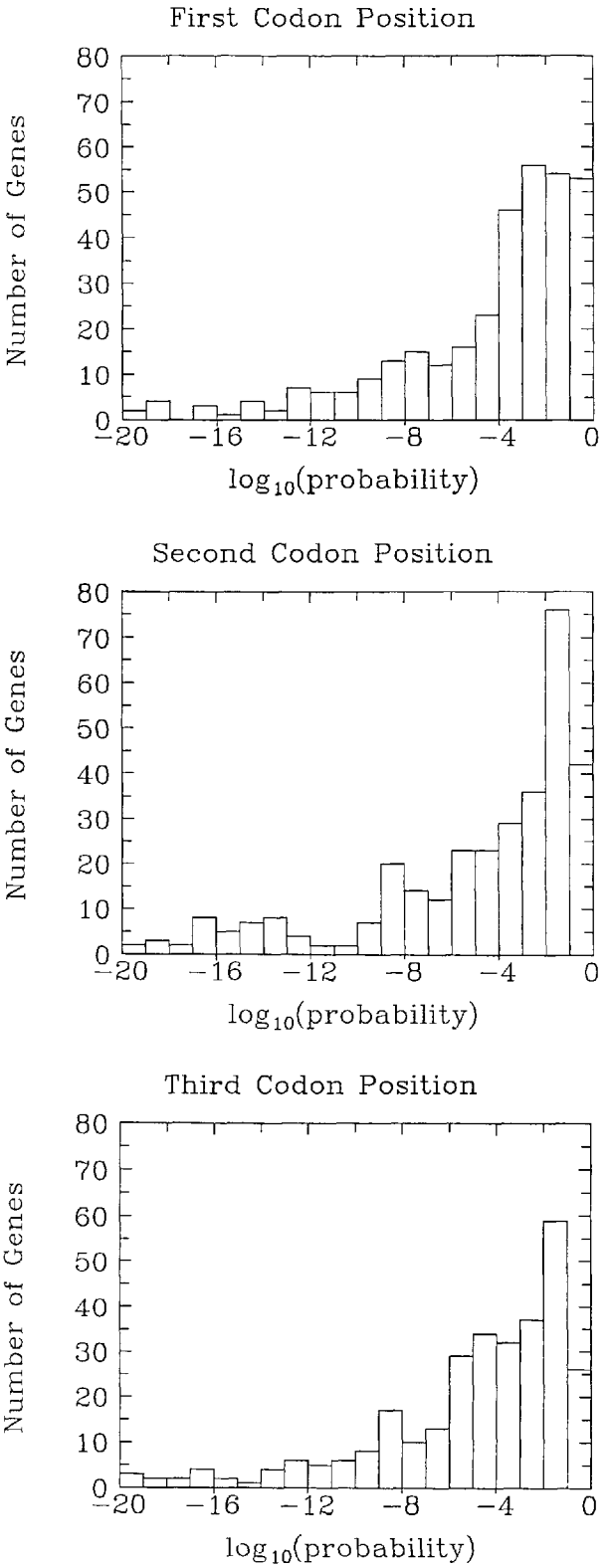


Fig. 2. Histograms of the numbers of genetic sequences found in regions of codon space. The probability of finding a random sequence with the same base composition is plotted on the abscissa

this study, $\nu = 9$. This is consistent with the intersequence distance defined in evolutionary tree work, except we are comparing groups of bases instead of single base pairs. The clustering obtained is independent of the metric used provided all coordinates enter equally into x_k and x_m. If we use weighting, e.g.,

$$d'_{km} = \sum_{i=1}^{\nu} w_i |x_{k_i} - x_{m_i}| \qquad (9)$$

where w_i is the weight for the i^{th} coordinate, the choice of metric is important.

The cluster analysis we present here uses equal weights. Essentially the same clusters are obtained if any one codon frame is omitted (i.e., if the three w_i values cooresponding to a particular codon frame are set to zero).

The final clusters obtained depend to a certain extent on the location on the starting centroids. After the general trends in the clustering became apparent to us, we tried to choose our initial centroids to correspond roughly to where we expected the final centroids to lie. No matter how much we distorted the initial choices for the centroids, however, the general trends in the clusters persisted. We used 16 clusters, since this number gave the clearest separation into groups whose members had distinct properties. Fewer clusters resulted in unrelated genes being grouped together, while a number greater than 16 split groups of common genes into separate clusters.

