Indexing Scheme and Similarity Measures for Macromolecular Sequences

C. Raychaudhury and A. Nandy*

Computer Division, Indian Institute of Chemical Biology, 4 Raja S C Mullick Road, Calcutta 700032, India

Received April 15, 1998

Nucleotide composition and distribution along a DNA sequence is known to play a vital role in the determination of gene functions. Protein coding regions, regulatory sequences, and other functional regions are determined generally by homology studies with comparable genes from other species or specific experimental verification. With the rapid and explosive increase in sequence information, new computational techniques for rapid determination of such information and comparative studies of different genes are becoming necessary which ideally should encompass not only DNA sequences but other macromolecular sequences as well.

Geometrization of macromolecular sequences in the form of a graphical representation provides one such technique where the nucleotides in a gene sequence can be viewed as objects in a four-dimensional space; the method can be extended, in principle, to include, say proteins, in a 20-dimensional space. We have found a reduced two-dimensional representation of DNA sequences very useful in studies of nucleotide distribution and composition, especially in comparative studies of homologous sequences, providing rapid visual clues to sequence identity and similarities and differences in distribution of bases along the sequence. Our studies of gene sequences of the globin family, the myosin heavy chain genes, histones, and others in such a representation has shown that evolutionary changes produce shifts in base distribution that appear to reflect evolutionary distances in the dispersion of the graphical form of the sequences. However, a mathematical scheme to quantitatively evaluate the various graphical representations and provide a useful classification method has not been available so far.

In a previous paper we had evolved a method of measuring the dispersion in the coordinate data of the 2D graphical representation of these sequences that is suitable for individual intron and exon segments. We here propose a new measure of the dispersion of DNA graphs that can be used to quantify the differences between two or more graphs of genes of various organisms. This provides a method to index a graph and formulate similarity measures to compute distances between two or more graphs thus generating comparison tables. The technique presented here is an improvement over the existing ones in that differences in sequence lengths are normalized out for a more acceptable comparison between sequences of differing lengths; methods such as the Euclidean and Manhattan distance measures are applicable only to the rare cases of equal length sequences. We report here that application of this technique to the α-globin and histone genes provides quantitative support to conclusions arrived earlier through qualitative discussions. It also appears that, once standardized, the proposed scheme may help in rapid identification and retrieval of molecular sequences from electronic sequence libraries and to study molecular phylogeny in evolutionary time scales.

1. INTRODUCTION

Effective representation of long DNA sequences has led to several innovative techniques to provide useful ways of viewing, sorting, analyzing, and comparing various sequences.[1−5] We have used a graphical representation that has proved very useful in sequence comparison and homology studies.[4] The graphs obtained therefrom for individual genes have been found to vary slightly from organism to organism, which has been interpreted to be a consequence of evolutionary divergence.

The method of representing DNA sequences graphically using a two-dimensional Cartesian coordinate system has been explained elsewhere.[4] The shapes of these DNA graphs depend on the base distribution in the sequence. The plot is generated by moving one step in the positive *x*-direction for a guanine (G) in the sequence, the negative *x*-direction for an adenosine (A), the positive *y*-direction for a cytosine (C), and the negative *y*-direction for thymine (T), the succession of such steps producing a shape characteristic of the sequence; while much work has been done using this ACGT-axis system (counting clockwise), other permutations have also been considered.[2,5] We have shown elsewhere[4] that for conserved genes such plots are shape similar thereby making identification of a new sequence of the family possible rapidly and easily by visual inspection alone.

It is clear that the differences in the base composition and distribution of individual members of a homologous family will induce changes in the plots of the sequences in the graphical representation method outlined. Changes in base composition will result in changes in the end-points represented by the coordinates (G−A, C−T), where the letters represent the total number of each base in the sequence; changes will also arise from the differences in distribution of the bases along the sequence as measured by the instantaneous coordinates (g−a, c−t) where the lower case letters represent the number of each base up to the nucleotide number under consideration.

Gates[2] had proposed a Manhattan approach to estimate such differences but only for equal length sequences.

