OPEN ACCESS Freely available online

Journal of Genetic Syndromes and Gene Therapy

Review Article

A Critical Survey of Mathematical Approaches towards Genome and Protein Sequence Comparison

Bhattacharya DK

Department of Pure Mathematics, Calcutta University,Kolkata, India

ABSTRACT

The present review highlights the very purpose of comparing genome and protein sequences and examines critically the different types of methodologies involved in the process leading to the final results of comparison.

Keywords: Numerical representation; Genome sequence; Protein's primary structure; classified groups

NUCLEOTIDES AND THEIR CLASSIFICATIONS

Genomes are sequences of 4 nucleotides: (A-adenine; C-cytosine; G- guanine; T- thymine). They are polynucleotide chains. Two genome sequences differ in the number and relative positions of nucleotides in the chain. They may also differ in their lengths.

Classification of nucleotides on the basis of Bio-chemical properties

Based on their bio-chemical properties, the 4 nucleotides are classified in three distinct groups R/Y) [Purine-Pyrimidine], (M/K) [Amino-Keto] and (W/S) [Weak-Strong H-Bonds], where R = (A, G) and Y = (C, T), M=(A, C) and K=(G, T), W=(A, T) and S=(C, G).

AMINO ACIDS AND THEIR CLASSIFICATIONS

Protein's primary structures are sequences of 20 peptides (amino acids), which are given by Alanine(A), Cysteine(C), Aspartic acid(D), Glutamic acid(E), Phenylalanine(F), Glycine(G), Histidine(H), Isoleucine(I), Lysine(K), Leucine(L), Methionine(M), Asparagine(N), Proline(P), Glutamine(Q), Arginine(R), Serine(S), Tyrosine(T), Valine(V), Tryptophan(W) and Threonine(Y). When a protein's primary structure is taken up as a sequence of amino acids, it is understood that only the backbone structures of the amino acids are taken into consideration. It is a polypeptide chain. As sequences, Protein's primary structures differ externally due to the number and relative position of the peptides in the chain. They may also differ in their lengths.

CLASSIFICATION OF AMINO ACIDS ON THE BASIS OF PHYSIO-CHEMICAL PROPERTIES

Primary Classification of Amino acids

(i) Nonpolar: Alanine (Ala), Glycine (Gly), Isoleucine (Lle), Methonine (Met), Tryptophan (Trp), Phenylalanine (Phe), Proline (Pro), Valine (Val)

In America, about 75% of computer users who worked for long hours at the computer had complaints of visual symptoms. This is expected to be worse in developing countries where fewer people are aware and take treatment, but the majority is less aware and ignorant about the condition.

In Africa, limited studies on CVS have been carried out in spite that computer use has attained a significant increased especially as technology is advanced.

In Ethiopia, the prevalence of CVS ranges from 69.5% to 73.9% . Consequently, many organizations can facilitate and manage their businesses using a computer.

It reduces the quality of life of computer users.Therefore, adjusting ergonomics is very important. Therefore, as the number of postgraduate students are increasing, studying CVS among them is indispensable.

(ii) Polar: Cysteine (Cys),Serine (Ser),Threonine (Thr), Aspargine (Asn),Glutamine (Gln)

(iii) Polar Basic (Positively Charged): Histidine (His), Lysine (Lys), Arginine (Arg)

(iv) Polar Acidic (Negatively Charged): Aspartate (Asp), Glutamate (Glu)

Correspondence to: D K Bhattacharya, Department of Pure Mathematics, Calcutta University, Kolkata, India, E-mail: dkb_math@yahoo.com

Received: June 25, 2020; **Accepted**: July 13, 2020; **Published**: July 20, 2020

Citation: Bhattacharya DK (2020) A Critical Survey of Mathematical Approaches towards Genome and Protein Sequence Comparison. J Genet Syndr Gene Ther. 11: 329. DOI: 10.4172/2157-7412.20.11.329

(B) Special classifications of amino acids used in protein sequence comparison

(i)(a) 3 group Classification [1]: Dextrorotatory E, A, I, K, V ; Levorotatory N, C, H, L, M, F, P, S;

Irrotational G, Y, R, D, Q

(b) 3 group Classification [2]: Strongly Hydrophilic R, D, E, N, Q, K, H ; Strongly Hydrophobic L, I, V, A, M, F ; others S, T, Y, W, C, G, P [derived from (ii)(a)]