We found distinct clusters for (i) vertebrate mitochondria, (ii) invertebrate (yeast) mitochondria, and (iii) phages, and a trend for the genes from more "advanced" organisms to have a higher G content in all three codon frames, at the expense of A and T content. If the centroids are projected onto the AT plane, one finds phage and mitochondria lying furthest from the origin with a progression through viruses, bacteria, lower eukaryotes, and higher eukaryotes as one proceeds toward the origin. This trend is most noticeable in the third codon frame.

That the clusters we obtained contain genes corresponding to particular types of genes and not random samples is not surprising, since similar clustering of genetic sequences based on codon usage has been observed previously (Grantham 1980). However, our analysis is conceptually simpler, since the meaning of each of our dimensions is well defined.

We present in Fig. 3 three three-dimensional projections of the nine-dimensional clusters. Each plot corresponds to one of the three codon positions, say i, with the a_i, t_i, and g_i values from the matrix M as the x, y, and z coordinates, respectively. The projection of the centroid positions onto the three-dimensional subspace gives the points plotted in each graph. Some caution is required in the interpretation of these plots, since the regions of space in 3-space cooresponding to two of the centroids may overlap (in the original nine-dimensional space

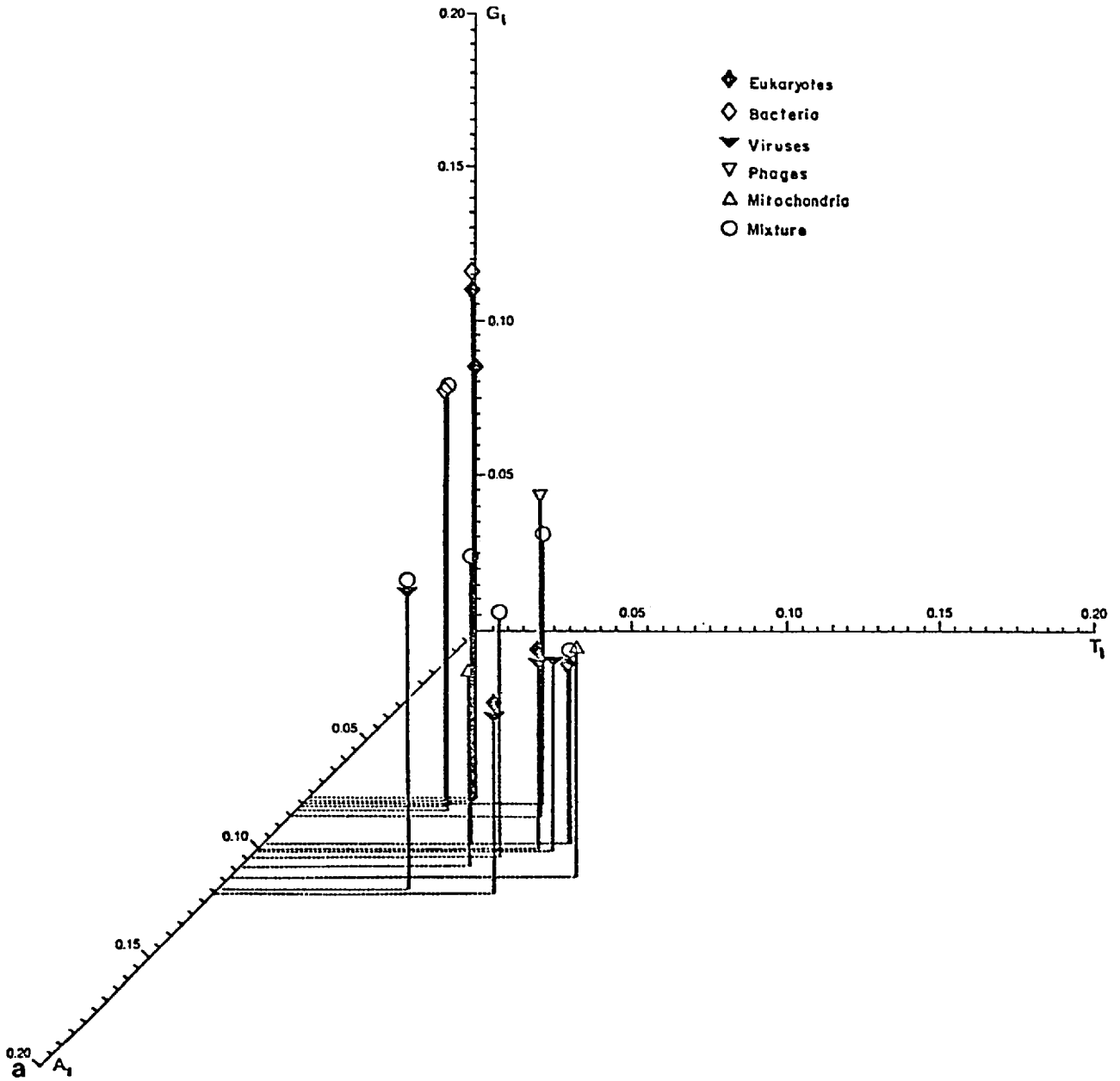


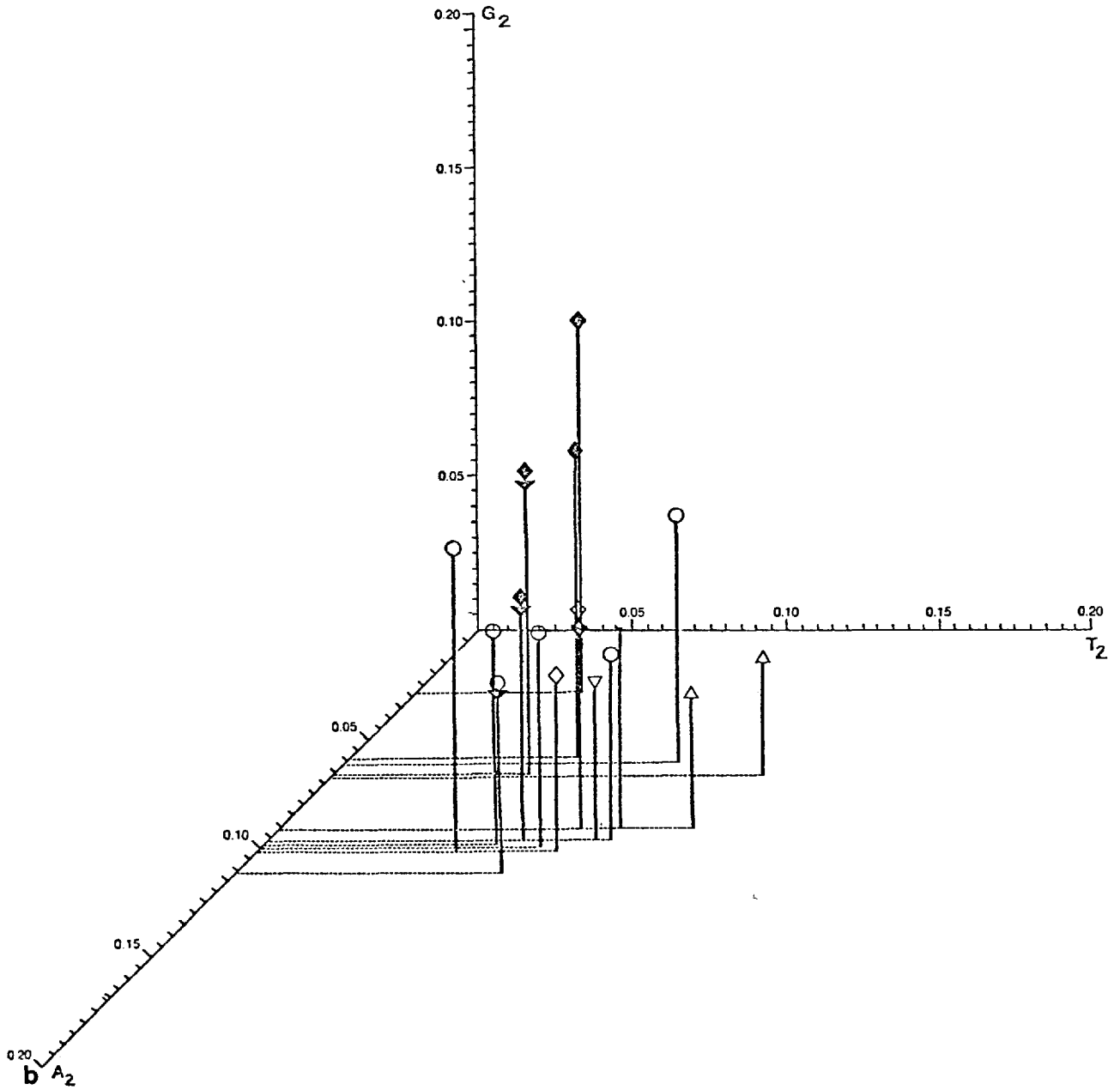
Fig. 3a–c. Locations of the cluster centroids obtained by the cluster analysis. Projections in three three-dimensional spaces corresponding to base compositions in the **a** first, **b** (facing page) second, and **c** (following page) third codon positions. The numbers on the axes correspond to entries from the matrix M in the text. Where two symbols appear at a centroid the corresponding cluster contains genes from both categories. A mixture contains more than two categories of genes