* Corresponding author. E-mail: anandy@giascl01.vsnl.net.in

10.1021/ci980077v CCC: $18.00
Published on Web 11/24/1998

Another attempt was made by us[6] to estimate the divergence of two graphs by calculating a measure of plot density, i.e., the ratio of the number of points and the enclosing area for the points, but this method misses out on differences arising out of shape changes within the same overall distribution. In this paper we propose a method to compare two or more such graphs in order to provide a quantitative estimate of the divergence of the different sequences.

2. METHOD

The map of a gene sequence represented by a sequence of points on our two-dimensional graph may be mathematically characterized by a function of terms of various orders as follows

$$C_{\text{map}} \sim C\,(S_0, S_1, S_2, S_3, ...)$$

where S_0 is the zeroth-order term representing the coordinates x_f, y_f of the end points, S_1 is the linear term representing the first-order moments about the two axes, S_2 the second-order term represents the variance about the mean, and S_3 the third-order term represents the skewness, etc. The zeroth-order term will distinguish between the base compositions of different gene sequences since in our ACGT-axes system the end-co ordinates are given by the total G−A and C−T. Higher order terms are required to trace the differences between the maps representing the base distributions within the sequences. While the first-order moment, variance, skewness, and moments of higher order are required for a proper measure of these differences, for the purposes of the present paper we restrict ourselves up to the first-order term only as a first-order measure of the sequence of the points representing the base distribution patterns.

To obtain a measure of the overall shape of a DNA sequence graph, we define

$$\mu_x = \sum_{i=1}^{N} x(n_i)$$

$$\mu_y = \sum_{i=1}^{N} y(n_i)$$

where $x(n_i)$ and $y(n_i)$ represent the values of the x and y coordinates of the ith nucleotide in the sequence being plotted in the chosen axes system. The μ are then normalized to adjust for the sequence length N as

$$\bar{\mu}_x = \mu_x/N$$

$$\bar{\mu}_y = \mu_y/N$$

where $\bar{\mu}_x$ and $\bar{\mu}_y$ are the first-order moments of the x and y coordinate values respectively, giving the average displacements in the x and y directions. We also define the radius of a DNA sequence graph, g_R by

$$g_R = [\bar{\mu}_x{}^2 + \bar{\mu}_y{}^2]^{1/2}$$

These measures are for individual sequences and their graphs. However, for the purpose of comparison, we define distance between two sequences and hence their graphs g_1 and g_2 as

$$d\,(g_1, g_2) = [(\bar{\mu}_x - \bar{\mu}'_x)^2 + (\bar{\mu}_y - \bar{\mu}'_y)^2]^{1/2}$$

where $\bar{\mu}_x$ and $\bar{\mu}_y$ correspond to the graph g_1 while $\bar{\mu}'_x$ and $\bar{\mu}'_y$ correspond to g_2.

It is clear that the values of the μ are dependent upon the base distribution pattern; the values of the first two measures of map characterization for a hypothetical sequence segment consisting of 5 A's and 5 G's are

sequence	G−A	μ_x	$\bar{\mu}_x$
AAAAAGGGGG	0	25	2.5
AAAGGAAGGG	0	17	1.7
AAAGGGAAGG	0	13	1.3
AGAGAGAGAG	0	5	0.5

Thus, the measure that we have adopted provides a way to differentiate between different distribution patterns of the same sets of bases, which is not necessarily unique, but clearly dependent upon base distribution. Since our interest in the present type of applications lies in comparing sequence graphs that are essentially shape similar, the measure defined above would be sufficient in a first approximation to quantify the differences.

For two shape similar plots where, e.g., slight differences in base composition lead to a small rotation between them, this measure will arrive at a quantitative difference. Thus in the case of two sequences with one having, say, less guanine than the other, the former will generally tend to be shrunk along the x-axis and g_R will thus turn out to be smaller than the sequence with more guanines. The same will hold true for decrease in any one or more of the four bases in one sequence compared to another, and the drop in our dispersion index will be related to the decrease in the compositional differences between A and G and C and T in our ACGT axes system plots.

Comparison of two such graphs is facilitated by invoking the "distance" concept defined earlier. The distance will be zero for two identical and overlapping sequences. As graphs diverge due to the cumulative aspect of base differences, the first-order moments will diverge and lead to increasing distances between the sequences. Closely related sequences can be expected to have small distance values, while unrelated sequences will lead to large distances. Clearly, only relative distances will be meaningful, and absolute values will not be important at this stage of our analysis.

