Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Geometrical Method of Exhibiting Similarity/ Dissimilarity under New 3D Classification Curves and Establishing Significance Difference of Different Parameters of Estimation

Subhram Das*, Jayanta Pal
Dept. of Computer Science & Engineering
Narula Institute of Technology
West Bengal, India

D. K. Bhattacharya
Dept. of Pure Mathematics
University of Calcutta
West Bengal, India

Abstract— We introduce a new 3D graphical representation of DNA sequences using three classes of classification curves defined in terms of three types of classifications of bases of nucleic acids. This is an extension of the corresponding known 2D representation under classification of curves. The new graphical representation can completely avoid loss of uniqueness in the transfer of data from a DNA sequence to its mathematical representation. The objective is to examine similarities/ dissimilarities among the coding sequences of the first exon of β-globin gene of 11 species based on geometrical descriptors, which are taken as 3D graph radius, associated average angle measure and relative departure from the DNA curve and also to check whether there exists any significant differences between the present measures and similar measures as taken up in 2D cases earlier. We record our results against those of the existing ones as derived under similar three geometrical descriptors in 2D cases separately. We establish significant difference between the descriptors based on statistical hypothesis at .05 level of significance error.

Keywords—3D Graphical representation; Similarity/ dissimilarity of DNA sequences; Computational complexity; 't' test for hypothesis testing

I. INTRODUCTION

DNA is usually presumed to be the critical macromolecular target for carcinogenesis and mutagenesis [1]. To predict sequence changes induced by different agents, it is imperative to have quantitative measures to compare and contrast the different DNA sequences [2,4]. In addition, the very rapid rise in available DNA sequence data has also made the problem more emerging and interesting too. This is why over the last few years several authors have presented various methods to assign mathematical descriptors to DNA sequences in order to quantitatively compare the sequences and determine similarities and dissimilarities amongst them. As the methods are different and there are many such methods, so it becomes necessary to compare these methods and determine which one(s), if any, is best in characterizing DNA sequences. If not, better numerical representation and characterization are to be developed.

For numerical representation of DNA sequences, there are two approaches; one is called graphical and the other one is called direct. In the former case, the DNA sequence is first embedded in a graph of finite dimension and then quantification is made with coordinates of the data points of the graph. In the latter case, sequence of natural numbers is associated with the nucleotides of DNA sequence and quantification analysis is made straightway from these natural numbers. Again the latter approach is limited [3-5], whereas the former approach is most common. In fact, researchers have outlined different graphical representations of DNA sequences. These are two dimensional, three dimensional, four dimensional and even six dimensional. Some of the different types of 2D graphical representations are listed in [2,6-12]. Some of the different types of 3D graphical representations are listed in [14-22]. 4D graphical representation is found in [23]. Other graphical representations are found in [24,25].

Again so far as numerical characterization of DNA sequences is concerned, there are mainly two types of methods. One is called matrix association method and the other one is called geometrical method. Usually all these methods are tried with the first exon of the DNA sequence of the β-globin gene in order to compare the sequences from different species for their similarities and dissimilarities. The idea behind numerical characterization of DNA sequences is to devise mathematical descriptors that would capture the essence of the base composition and distribution of the sequence in a quantitative manner, which would facilitate sequence identification and comparison of similarities and dissimilarities of sequences. Base composition provides gross information of the total content of each base in the sequence and is easily determined. Base distribution is more informative and is capable of differentiating among various genes and species, even if the base composition numbers are identical, as is the case with highly conserved genes like Histone H4 or if there are many mutational variations of viral genes. Since the sequence of a gene is almost unique in the DNA of a species, and it bears close homology with the same gene of other species, but are quite different from other genes, so it is expected that the base composition and distribution characteristics would form part of a set of descriptors, which can quantify each gene sequence.

In matrix association method a matrix is associated with the coordinates of the points of the graph, if it is a graphical representation. The matrices are of the following types D/D, L/L, M/M, and they are called distance matrices, as they are defined using Euclidean distances of points of 2D or 3D spaces. As we are interested in geometrical methods, so we are not giving the details.