→

each centroid is the center of a region of space that is not shared with any other centroid). This may be seen in a two-dimensional analog (Fig. 4). However, the graphs may be used as a guide to show, on average, how far the members of each cluster are from each other in each codon position.

As can be seen, little information is contained in the projection onto first-codon-position coordinates, whereas the second and third positions show quite striking partitioning of the space. Let us examine the third position first. Starting with those clusters furthest from the g_3 axis, we see that mitochondria have very large a_3 values, but that mammalian mitochondria have very different t_3 values from yeast and mold mitochondrial genes. (The subdivision of the mammalian mitochondria cluster reveals that human and mouse mitochondria form two completely separate clusters, with mouse mitochondria tending to higher a_3 values.)

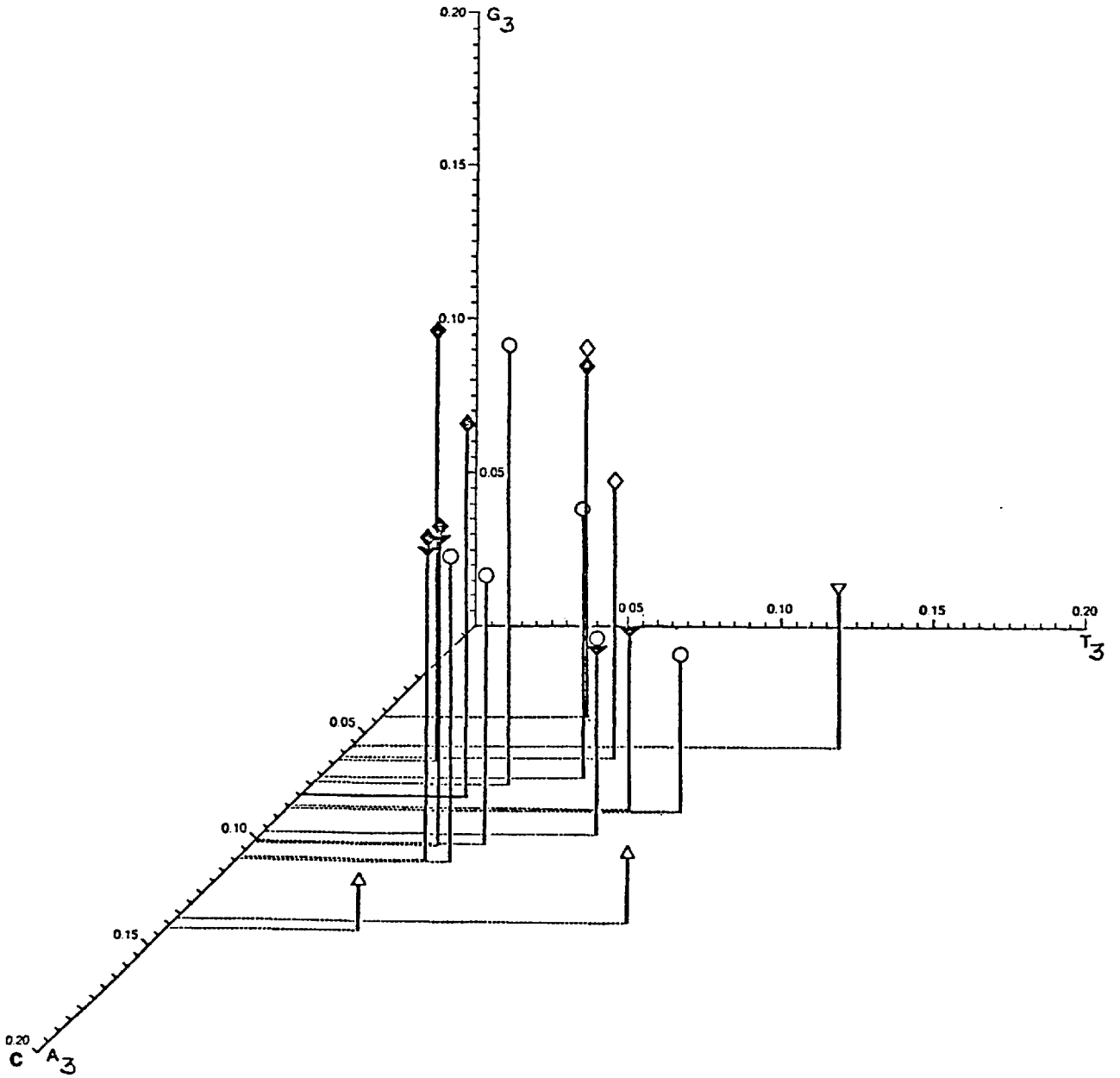
The phages tend to have high t_3 but low a_3 values.