The method outlined here can be generalized to apply to the case of protein and other sequences where one may construct a multidimensional hyperspace to represent the sequences. In a fully generalized model one can then define moments M_k for each dimension k and a generalized graph radius Γ_R

$$M_k = \sum_{i=1}^{N} x_k(n_i)$$

$$\Gamma_R = [\Sigma_k(M_k{}^2)]^{1/2}$$

Distance between two such sequences can then be defined as

$$\Delta(g_1, g_2) = [\Sigma_k(M_k{}^2 - M'_k{}^2)]^{1/2}$$

Table 1

a: The Graph Radius g_R and $\bar{\mu}_x$ and $\bar{\mu}_y$ for the α-Globin Sequences of Various Species

species	EMBL code	$\bar{\mu}_x$	$\bar{\mu}_y$	g_R
horse	ECHBA22	64.82	126.37	142.03
rhesus monkey	MMHBA	56.63	84.64	101.84
orangutan	PPHBA02	47.85	87.97	100.14
goat	CHHBAI	61.13	70.09	93.00

b: The Distance d (g_1, g_2) between Each Pair of the Four Species of Part A

	rhesus monkey	orangutan	goat
horse	42.53	41.98	56.40
rhesus monkey		9.39	15.23
orangutan			22.27

where the M' represent the moments for the second sequence. Again, the distance will be zero for two identical sequences and will diverge for diverging sequences. In this preliminary work, however, we restrict ourselves to the applications with two-dimensional representation of DNA sequences.

3. RESULTS AND DISCUSSION

The prime objective of the present study is to develop quantitative descriptors for nucleotide sequences and the corresponding 2D graphs and to use them in studying similarities of the sequences. In this paper we have proposed three indices, viz. $\bar{\mu}_x$, $\bar{\mu}_y$, and g_R computed from the mean values of the *x*- and *y*-co ordinate data, and a distance measure, d (g_1, g_2), to compare two graphs. From the data in Table 1 we find that the g_R have relatively higher values for the horse α-globin compared to those of the three other species, and the distance indices d (g_1, g_2) have higher values between the α-globins of horse and each of the three other organisms. A deeper look into the background of Table 1 reveals that the $\bar{\mu}_x$ values for the four species are not too different, while the $\bar{\mu}_y$ value for the horse α-globin is quite higher than the three other $\bar{\mu}_y$ values producing appreciably high g_R value for the horse α-globin. The higher index values for horse adequately reflect the difference in shape of the graph for horse (Figure 1) in comparison with the three other species; at the same time, the sequences for rhesus monkey, orangutan, and goat have relatively close index values which also reflect the closeness of the corresponding graphs.

In the case of the histones also, the data given in Table 2 shows that the $\bar{\mu}_x$ and $\bar{\mu}_y$ values for the mouse and human histones are lower compared to those for chicken, wheat and maize resulting in lower g_R values for the first two species. These values also reflect the shapes and sizes of the corresponding graphs in Figure 2. Considering the distance measures d (g_1, g_2) values for these five sequences we see that the indices have low values between pairs of sequences for chicken, wheat, and maize and relatively higher values for the sequences between human and each of these three species. These are also in agreement with the corresponding graphs.

Thus g_R and d (g_1, g_2) can be used in differentiating and quantifying graphs of comparable sequences and lengths. Moreover, sequences and their graphs which move abruptly in opposite directions might be better quantified by g_R and d (g_1, g_2) since in these cases the end coordinate values may