In geometrical method, some geometrical invariants of the embedded curve are taken as mathematical descriptors. They may be first order moments (μ_x, μ_y) and a graph radius g_R defined for each sequence by the formulae $\mu_x = \frac{\sum x_i}{N}, \mu_y = \frac{\sum y_i}{N}, g_R = (\mu_x^2 + \mu_y^2)^{1/2}$ where (x_i, y_i) represent the co-ordinates of points on the plot and N is the total number of bases in the segment. Here g_R represents the Base Distribution index and is critically dependent on the position of each base in the sequence. Again g_R and the first order moments also enable computation of graph similarity/ dissimilarity index $\Delta g_R = [(\mu_{1x} - \mu_{2x})^2 + (\mu_{1y} - \mu_{2y})^2]^{1/2}$ where μ_1, μ_2 refer to two different DNA sequences. g_R and Δg_R have been found to be very sensitive measures of the sequence composition and distribution [2,8], the values depending on the type of mutations and the positions in the sequence. g_R is especially useful in comparing equal length sequences [13].

The descriptors may be first order moments, and standard deviation. If we consider three types of curves viz., W-S curve, where W = {A, T}, S = {G, C}; M-K curve, where M = {A, C}; K = {G, T}; R-Y curve, where R = {A,G}, Y = {C,T} for each species, then for every type of curve, each of first order moments, and standard deviation is a 2-component vector for 2D representation and a 3-component vector for a 3D representation. So for 2D and 3D representation of DNA sequence, the descriptors are taken as corresponding 6- component and 9-component vectors respectively. Sometimes graph radius, angle made by the graph radius with the x-axis, average departure of the points of the graph from the line of identity are also taken as descriptors separately [11]. To sum up all such ways of numerically representing DNA sequences and developing descriptors for quantitatively analyzing DNA sequences, we may make the following remarks:

All methods that require plotting systems in four dimensions and above are difficult to visualize;

2D and 3D graphical representations have additional advantage over other dimensional ones in the sense that they allow visual inspection of data, which helps in recognizing major differences among similar DNA sequences. Again 2D methods that do not exclude repetitive walks necessarily lose some amount of information. So researchers tried for non-degeneracy or at least for lower degeneracy of 2D representation [26]. Of course those 2D representations that do completely meet requirements of non-degeneracy have not yet been used to demonstrate any identifiable and useful visual clues to DNA or gene properties. Nevertheless it is obvious that Quantification analysis based on non-degenerate representation is expected to give better results compared to the corresponding degenerate ones. Quantification analysis based on 3D representation is expected to give still better results if they are non-degenerate and so researchers pursued with such 3D plots also [14-22]. Lastly it is noted that even now comparison of a specific exon of DNA sequences of different genes is not fully known. This is why the representation of a particular exon of DNA sequence and the process of finding out its proper descriptor is still open. We have considered 3D generalization of a 2D representation [given in Section II] of DNA sequence; this is non-degenerate. We have used three descriptors for proper analysis. Lastly by statistical hypothesis testing we have established significant difference between our descriptors and the corresponding descriptors used earlier.

II. A KNOWN 2D REPRESENTATION OF DNA SEQUENCES UNDER CLASSIFICATION CURVES [11]

A. Outline of the classification curve

In DNA primary sequences, the four bases A (adenine), C (cytosine), G (thymine) and T (guanine) can be divided into two classes according to the strength of the hydrogen bond, i.e. weak H-bonds WZ{A, T} and strong H-bonds SZ{G, C}. The bases can also be divided into another two classes, amino group MZ{A, C} and keto group KZ{G, T}. Besides, the division can also be made according to their chemical structures, i.e. purine RZ{A, G} and pyrimidine YZ{C, T}. Based on the above three classifications of bases of nucleotides, a 2D graphical representation of DNA sequences is made [33-38]. It consists of three characteristic curves. For example, for W–S curve on the Cartesian coordinate system, G and C are assigned to Cx, A and T to Cy, while the corresponding curves extend in the first quadrant. Similar rules are also followed for describing other curves.

