Novel Method for Discrimination of Conserved Genes through Numerical Characterization of DNA Sequences

Ashesh Nandy,*

[1] Environmental Programme in Science, Jadavpur University, Kolkata 700068, INDIA

Received xxx; Preprint published xxx; Accepted xxx ; Published xxx

Internet Electron. J. Mol. Des. **2003**, *1*, 000–000

Abstract

Motivation. One of the major motivations in developing a model to characterize gene sequences numerically is to enable a clear distinction to emerge between different sequences on the basis of base distribution patterns. While a number of different approaches have been developed to date on numerically characterizing DNA sequences, resolution of the methods to practical results still remain a formidable problem. This author's approach using the 2D graphical representation technique for such sequences provide one possibility of using DNA numerical descriptors to discriminate amongst gene sequences. In this paper we apply the method and describe preliminary results to show such discrimination between some conserved gene sequences.

Method. We have used the 2D graphical representation system for DNA sequences and the DNA graph descriptors method to calculate the normalized mean moments for the coding segments of several genes. The mean moments about the x- and y-axes of the different genes are then plotted on a 2D grid to examine their scattering and possible groupings for gene categories..

Results. This paper presents preliminary results that show that the selected set of DNA numerical descriptors is able to discriminate between conserved mammalian gene sequences.

Conclusions. The preliminary results of the DNA descriptor and numerical characterization methods to discriminate between gene sequences described in this paper shows that in principle it might be possible and useful to identify functions of newly identified sequences. For this we need to build up a large set of descriptor tables for the known genes.

Keywords. Genes discrimination, numerical characterization of DNA sequences, 2D graphical methods, DNA descriptors

1 INTRODUCTION

The rapid growth of data in the DNA sequence databases have led to intensive research to determine different ways to identify new gene sequences and functions. Quantitative methods represent an important technique in this quest and many schemes have been proposed to numerically characterize DNA sequences in the hope that such characterizations will pave the way for rapid selection and identification of coding sequences. Raychaudhury and Nandy [1] employed quantitative techniques from 2D-graphical representation of DNA sequences [2] to develop DNA descriptors and showed that the resulting numbers tallied well for the species considered. Several refinements have been made on the 2D representation method [3-6] and the difficulty with its degeneracy feature has been shown to be restrictive in only a small number of cases [7]. The importance of the problem of numerical characterization of DNA sequences has led to many other attempts using matrix invariants, compact representations, 3D displays, characteristic sequences,

* Correspondence author; phone: 91 (33) 2414-6110 (Univ) 2473-0577 (Res); E-mail: anandy43@yahoo.com

and other methods [8-14]. However, the computations remain a formidable problem; in most instances results have been published for small segments of gene sequences.

While a reasonably good and unique scheme to characterize a DNA sequence segment remains to be developed, applications of the 2D graphical representation technique to identify new genes in human chromosomal sequences have yielded good results [15]. However, the need to identify possible gene candidates for their functions remains a predominant and important problem. In this paper we address the problem of identification of new genes using computational methods with existing libraries of categories of gene sequences and show that an application of the DNA descriptors arising out of the 2D graphical representation model provides one method to differentiate between various conserved gene sequences. This can lead us to explore the possibility of applying such techniques for function identification of newly discovered genes.

2 MATERIALS AND METHODS

In the graphical representation method, a DNA sequence is represented as a series of points in a 2-dimensional Cartesian plot using the following algorithm [2]: we move one step to the left in case the base is adenine, one step up for cytosine, one step to the right for guanine, and one step down for thymine. This draws a running plot with the instantaneous difference between the guanine and adenine residues along the x-axis and that between cytosine and thymine along the y-axis. Two other orthogonal systems are possible depending on the association of the bases with the cardinal directions but we consider here only the system described above, viz. the ACGT association going clockwise from the negative x-axis. The cumulative effect of this representation scheme is a graph of the sequence that is characteristic of the local and global base distribution in the sequence.

For the numerical descriptors of the graphical representation of a gene sequence, we proceed as in our Ref [1] and define the normalized mean moments $\overline{\mu_x}, \overline{\mu_y}$ about the two axes as

$$\overline{\mu_x} = \sum_i x_i / N, \quad \overline{\mu_y} = \sum_i y_i / N$$

where the x_i and the y_i represent the x- and y-co-ordinates of the representative points of each base on the graph and the sum runs over the total sequence length *(N)* represented. We have shown previously [1,16,17] that these numerical descriptors and their derivatives are useful parameters to characterize and compare DNA sequences.

In this paper we compute the mean moments for the coding sequences of the alpha globin, beta globin and histone H4 genes of several mammalian species. All data are taken from the EMBL DNA database. For ease of readability, we omit the bars on the mean moments henceforth in all references to them.