Table 2

a: The Graph Radius g_R and $\bar{\mu}_x$ and $\bar{\mu}_y$ for the Histone H4 Sequences of Various Species

species	EMBL code	$\bar{\mu}_x$	$\bar{\mu}_y$	g_R
wheat	TAH4091	31.28	43.88	53.89
maize	ZMH4C14	37.33	35.54	51.54
chicken	GGHIST4A	33.47	42.39	54.01
mouse	MMHIST4	22.16	23.93	32.61
human	HSHISAD	15.35	11.51	19.18

b: The Distance d (g_1, g_2) between Each Pair of the Five Species of Part a

	maize	chicken	mouse	human
wheat	10.30	2.65	21.94	36.09
maize		7.86	19.10	32.57
chicken			21.65	35.81
mouse				14.17

not be appropriate indicators of the base distribution, e.g., λ repressor gene sequence whose graph moves upward almost monotonically and comes down in a similar fashion [Nandy 1996]. However, for the sequences of the present study, g_R has been adequate for determining the differences in the sequences and the corresponding graphs under consideration. It is noteworthy that the g_R values for α-globin gene sequences have much higher values compared to those for histone H4 sequences indicating that such an index can effectively differentiate sequences of different size (for α-globin genes the sequence length N is in thousands, whereas the length of all the five histone H4 sequences is identical at 312 only). Thus along with the sequence length N, the proposed indices and other sequence related information such as the number of each type of nucleotide, their ratios, etc. may be incorporated in computer based systems meant for pattern searching/retrieval of similar sequences from sequence libraries for the prediction of gene functions. This may be done by assigning these sequence related information as a separate item of information in the sequence libraries such as EMBL, GENEBANK, etc. and prescribing specified search conditions for identifying comparable segments in any newly sequenced DNA. Such rules or constraints may be determined by examining the sequence information stored in the sequence library databases preferably using machine learning algorithms of artificial intelligence techniques. The corresponding sequence graphs may then be analyzed from the underlying patterns.

The technique can, in principle, be extended with appropriate parameters to cover multidimensional sequences, such as protein sequences and define distance measures that can be used to correlate close sequences. Although preliminary results in these applications are still under process, it is evident that several problems regarding choice of axes systems and multivariate analysis need to be resolved before a consensus on the extent of utility of such an approach can be reached. We expect to cover such extensions at a later date.

Thus, the foregoing results on analysis of gene sequences indicate that the proposed indices are useful for quantifying differences in base distribution of gene sequences and effective in discriminating closely related sequences (and graphs) arising from evolutionary divergence, DNA polymorphism, etc. Considering all these it seems plausible that the similarity indices may find application in retrieving

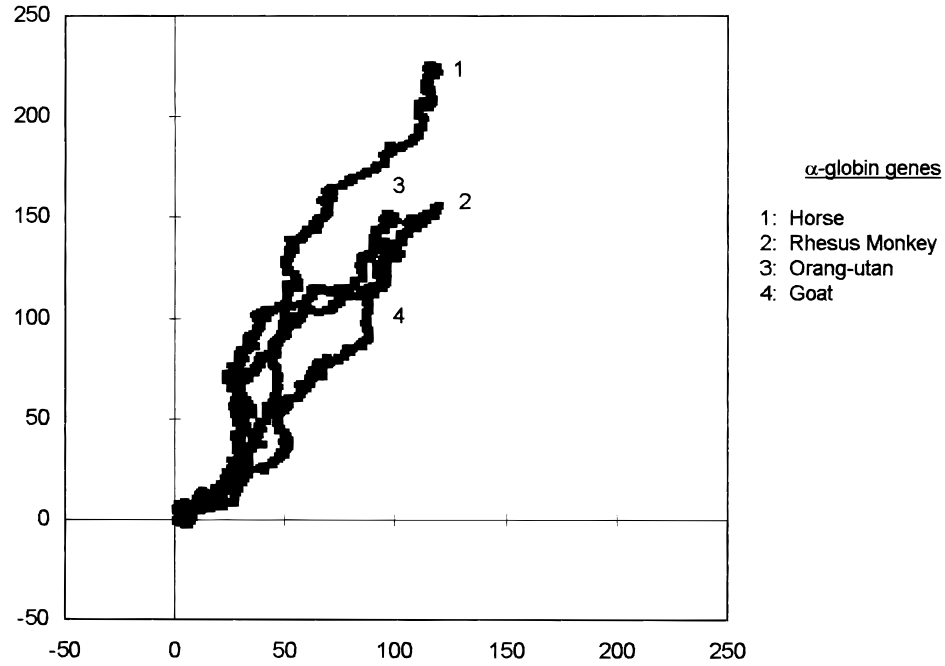

Figure 1. Superposition of graphs of the α-globin genes of horse, rhesus monkey, oranguntan, and goat.