The 2D represented points on the curves are given by $P_i = (x_i, y_i)$ where $x_i = G_i + C_i, y_i = A_i + T_i$; $x_i = G_i + T_i, y_i = A_i + C_i$; $x_i = C_i + T_i, y_i = A_i + G_i$; for W-S curve, W = {A, T}, S = {G, C}; M-K curve, M = {A, C}; K = {G, T}; R-Y curve, R = {A, G}, Y = {C, T}respectively. Further G_i, C_i, A_i, T_i are the cumulative occurrence numbers of G, C, A and T respectively, in the subsequence from the first base to the ith base in the sequence.

The main difference between this method and other existing ones [3,4,5,31,32] are that the coordinates in this method reflect the summation of the cumulative occurrence numbers of some bases in the subsequence from the first base to the ith base in the sequence, while the coordinates in other methods reflect the difference between the cumulative

occurrence numbers of some bases, which cause degeneracy in the curves representing DNA sequences. This is evident from the following curves. For example for Gates [3], for Nandy [4] and for Leong [5],

2D curves are respectively given by $x_i = C_i - G_i, y_i = T_i - A_i$, $x_i = G_i - A_i, y_i = C_i - T_i$, and $x_i = A_i - C_i, y_i = T_i - G_i$.

For Randiac [15] and Zhang [29], the 3D curves are respectively given by

$x_i = (A_i + T_i) - (G_i + C_i)$ $x_i = (A_i + G_i) - (C_i + T_i)$

$y_i = (G_i + T_i) - (A_i + C_i)$ $y_i = (A_i + C_i) - (G_i + T_i)$

$z_i = (C_i + T_i) - (A_i + G_i)$ $z_i = (A_i + T_i) - (G_i + C_i)$

As the coordinates in the present method reflect the summation of the cumulative occurrence of numbers, so it is claimed that this causes the non-generacy in the curves representing DNA sequences. But if we look very carefully then its non-degeneracy may be questioned. We consider the following table, which shows the first exon of β-globin gene of eleven different species.

TABLE 1 THE CODING SEQUENCES OF THE FIRST EXON OF B-GLOBIN GENE OF ELEVEN DIFFERENT SPECIES

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACG TGGATTA AGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACG TGGATGA AGTTGGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAA
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACG TGGATGA AGTTGGTGGTGAGGCCCTGGGCAGG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACC CTGATAA TG TTGGCGCTGAGGCCCTGGGCAG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATG TGGAAGA AGTTGGTGGTGAGGCCCTGGGC
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACC CCGATGAA GTTGGTGGTGAGGCCCTGGGCAGG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACC TCTCTGTGGGGCAAGGTGGATGTAGAGAAA GTTGGTGGCGAGGCCTTGGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACC GGCCTCTGGGGCAAGGTCAATGTGGCCGA ATGTGGGGCCGAAGCCCTGGCCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCAC TACCATCTGGTCTAAGGTGCAGGTTGACCA GACTGGTGGTGAGGCCCTTGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATG AAGTTGG TGCTGAGGCCCTGGGCAG

B. Limitation of [11] and necessary 3D generalization ATGGTGCACCTGACTCCTGA

To show the degeneracy of the method [11], we consider 2D graphical representation of the Sequence S = ATGGTGCACCTGACTCCTGA of the first 20 nucleotides of first exon of B-globin gene of Human as taken up in [11]. For W-S curve, the rule is $\phi(s_i) = (G_i + C_i, A_i + T_i), S = s_1 s_2 s_3 ... s_n$, Obviously the points of representation are sequentially (0+0, 1+0) = (0,1), (0+0,1+1) = (0,2),(0+1,1+1)=(1,2), (1+1,1+1) = (2,2), (1+1,1+2) =(2,3) etc.

Now let us see what happens if we change the first five nucleotides as TAGGA and keep others unaltered? It is seen that the same points of representation are obtained for a different sequence also. This is true for change of other parts

also. The reason is that it is the sum which matters, and not the individuals. The same observation may be made for other types of curves. So we note that the aforesaid 2D representation is degenerate, although apparently it looks non-degenerate.

3D generalization

We note that the degeneracy could be avoided by considering the frequencies of the nucleotides and putting them in the third coordinates of each point. For example for W-S 3D curve of S = ATGGTGCACCTGACTCCTGA, the pints of representations are (0,1,1),(0,2,1),(1,2,1), (2,2,2),(2,3,2) and so on. Obviously it is the inclusion of the frequency, which will make the 3D representation non-degenerate.