3 RESULTS AND DISCUSSION

The data for the mean moments of the coding sequences of the α-globin, β-globin and histone H4 were calculated from the extracts of the sequence data from the EMBL DNA database. For our purpose only the CDS sequences of the α- and β-globin genes were used in order to relate closely to the conserved sequences. The results of the calculations of the mean moments are given in Table 1. The data are grouped by gene types: Histone H4, α-globin and β-globin. Column 1 lists the species common name, column 2 lists the EMBL ID of the corresponding DNA sequence and the moments μ_x and μ_y (representing the barred quantities defined earlier now without the bars) are given in columns 3 and 4.

Table 1: Normalized Mean Moments of the Coding Regions of three conserved Genes

Species	**EMBL ID**	**Normalized Mean Moments**	
		μ_x	μ_y
HISTONE H4			
MOUSE	MMHIST4	17.86218	19.24678
RAT	RR4HIS	21.23397	22.87819
HUMAN	HSHIS	17.85209	17.55627
HUMAN	HSHISAD	12.34296	9.22757
ALPHA GLOBINS - Exons			
HORSE	ECHBA22	23.01630	38.12354
GOAT	CHHBAI	26.00936	33.03265
RH.MONKEY	MMHBA	31.01166	36.42191
MOUSE	MMAGL1	15.39610	14.61502
RABBIT	OCHBAPT	12.93940	36.66670
ORANGUTAN	PPHBA02	23.42657	40.96742
BETA GLOBINS - Exons			
Human Betaglobin	HSHBB	31.78600	-0.26351
Mouse	MMBGL1	19.52833	2.00680
Rat	RNGLB	16.73923	-4.27438
Goat	CHHBBAA	28.28989	-4.29680
Oppossum	DVHBBB	16.86936	-9.89639
Lemur	LMHBB	28.94600	-6.10800
Chimp	OCBGLO	27.73000	-5.58300

A two-dimensional plot of μ_x vs μ_y shows three distinct regions for the three gene types except for the mouse gene data for alpha globin. Mouse DNA sequence has many genes closely related to the human genome but homologies are often low. In the 2D graphical representation of the mouse alpha globin gene the plot turns out to be quite compact compared to the other species and this is reflected in the outlier effect of the point on our discrimination chart. A similar, although less acute, outlier effect is seen with the mouse data in the case the beta-globin gene also. If we omit the mouse data point from the beta-globin moment region, the area outlined in the map would be significantly smaller and show better resolution.

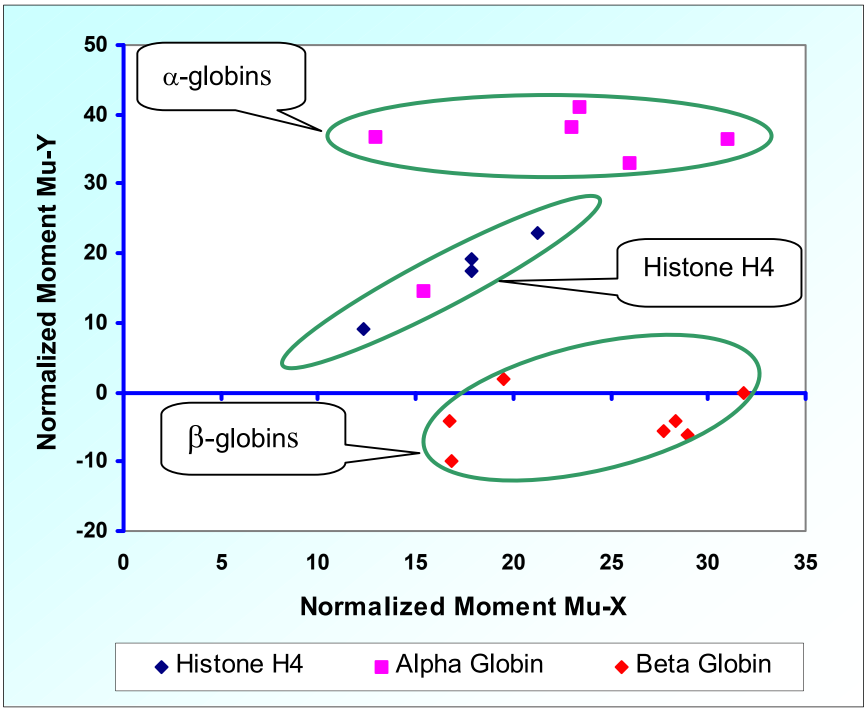

Fig.1 The characteristic map of the 2D gene descriptors of α-globin, β-globin and histone H4 sequences. The green ellipses outline the areas of scatter of the points arising from the mean moments of the various species and genes given in Table 1.