(a) 4 group Classification [2]: Strongly Hydrophilic (POL) R, D, E, N, Q, K, H; strongly hydrophobic (HPO) L, I, V, A, M, F; Weakly Hydrophilic or weakly Hydrophobic (Ambiguous) Ambi S, T, Y,W; Special (none) C, G, P

(b) 4 group Classification [3]: Hydrophobic (H) Non-polar A, I, L, M, F, P, W, V; Negative polar class D, E; Uncharged polar class N , C, Q, G, S, T, Y; Positive polar class R, H, K

5 group Classification [4]: I = C, M, F, I, L, V, W, Y; A = A, T, H; G = G, P; E = D, E; K = S, N, Q, R,

(a) 6 group Biological Classification based on side chain conditions : Side chain is aliphatic G, A, V, L, I; Side chain is an organic acid D, E, N, Q; Side chain contains a sulphur M, C; Side chain is an alcohol S, T, Y; Side chain is an organic base R, K, H; Side chain is aromatic F, W, P

(iv) (b) 6 group Theoretical Classification [5]:I = I; L = L,R; A = V A, G, P, T; E = F, C, Y, Q, N, H, E, D, K; M = M,W;S = S.

Objective of Genome and Protein Sequence comparison

(a) To find proper clustering of genome/protein sequences by similarity/ dissimilarity analysis.

(b) To obtain from the clusters proper phylogeny of family of species to know their family history.

MOTIVATION OF GENOME AND PROTEIN SEQUENCE COMPARISON

As the data base of both Genome and Protein sequences is increasing rapidly; it becomes necessary to identify similarity and dissimilarity of the sequences in order to cluster them properly. Once clustering is done, it becomes clear whether data base contains a new sequence or not. At the same time, the phylogeny of the family of species has to be known from the clusters in order to know their family history. That is why phylogenetic trees are constructed from large number of sequences. It may be noted that if a new sequence is found to be similar to a known sequence, then it suggests that we need not study the second sequence separately, as there is no difference in their properties. But if the two sequences are found to be dissimilar, then it is a real problem. We are to study the new sequence separately, as the two sequences have different properties. Now from the phylogenetic tree, it may be checked to which cluster it is closest. This will give some hints about its properties from the properties of the sequences of the cluster it is closest to. This is the general scenario for comparison of genome as well as protein sequences. For protein sequence comparison, the problem is something more. As properties of proteins depend only on their tertiary structures, so apparently protein's primary structure comparison may not be of any use in this context. In fact, if we show that two proteins are similar in their primary structures, then it cannot be concluded that they are similar in their tertiary structures. So their properties may be different. Hence similarity of primary structures of two proteins is not useful. But dissimilarity of primary sequences of two proteins is very meaningful; it ensures that the two proteins also differ in their tertiary structures and therefore they differ in their properties too. Thus in this case, it is necessary to study the new protein separately.

METHODOLOGIES IN GENOME AND PROTEIN SEQUENCE COMPARISON

For comparison of both types of sequences there are two types of methods: Alignment based and alignment free. Alignment-based approaches generally give excellent results when the sequences under study are closely related so that the sequences can be reliably aligned, but when the sequences are divergent, a reliable alignment cannot be obtained and hence the applications of sequence alignment become limited. Another limitation of alignment-based approaches is that it has more computational complexity, and it is more time-consuming. Therefore alignment free methods are of current use.

Basics of alignment free methods for genome and protein sequence comparison

- To obtain Numerical Representation (crisp, probabilistic, fuzzy) in different dimensions
- To obtain proper Descriptors for comparison
- To choose proper Distance Measures of comparison
- To obtain Distance matrix [similarity/dissimilarity matrix]
- To obtain cluster and phylogeny of species from the Distance Matrix by UPGMA software

NUMERICAL, GRAPHICAL AND GRAPH THEORETICAL REPRESENTATION

There is a distinction between numerical representation and graphical representation. Numerical representation means association of one dimensional, two dimensional or higher dimensional points of Euclidean space or even complex numbers to each of the 4 nucleotides in a genome sequence and to each of the 20 amino acids in a protein sequence. In case of complex representation it is to be associated with points on the Complex Argond plane. But graphical representations are those numerical representations, by which the whole sequence may be plotted on a two dimensional or on a three dimensional curve. Necessarily the representations must be non-degenerate in the sense that corresponding to each element of the sequence, which may be a nucleotide in a genome sequence or an amino acid in a protein sequence, there is just one single two dimensional point or a three dimensional point. Graphical representations are also called geometrical representations. Graph theoretic representations are those, where a graph can be drawn out of the given points of representations. In this sense, those representations, which are only numerical, may help in graph theoretic representations.