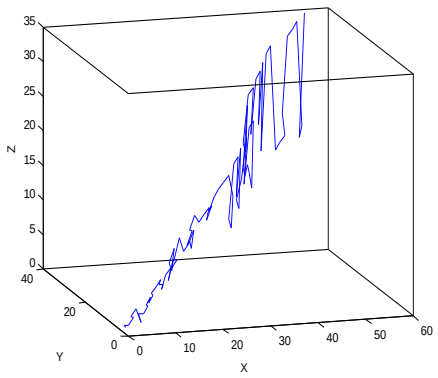

Fig. 1. 3D Graphical representation of classification curve of the DNA sequences of the first exon of β-globin gene of Human (W-S Curve)

Let xij be the points on the curves (W-S, M-K & R-Y) where j=1,2..Ni, i=1,2,3.

III. NUMERICAL CHARACTERIZATION OF DNA SEQUENCES

3-component Descriptors

A. Graphical Radius

$$g_{R1} = (\mu_{x1} + \mu_{y1} + \mu_{z1}) / N_1$$

$$\text{where } \mu_{x1} = \sum_{j=1}^{N_1} x_{1j}, \mu_{y1} = \sum_{j=1}^{N_1} y_{1j}, \mu_{z1} = \sum_{j=1}^{N_1} z_{1j},$$

$$g_{R2} = (\mu_{x2} + \mu_{y2} + \mu_{z2}) / N_2$$

$$\text{where } \mu_{x2} = \sum_{j=1}^{N_2} x_{2j}, \mu_{y2} = \sum_{j=1}^{N_2} y_{2j}, \mu_{z2} = \sum_{j=1}^{N_2} z_{2j},$$

$$g_{R3} = (\mu_{x3} + \mu_{y3} + \mu_{z3}) / N_3$$

$$\text{where } \mu_{x3} = \sum_{j=1}^{N_3} x_{3j}, \mu_{y3} = \sum_{j=1}^{N_3} y_{3j}, \mu_{z3} = \sum_{j=1}^{N_3} z_{3j}$$

B. Average Angle Measure

$$\theta_{A1} = (\frac{\mu_{x1}}{g_{R1}}, \frac{\mu_{y1}}{g_{R1}}, \frac{\mu_{z1}}{g_{R1}})$$

$$\theta_{A2} = (\frac{\mu_{x2}}{g_{R2}}, \frac{\mu_{y2}}{g_{R2}}, \frac{\mu_{z2}}{g_{R2}})$$

$$\theta_{A3} = (\frac{\mu_{x3}}{g_{R3}}, \frac{\mu_{y3}}{g_{R3}}, \frac{\mu_{z3}}{g_{R3}})$$

$$\Phi = \frac{\frac{\mu_{x1}}{g_{R1}} + \frac{\mu_{y1}}{g_{R1}} + \frac{\mu_{z1}}{g_{R1}}}{3}, \Psi = \frac{\frac{\mu_{x2}}{g_{R2}} + \frac{\mu_{y2}}{g_{R2}} + \frac{\mu_{z2}}{g_{R2}}}{3},$$

$$Ж = \frac{\frac{\mu_{x3}}{g_{R3}} + \frac{\mu_{y3}}{g_{R3}} + \frac{\mu_{z3}}{g_{R3}}}{3}$$

C. Relative Departure

$$\rho_i = \frac{1}{N_i\sqrt{3}}\left[3\sum_{j=1}^{N_i}(x_{ij}{}^2 + y_{ij}{}^2 + z_{ij}{}^2) - \sum_{j=1}^{N_i}(x_{ij} + y_{ij} + z_{ij})^2\right]^{1/2}$$

$$i = 1, 2, 3$$

TABLE 2 THE GRAPH RADIUSES ASSOCIATED WITH THREE DIFFERENT PATTERNS OF THE CLASSIFICATION CURVES

Curves	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
W-S	35.6785	33.7047	35.1024	35.9690	35.4381	36.1247	35.0499	35.2776	36.2237	33.5909	40.8063
M-K	35.7358	33.6760	35.1788	35.4941	36.2615	36.6389	35.6112	35.6181	36.2300	33.7118	40.9014
R-Y	35.4593	33.7021	35.1715	35.7592	35.5857	36.0631	35.1775	35.5485	35.9759	33.6164	40.5463