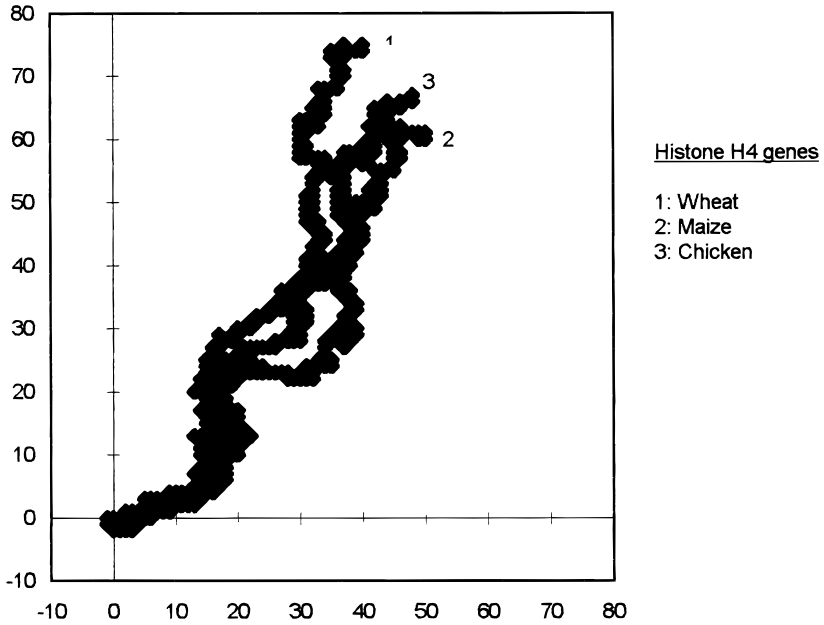

Figure 2. Superposition of graphs of histone H4 genes of wheat, maize, and chicken. Graphs of mouse and human not shown due to high degree of overlap.

similar sequences from a sequence library and also in determining functions of unknown sequences from comparative studies with their values for the known sequences side by side with other similarity measures.[7,8] Unlike quantifying a sequence on the basis of discrete segments as in other similarity measures, the present approach deals with a sequence in its entirety. Such an approach complements existing quantitative methods and would be useful in bioinformatics research with genomic sequences.

4. SUMMARY

These results show that the calculation technique outlined in this paper provides a reasonable method to index gene sequences and through a distance measure correlate with evolutionary changes. While much more work needs to be done to forge this technique into a precision tool for characterization of gene sequences, preliminary results appear encouraging in establishing a quantitative basis to visual cues in graphical representations on differences between closely related gene sequences. The indices developed here can, along with other parameters, be incorporated in sequence databases such as the EMBL to provide tools for searching gene banks for better retrieval of required sequences. The possibility exists also of extending this methodology to the case of protein sequence data to index such sequences and calculate distances of homologous genetic and protein sequences with a view to eventually determining phylogenetic trees on molecular basis.

ACKNOWLEDGMENT

This work was done under the project grant SP/SO/D31/94 of the Department of Science and Technology, Government of India, which financial assistance is gratefully acknowledged.

REFERENCES AND NOTES

(1) Hamori, Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Bio. Chem.* **1983**, *258*, 1318−1327.
(2) Gates, M. A. A simple way to look at DNA. *J. Theor. Biol.* **1986**, *119*, 319−328.
(3) Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Res.* **1990**, *18*, 2163−2170.
(4) Nandy, A. A new graphical representation and analysis of DNA sequence structure I. Methodology and application to globin genes. *Curr. Sci.* **1994**, *66*, 309−313.
(5) Leong, P. M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Applic. Biosci.* **1995**, *11*, 503−507.
(6) Nandy, A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci.* **1996**, *70*, 661−668.
(7) Baranidharan, S.; Sankaranarayanan, B.; Brahmachari, S. K. Chaos game representation of similarities and differences between genomic sequences. *Int. J. Genome Res.* **1994**, *1*, 309−319.
(8) Blaisdell, B. E.; Campbell, A. M.; Karlin, S. Similarities and dissimilarities of phage genomes. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5854−5859.

CI980077V