Moving progressively closer to the g_3 axis, we find viral clusters, bacterial clusters, and finally eukaryote clusters. The eukaryote gene clusters are spread over a large area, but the only clusters containing almost exclusively eukaryotic genes are found closest to the g_3 axis and have the highest g_3 values.

Although the points are not as widely spaced in the second codon position, the same trend is evident here as well. Omitting mitochondria, which have their own genetic code and thus cannot be compared directly with those organisms using the "normal" code, an evolutionary sequence of the main groups of organisms would read: phages, viruses, bacteria, invertebrates, vertebrates. Although we have insufficient data to distinguish vertebrates from invertebrates we can observe an inverse correlation between evolutionary complexity and distance from the g axis in both the second and third codon positions. Also, the closer one gets to the g axis, the higher the value of g.

On the basis of the preceding arguments we infer that the main determining factor for the clusters is the kind of organism, not the kind of gene. Some discrimination of the basis of gene function is possible, however. Almost all the globin genes fall into the same cluster, and although immunoglobulin genes do not fall into a single cluster, they tend to avoid regions of space (such as the globin cluster) occupied by other eukaryotic genes. More data will help to clarify this situation, and we are currently using extended sequence banks to update this study.



What causes these striking biases? Several possibilities are reviewed by Grantham (1980):

1. Optimization of error rates with availability of each of the four bases in the environment, i.e., varying availability of the mononucleotides.
2. Regulation of replication or transcription.
3. Regulation of expression of mRNA in translation.
4. Optimization of mRNA secondary structure.
5. Harmony of codon–anticodon populations.

All of these options have their problems, be it inconsistency at some point in the argument or simply lack of supporting data.

Two of us (Rowe and Trainor 1983b) have proposed that viral codon bias is the remnant of a periodic sequence with which primitive DNA or RNA formed. We showed that the periodic sequence could be the most energetically favorable structure for a nucleic acid in a primordial soup. Varying environments could result in varying periodicities in the sequence. Because of the smooth transition in our codon space from simpler to more complex organisms as one traces a path toward the g_2 or g_3 axis, however, it seems likely that this variation in the environment was the increasing complexity of the media in which the DNA found itself as evolution progressed, rather than some random fluctuation in such parameters as nucleotide concentrations, salt concentrations, or temperature. As living systems became more complex, the nucleic acid sequence adapted so as to preserve as much stability as pos-