TABLE 3 THE AVERAGE ANGLES ASSOCIATED WITH THREE DIFFERENT PATTERNS OF THE CLASSIFICATION CURVES

Curves	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
W-S	0.5567	0.5554	0.5581	0.5523	0.5617	0.5594	0.5597	0.5598	0.5568	0.5574	0.5559
M-K	0.5594	0.5559	0.5569	0.5597	0.5490	0.5516	0.5509	0.5545	0.5567	0.5554	0.5546
R-Y	0.5601	0.5555	0.5570	0.5555	0.5594	0.5604	0.5577	0.5556	0.5606	0.5570	0.5595

TABLE 4 THE RELATIVE DEPARTURES ASSOCIATED WITH THREE CLASSIFICATION CURVES

Curves	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
W-S	0.1570	0.1664	0.1578	0.2130	0.1154	0.1544	0.2916	0.1386	0.1668	0.1519	0.1453
M-K	0.1821	0.1664	0.1693	0.1161	0.2087	0.1881	0.3237	0.1686	0.1801	0.1811	0.1906
R-Y	0.1343	0.1664	0.1629	0.1514	0.1570	0.1280	0.2987	0.1578	0.1413	0.1664	0.1453

TABLE 5 SIMILARITY/DISSIMILARITY MATRIX BASED ON THE EUCLIDEAN DISTANCES BETWEEN THE END POINTS OF THE 3-COMPONENT VECTORS OF THE NORMALIZED GRAPH RADIUSES

	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.0083	0.0093	0.0052	0.0064	0.0042	0.0092	0.0046	0.0025	0.0071	0.0016
Goat		0	0.0169	0.0067	0.0088	0.0114	0.0049	0.0110	0.0059	0.0017	0.0069
Opossum			0	0.0119	0.0131	0.0080	0.0177	0.0066	0.0116	0.0158	0.0108
Gallus				0	0.0103	0.0092	0.0102	0.0080	0.0044	0.0066	0.0051
Lemur					0	0.0055	0.0061	0.0072	0.0063	0.0071	0.0058
Mouse						0	0.0106	0.0039	0.0061	0.0098	0.0050
Rabbit							0	0.0113	0.0073	0.0039	0.0078
Rat								0	0.0065	0.0097	0.0057
Gorilla									0	0.0048	0.0011
Bovine										0	0.0057

Chimpanzee											0

TABLE 6 SIMILARITY/DISSIMILARITY MATRIX BASED ON THE EUCLIDEAN DISTANCES BETWEEN THE END POINTS OF THE 3-COMPONENT VECTORS OF THE AVERAGE ANGLES

	Human	**Goat**	**Opossum**	**Gallus**	**Lemur**	**Mouse**	**Rabbit**	**Rat**	**Gorilla**	**Bovine**	**Chimpanzee**
Human	0	0.00067	0.00005	0.00007	0.00013	0.00027	0.00020	0.00008	0.00013	0.00069	0.00134
Goat		0	0.00070	0.00073	0.00072	0.00093	0.00049	0.00071	0.00080	0.00003	0.00201
Opossum			0	0.00007	0.00010	0.00024	0.00022	0.00004	0.00011	0.00072	0.00131
Gallus				0	0.00016	0.00023	0.00027	0.00010	0.00010	0.00075	0.00128
Lemur					0	0.00022	0.00022	0.00008	0.00013	0.00074	0.00130
Mouse						0	0.00044	0.00022	0.00014	0.00095	0.00108
Rabbit							0	0.00023	0.00032	0.00051	0.00152
Rat								0	0.00011	0.00073	0.00130
Gorilla									0	0.00082	0.00121
Bovine										0	0.00203
Chimpanzee											0

TABLE 7 SIMILARITY/DISSIMILARITY MATRIX BASED ON THE EUCLIDEAN DISTANCES BETWEEN THE END POINTS OF THE 3-COMPONENT VECTORS OF THE RELATIVE DEPARTURES