REPRESENTATIONS FOR GENOME SEQUENCES, THEIR DESCRIPTORS AND DISTANCE MEASURES

Graphical representations

Examples of some graphical representations are 2D representations given in [6-8]. These are basically random walk of points moving along or parallel to coordinate axes. Later on these are modified in [9-12], where all the nucleotides are plotted not along the axes but always along a vector lying in one or more quadrants of the Euclidian plane. Other similar representations with slight modifications are found in [13-16].

But all such geometrical representations are not always non-degenerate. To illustrate some degenerate representations, we consider 2D representations given [17-20]. In each of the representations, the coordinates reflect the difference between the cumulative occurrence numbers of some bases, which cause degeneracy in the curves. So, this is improved in the sense that now the summation of the cumulative occurrence numbers of some bases in the subsequence from the first base to the ith base in the sequence is considered [21]. As summation is involved in place of difference, so it is claimed that the representation is non-degenerate. But, the authors obtain counter example to show that even this representation is also degenerate [22]. In this paper, they obtain corresponding non-degenerate representations under use of frequencies as the third coordinates. This is the first time the use of frequencies is found to make the graphical representation a non-degenerate one.

Anyway for such graphical representations, there are two types of descriptors- one is called geometrical descriptor, the other one is called matrix form of descriptor. In the former case, the descriptors are obtained directly from the data points of the curve. But in matrix form of descriptors, first of all some matrices are formed from the data points of the curve and then different forms of descriptors are obtained from the matrices themselves. Two dimensional geometrical descriptors are given in [21]. Three dimensional geometrical descriptors are developed [22]. In this case the distance measure used is the standard Euclidean measure. Matrix form of descriptors are of types D/D, L/L, M/M , J/J [23-29]. It is found that for all three dimensional matrix forms of descriptors, J/J is the most satisfactory one. In all these cases the distance measure is usually Euclidean. But sometimes Pearson's correlation coefficient is also used as the distance measure.

Representations, which are only numerical

(i) Real-number Representations are those where the four nucleotides bases are assigned four different real numbers arbitrarily [30].

(ii) Complex representation are those where the representation of four nucleotides in the four quadrants of the complex plane is obtained by assigning the complex numbers 1+i1, 1-i1, -1+i1, -1-i1. Naturally two nucleotides are mirror image of real axis or the mirror image of the imaginary axis [31-33].

(iii) Quaternion Representation is that, where the method of complex representation is extended to Quaternion [34]. This is also called the hyper-complex numbers representation of DNA sequences by using Quaternion of the form a+ ib+ jc+ kd where i2+ j2+ k2=1, i. j=0; j. k=0; k. i=0.

(iv) 4D Binary representation is the one, where the 4 nucleotides T, C, A, G are represented by (1,0,0,0), (0,1,0,0), (0,0,1,0) and (0,0,0,1) respectively [35]. These are four vertices of a 4-dimensional hypercube. This is a crisp representation.

Obviously for such representations geometrical forms of descriptors or matrix form of descriptors are not applicable. For 4D Binary transformation, the descriptors are obtained in from the frequency domain [36]. First of all the series under comparison are made equal by adding required number of zeroes. All these series are now taken as time series and by application of Fourier transform they are shifted in the frequency domain. There is no problem of additional zeroes, as zeroes have no contribution in the Fourier transformation. Now the descriptors are obtained by the method of Inter Coefficient Distance (ICD). The distance measure is Euclidean.

Obviously the method is applicable to sequences of equal and unequal lengths.

Another comparison of genome sequences based on 4D Binary transformation is found in paper [37]. Here also descriptors are found in the frequency domain. But the descriptors are now 12 dimensional moment vectors consisting of first order, second order and third order moments. The advantage is that it is applicable to genome sequences of equal and unequal lengths.

For complex representation this is a bit difficult problem, as we are to consider Fourier transform of a complex time series, which is not the standard one. In fact, for genome sequences such attempts are not made.

Some other representations for the whole genome sequence:

Condensed matrix representation of a genome sequence is the one where the whole genome sequence can be reduced to a 4 × 4 real matrix [38]. The descriptors are the respective 4 × 4 matrix to which it is reduced. The distance measure is Euclidean. The advantage is that it is applicable to comparison of genome sequences of equal and unequal lengths.