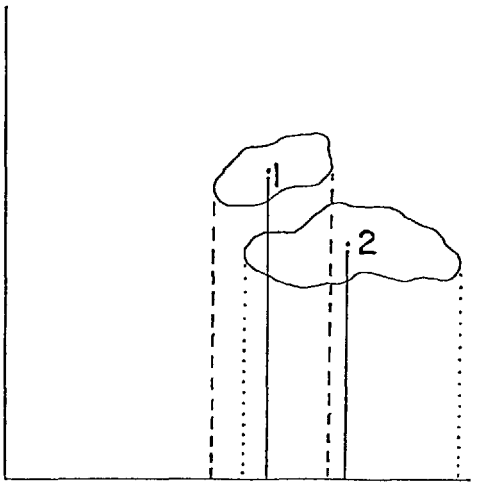


Fig. 4. Illustration showing that although clusters 1 and 2 occupy distinct regions of space in two dimensions, their projections onto one dimension overlap

sible while accommodating the increased information load it had to carry.

VI. Conclusions

Using the main result of our previous paper, that codon order in genes is random, we have constructed a nine-dimensional codon space in which the axes represent A, G, and T content in each codon frame. Natural protein-coding sequences are plotted in this space. The resulting distribution is highly nonrandom. Cluster analysis reveals the tendency of more "primitive" genes to lie further from the origin than nuclear genes from higher organisms.

This work, we believe, is an improvement on previous attempts to correlate codon bias with gene origin. The most comprehensive of these earlier studies, that of Grantham et al. (1981), plots sequences in a 61-dimensional space. Further, the clusters obtained have no easily visualizable interpretation—the axes of the plots are unlabeled. Our work shows clearly how the clusters relate to base content and codon frame.

The question remains: How and why did such biases originate? Two of us (Rowe and Trainor 1983b) have proposed that the prebiotic form of DNA was a periodic polymer (e.g., GCTGCT), because such a sequence was the most stable. The biases observed in modern coding regions are the remnants of this ancient sequence.

Once such a sequence becomes incorporated in primitive organisms, the "environment" seen by it is the internal environment of the cell, an environment that will, in general, be different from the external environment. DNA sequences will then be under an evolutionary pressure to form stable sequences appropriate to the new environment, while at the same time containing enough variability to be useful as information carriers. Perhaps the strategy in developing more complex information-processing capabilities for the genome is to adjust the environment of DNA such that the base stacking energies are most nearly equal, thus allowing maximum sequence variety.

The fact that different organisms have different codon biases requires a deeper analysis. Because of the trend for the DNA of more primitive organisms to be further from the origin of codon space than that of higher organisms (which would be expected if primitive DNA were indeed periodic, since a point far from the origin indicates a strong bias in favor of one or two bases at the expense of the others), we are currently exploring a model of DNA sequence evolution through codon space. The model envisages codon space containing islands of stability, the locations of which are determined by evolutionary constraints and driving forces. Thus there is no unique "ideal point" in codon space where a gene would be optimally suited to all environments; rather, genes evolve in space as their surroundings dictate.

Acknowledgments. The authors would like to thank Dr. Charles Lumsden for discussion; Division of Medical Computing, University of Toronto, for computing support on the VAX 11/780 system; and Anne Hansen-Johnston for her careful preparation of the manuscript. The research reported here was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Medical Research Council of Canada. Glenn W. Rowe is a Post-Doctoral Fellow of the Medical Research Council. Victor Szabó holds a Natural Sciences and Engineering Research Council of Canada Post-Graduate Scholarship.

References

Anderberg MR (1973) Cluster analysis for applications. Academic Press, New York

Gillespie DT (1975) The Monte Carlo method for evaluating integrals. Naval Weapons Center, China Lake, California

Grantham R (1980) Workings of the genetic code. Trends Biochem Sci 5:327–331

Grantham R, Gautier C, Gouy M (1980) Codon frequences in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res 8:1893–1912

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage as a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Rowe GW, Trainor LEH (1983a) On the informational content of viral DNA. J Theor Biol 101:151–170

Rowe GW, Trainor LEH (1983b) A thermodynamic theory of codon bias in viral genes. J Theor Biol 101:171–203

Shepherd JCW (1981) Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a commaless genetic code. J Mol Evol 17:94–102

Received September 6, 1983/Revised January 6, 1984