	Human	**Goat**	**Opossum**	**Gallus**	**Lemur**	**Mouse**	**Rabbit**	**Rat**	**Gorilla**	**Bovine**	**Chimpanzee**
Human	0	0.00053	0.00034	0.00089	0.00059	0.00012	0.00290	0.00036	0.00011	0.00049	0.00037
Goat		0	0.00029	0.00083	0.00079	0.00065	0.00253	0.00049	0.00044	0.00024	0.00078
Opossum			0	0.00084	0.00063	0.00045	0.00279	0.00022	0.00028	0.00032	0.00051
Gallus				0	0.00146	0.00104	0.00302	0.00099	0.00086	0.00105	0.00112
Lemur					0	0.00058	0.00288	0.00050	0.00066	0.00058	0.00057
Mouse						0	0.00299	0.00041	0.00023	0.00060	0.00032
Rabbit							0	0.00295	0.00284	0.00251	0.00322
Rat								0	0.00036	0.00044	0.00035
Gorilla									0	0.00045	0.00045
Bovine										0	0.00073
Chimpanzee											0

IV. RESULTS ON COMPARISON

The similarity/dissimilarity of the coding sequences of the first exon of the human β-globin gene based on the Euclidean distances between the end points of (A) the 3-component vectors of normalized graph radiuses, (B) the 3-component vectors of angles, (C) the 3-component vectors of normalized relative departures are given in the following

Table 8 The similarity/dissimilarity of the coding sequences of the first exon of the human β-globin gene based on the Euclidean distances between the end points of (A) the 3-component vectors of normalized graph radiuses, (B) the 3-component vectors of angles, (C) the 3-component vectors of normalized relative departures

	A	B	C	A1	B1	C1
Human	0.00827	0.00067	0.00053	0.00567	0.2896	0.0259
Opossum	0.01685	0.0007	0.00029	0.00877	0.4667	0.0485
Gallus	0.00667	0.00073	0.00083	0.00462	0.0235	0.1558
Lemur	0.00881	0.00072	0.00079	0.00822	0.3671	0.0401
Mouse	0.01136	0.00093	0.00065	0.00664	0.2902	0.0315
Rabbit	0.00489	0.00049	0.00253	0.00529	0.1369	0.0212
Rat	0.01102	0.00071	0.00049	0.00634	0.2726	0.036
Gorilla	0.00595	0.0008	0.00044	0.00514	0.254	0.0215
Bovine	0.00171	0.00003	0.00024	0.00191	0.0678	0.008
Chimpanzee	0.00691	0.00201	0.00078	0.00509	0.2189	0.0185

A1 corresponds to 3-component graph radius of [27]
B1 corresponds to 3-component vector angles of [28]
C1 corresponds to 3-component relative departures [30]

D. Comparison of our different measures with the corresponding earlier measures:

Basic Problem:

We have taken only 11 species for the purpose of comparing their β-globin part. But in reality the number of such species is always more. So a fundamental problem is to see whether we can make some interpretation on the whole populations of such β-globin part based on the analysis of this sample of 11 species only. Obviously when we compare two column vectors of, A and A1, say, each of size 11, we look for equality of means of their populations. This leads to the following hypothesis testing under .05 level of significance error:

H0: There is no significance difference between the means of the populations of A and A1.

H1: There is a significance difference between the means of the populations of A and A1.

Again H0 is effective, if the calculated value of't' statistics is less than the prescribed value, and H1 is effective, if the calculated value of 't' statistics exceeds the prescribed value.

(1) Now $t = \dfrac{\overline{x} - \overline{y}}{S.E.(\overline{x} - \overline{y})}$, where $\overline{x}, \overline{y}$ are the sample means and $S.E.(\overline{x} - \overline{y})$ represents the standard error of $(\overline{x} - \overline{y})$. Again $S.E.(\overline{x} - \overline{y})$ for samples of sizes n1 and n2 is determined by two formulae, according as (i) the sample variances are assumed to be the same and (ii) the sample variances cannot be assumed to be the same. Again the condition of equality of sample variances is determined by F-tests (Fisher's F-statistics). In this case also two hypothesis are taken, viz., H0: There is no difference between the sample variances and HA: There is a difference between the sample variances. F-statistics is given by $F = s_1^2 / s_2^2$ where $s_1^2 = \dfrac{n_1}{n_1 - 1} S_1^2$ and $s_2^2 = \dfrac{n_2}{n_2 - 1} S_2^2$ are the unbiased estimators of population variances; S_1^2, S_2^2 are the sample variances. Value of F is to be compared with the prescribed value at (n1-1, n2 -1) degrees of freedom. Now if the value of the F statistics is less than the prescribed value, H0 holds, otherwise HA holds. Thus first of all by F-test, it is to be decided, which of the above two cases (i) or (ii) is to be considered. Accordingly the following values of $S.E.(\overline{x} - \overline{y})$ are to be taken,