Neglecting the alpha-globin mouse moment data, the balance points representing the different species are seen to form well-defined groups where the *intra*-genic differences are much smaller than the *inter*-genic differences. This is to be expected from the fact that similar gene sequences from different species bear close homologies and are distinctly different from other genes by virtue of the base composition and distribution patterns. This in fact has been also seen from the various schemes of numerical characterizations based on the 2D graphical respresentation methods [1], the matrix invariants methods worked out by Randic et al [11,12] and the characteristic sequences methods of He and Wang [14] in which the authors have worked out distance parameters for the gene sequences and demonstrated different degrees of correspondence with the known phylogenetic spectrum.

To provide a broader basis for our hypothesis of gene discrimination through numerical characteristics parameters of DNA sequences, we have also tested the model with plant and avian descriptors. The results are commensurate with our expectations but not as compact as is observed for the mammalian species genes. This could be because of the sparse database we have tried out to date, but could also arise from the probable dispersion within the base distribution patterns that may have arisen in the longer evolutionary time span for these kingdoms as compared to mammals. Further calculations are being carried out in this area and will be reported in due course.

4 CONCLUSIONS

We have shown in this brief note that the numerical descriptors defined for DNA sequences in a 2D graphical representation model can be used to discriminate between some conserved mammalian gene sequences. Numerical estimates derived from 2D methods have been found to be quite sensitive to small variations in base distributions [16,17] and the method has also been used in gene discovery in newly sequenced segments of the human genome [15]. An important aspect of such gene discovery is also to determine its probable function and homology with known gene sequences. In this context computational methods would be useful in a rapid scan and analysis of genome length sequences for discovery and identification of coding regions. The method presented in this paper is an attempt to utilize the power of numerical characterizations of DNA sequences to achieve this goal. While the results described here are preliminary, we believe that the method appears promising in this respect.

5 REFERENCES

[1] C. Raychaudhury and A. Nandy, Indexing scheme and similarity measures for macromolecular sequences, *J Chem Inf Comput Sci* **1999**, *39*, 243-247.

[2] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes, *Curr Sc* **1994**, *66(4),* 309-314.

[3] S. Tarafdar, P. Nandy, S. Sahoo, A. Som, J. Chakrabarti and A. Nandy, Self-similarity and scaling exponent for DNA walk model in two and four dimensions, *Indian J Phys* **1999**, *73B(2),* 337-343.

[4] X. Guo, M. Randic and S. C. Basak, A novel 2D graphical representation of DNA sequences of low degenracy, *Chem Phys Lett* **2002**, *350*, 106-112.

[5] Y. Liu, X. Guo, J. Xu, L. Pan and S. Wang, Some notes on 2-D graphical representation of DNA sequences, *J Chem Infor and Comput Sc* **2002**, *42*, 529-533.

[6] X Guo and A Nandy, Numerical Characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy, Chem. Phys Letters **369**, 361-366, 2003

[7] A Nandy and P Nandy, On the Uniqueness of Quantitative DNA Difference Descriptors in 2D Graphical Representation Models, Chem. Phys Letters **368**, 102-107, 2003

[8] M. Randic, A. Nandy and S. C. Basak, On the numerical characterisation of DNA primary sequences, *J Math Chem*, submitted.

[9] M. Randic, M. Vracko, A. Nandy and S. C. Basak, On 3-D representation of DNA primary sequences, *J Chem Infor and Comput Sc* **2000**, *40*, 1235-1244.

[10] M. Randic, On characterisation of DNA primary sequences by a condensed matrix, *Chem Phys Letters* **2000**, *317*, 29-34

[11] M. Randic, X. Guo and S. C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *J Chem Inf Comput Sci* **2001**, *41*, 619-626.

[12] M. Randic and M. Vracko, On similarity of DNA primary sequences, *J Chem Inf Comput Sci* **2000**, *40*, 599-606.

[13] P-a He and J Wang, Characteristic sequences for DNA primary sequences, *J Chem Infor and Comput Sc* **2002**, *42*, 1080-1085.

[14] P-a He and J Wang,Numerical characterization of DNA primary sequences, *Internet Electronic J. Mol. Design,* **2002**, *1*, 668-674.

[15] S. Ghosh, A. Roy, S. Adhya and A. Nandy, Identification of New Genes in Human Chromosome 3 Contig 7 by Graphical Representation Technique, Current Science **84 (12)**, 1534 – 1543, 2003 June 25.

[16] A. Nandy and S. C. Basak, A simple numerical descriptor for quantifying effect of toxic substances on DNA sequences, *J Chem Inf Comput Sci* **2000**, *40***,** 915-919.

[17] A Nandy, P Nandy and S C Basak, Quantitative Descriptor for SNP Related Gene Sequences, *Internet Electronic J. Mol. Design* **2002**, *1*, 367-373..