(i) $S.E(\overline{x} - \overline{y}) = s_p \sqrt{(\dfrac{1}{n_1} + \dfrac{1}{n_2})}$

S_1^2, S_2^2 are the variances of n1 and n2 respectively, and s_p^2 is the pulled variance given by $s_p^2 = \dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$

(ii) $S.E(\overline{x} - \overline{y}) = \sqrt{(s_1^2 / n_1) + (s_2^2 / n_2)}$

In the two cases the value of 't' is to be compared with standard values at degrees of freedom given by $\nu = n_1 + n_2 - 2$ and

$$\nu' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{((s_1^2/n_1)^2/(n_1-1)) + ((s_2^2/n_2)^2/(n_2-1))}$$

respectively. In the latter case (ii), the 't' test is called Welch's approximation 't' test [39].

For comparison of A with A1 as given in (TABLE 8) by 't' test, we first apply F statistics for equality of population variances. We have F=4.647, this value is greater than the standard value of $F_{.05,9,9} = 3.18$. So, we take $S.E(\overline{x} - \overline{y}) = .\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$. Now we calculate the value of t given by $t = \frac{\overline{x} - \overline{y}}{S.E.(\overline{x} - \overline{y})}$ at degree of freedom

$$\nu' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{((s_1^2/n_1)^2/(n_1-1)) + ((s_2^2/n_2)^2/(n_2-1))}$$

On calculation the value of ν' is found to be 0.0000000034. As it is not an integer so we take the next least integral for ν' . This gives us $\nu' = 1$.

The value of 't' is found to be t=443.53. As this value of 't' is greater than the prescribed value of $t_{(2),.05,\nu'} = 6.31$, so the results of A and A1 significantly differ. Similarly we have proved that B and B1, C and C1 also differ significantly.

V. CONCLUSIONS

In this paper, based on the three DNA type bases, we outlined a 3D graphical representation of DNA sequences based on three types of classification curves, and presented a variant of a mathematical representation for DNA sequences in 3D graphs. Some advantages of classification curves are as follows:

1. The distributions of bases of different types are strictly displayed in these three characteristic curves of the corresponding DNA sequence.

2. In comparison to previous works that used combinations of sums and subtractions, and even that, which involves only sum [11], it properly eliminates plot degeneracy and can completely avoid loss of information in the transfer of data from a DNA sequence to its mathematical representation.

3. The three characteristics used are graph radiuses, angles and relative departures, which are extracted from the 3D graphs and then used to calculate distance values among sequences. Comparison of the results of the examination of similarities/dissimilarities among the coding sequences of the first exon of β-globin gene of 11 species with other geometrical methods illustrates the utility of the approach.

4. Statistical analysis shows that although our measures are similarly obtained as in previous cases, but our measures are significantly different from the earlier ones. This proves the necessity of introducing such new measures.

5. Lastly the method is sound and one can find that the computational complexity is only O(N); this greatly reduces the computational complexity.

REFERENCES

[1] L. Rhomberg, V. L. Dellarco, W. H. Farland, and R. S. Cortesi, The Significance of DNA Damage and Repair Mechanisms in Health Risk Assessment, in: *DNA Damage and Repair in Human Tissues*, Eds. M. Betsy Sutherland and D. Avril Woodhead, Plenum Press, New York 1990, pp 225-232.

[2] A. Nandy, P. Nandy, and S. C. Basak, Quantitative Descriptor for SNP Related Gene Sequences, *Internet Electronic Journal of Molecular Design*.2002, *1*, 367-37

[3] Gates, M. A. *J. Theor. Biol.* 1986,*119*, 319.

[4] Nandy, A. *Current Science* 1994, *66*, 309.

[5] Leong and Morgenthaler, *Comput. Appl. Biosci.* 1995, *11*, 503.

[6] Nandy, A. *Comput. Appl. Biosci.* 1996, *12*, 55.

[7] Nandy, A. Internet Electron. J. Mol. Des. 2002, 1, 545.

[8] Raychaudhury, C.; Nandy, A. *J. Chem. Inf. Comput. Sci*. 1999, *39*, 243.

[9] Nandy, A.; Basak, S.C. *J. Chem. Inf. Comput. Sci*. 2000, *40*, 915.

[10] Wu, Y.; Liew, A. W.; Yan, H.; Yang, M. *Chem. Phys. Lett.* 2003, *367*, 170.

[11] Yao, Y.; Nan, X.; Wang, T. *J. Mol. Struct. (Theochem)* 2006, *764*, 101.

[12] Ghosh, S.; Roy, A.; Adhya, S.; Nandy, A. *Current Science* 2003, *84*, 1534.

[13] Nandy, A.; Nandy, P. Chem. Phys. Lett. 2003, 368, 102.

[14] Hamori, E.; Ruskin, J. J.Biol.Chem. 1983, 258, 1318.

[15] Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. J. Chem. Inf. Comput. Sci. 2000, 40, 1235.

[16] Li, C.; Wang, J. Combinatorial Chemistry & High Throughput Screening 2004, 7, 23.

[17] Yao, Y.; Nan, X.; Wang, T. Chem. Phys. Lett. 2005, 411, 248.

[18] Yuan, C.; Liao, B.; Wang, T. Chem. Phys. Lett. 2003, 379, 412.

[19] Liao, B.; Wang, T. J. Mol. Struct. (Theochem) 2004, 681, 209.

[20] Liao, B.; Zhang, Y.; Ding, K.; Wang, T. J. Mol. Struct. 2005, 717, 199.

[21] Zhu, W.; Liao, B.; Ding, K. J. Mol. Struct. 2005, 757, 193.
[22] Bai, F.; Zhu, W.; Wang, T. Chem. Phys. Lett. 2005, 408, 258.
[23] Chi, R.; Ding, K. Chem. Phys. Lett. 2005, 407, 63.
[24] Liao, B.; Wang, T. J. Chem. Inf. Comput. Sci. 2004, 44, 1666.
[25] Randic, M.; Lers, N.; Plavsic, D.; Basak, S. C.; Balaban, A. T. Chem. Phys. Lett. 2005,407,205.
[26] Xiaofeng Guo, Milan Randic, Subhas C. Basak- Chemical Physics Letters 350 (2002) 106-112
[27] Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Chem. Phys. Lett. 2003, 371, 202.
[28] Liao, B.; Wang, T. J. Comput. Chem. 2004, 25, 1364.
[29] C.T. Zhang, J. Wang and R. Zhang(2001). A novel method to calculate the G+C content of genomic DNA sequences. *J Biomol Struct Dyn.* 19(2), 333-41.
[30] B. Liao, T.M. Wang, J. Comput. Chem. Phys. Lett. 388 (204) 195.
[31] Liao, B. Chem. Phys. Lett. 2005, 401, 196.
[32] Liao, B.; Tan, M.; Ding, K. Chem. Phys. Lett. 2005, 414, 296.
[33] He, P.; Wang, J. Internet Electron. J. Mol. Des. 2002, 1, 668.
[34] Randic, M.; Vracko, M.; Lers, N.; Plavsic, D. Chem. Phys. Lett. 2003, 368, 1.
[35] Yao, Y.; Liao, B.; Wang, T. J. Mol. Struct. (Theochem) 2005, 755, 131.
[36] Li, C.; Wang, J. Combinatorial Chem. & High Throughput Screening 2003, 6, 795.
[37] Liao, B.; Ding, K. J. Comput. Chem. 2005, 26, 1519, 1523.
[38] Wang, J.; Zhang, Y. Chem. Phys. Lett. 2006, 423, 50.
[39] Welch, B.L. (1938) The Significance of the Difference Between Two Means when the Population Variances are Unequal, Biometrika, 29, 350-362.