Fuzzy Representation is the one where each genome sequence can be represented by a 12 dimensional fuzzy vector with components lying on a 12 dimensional unit hypercube [38-40]. The idea comes from that of 4D representation [35]. Obviously due to 4D Binary representation a codon is always represented at one of the corners of 12 dimensional unit hypercube. It is a crisp representation. In fact, when a nucleotide in a genome sequence is represented as (1, 0, 0, 0), it is meant that the nucleotide is fully recognized as T. It is not C, not A or G. Similar meaning may be given against binary representations of C, A, G also. But due to chemical nature of the nucleotides, sometimes a component of a genome sequence may not be completely understood. There is a gradation in the understanding. For example for the codon XAU ofDNA, where X = (0.1, 0.2, 0.3, 0.4, 0, 0, 1, 0, 1, 0, 0, 0), the first letter X is unknown and corresponds to T to extent 0.1, C to extent 0.2, A to extent 0.3 and G to extend 0.4. Now 0.1, 0.2, 0.3, 0.4 are the

membership values whose sum is 1. So XAU is a fuzzy vector. Hence in such cases, crisp representation of codon in I12 fails. It is a fuzzy representation. It is noted that any genome sequence can be reduced to a fuzzy vector of 12 components only. The descriptors for genome sequences are the corresponding fuzzy vectors. Obviously the method is applicable to sequences of any length, equal or unequal. The distance measure is the known NTV metric. But other metrics are defined based on the NTV metric. Further three examples of genome sequences are cited, where all such metrics behave similarly [40]. As the number of such examples are only three, so question is raised , whether such similar behavior of the metrics is general [41]. In this paper counter examples are created to show that the behavior is not true, in general. This led to the introduction of Intutionistic Fuzzy Polynucleotide space in [42]. It is shown that under the Intutionistic distance measure, the results may be shown to be general.

Probalistic representation is based on a 2D numerical representation of A, G, C and T by 2D points (1, 0.8), (1, 0.6), (1, 0.4) and (1, 0.2) respectively [43]. Under this representation, for sequence of length n, this representation is used to define a probability vector,

$$(p_1, p_2, .., p_n) = \frac{x_i - y_i}{\frac{1}{2}n(n-1)y_n}$$

where (xi, yi) represents the position of the ith nucleotide in the DNA graphical curve, represents the choice of y-coordinate value at the ith nucleotide in the DNA geometrical curve. These probability vectors are used as the descriptors. As the length of the descriptor depends on the length of the sequence, so it is not applicable to sequences of unequal lengths. Again it could not be proved that.

$$\sum_{i=1}^{n} p_i = 1 .$$

So (p1, p2,..,pn) is not a probability vector at all. This is the drawback in the representation. Again the distance measure used is the symmetric form of Kull-back Leibller divergence measure. It may be remarked that this cannot be taken as a distance measure, in the proper sense of the term, as it fails to satisfy the property of triangular Inequality. So this is another drawback of the paper.

Chaos Game representations are those where the representation is made by the application of the rule of Chaos game [44].

k-mer representations

k-mer representations are called di-nucleotide, tri-nucleotide representations, where k = 2 and 3 respectively. K-mer means we consider k number of consecutive nucleotides at a time starting from the first position in the sequence, then from the second position, then from third position and so on. As a result, for sequence of length n, (n-k+1) number of such k-mers is generated. Now the question arises how to assign real numbers to such k-tuples of nucleotides. In particular, for di-nucleotide representation such distinct pair of nucleotides are 42 = 16 in number, for tri-nucleotide representations such triplets of nucleotides are 43 = 64 in number. Different types of di and tri nucleotide representations are found in the literature. But all such representations are simply numerical representations, and not geometrical representations. The reason is that for a sequence of length 100, number of di-nucleotides are 99 in number, but the only 16 values are at hand, which correspond to 16

di-nucleotide pair. Obviously the curve consists of 16 distinct points. Genome sequence comparison based on di- nucleotide representations are found in paper [45-47]. Similar representations for tri-nucleotide representations are available in paper [48,49]. Very recently use of tri-nucleotide based representation for comparing genome sequences is taken up in [50]. In this case, each of the genome sequences is expressed as a 64×10real matrix, which is used as the descriptor. The distance measure is Euclidean. The results are also satisfactory. This is the only paper, where tri-nucleotide based degenerate representation is made non-degenerate by taking frequencies of the tri-nucleotides as the additional component. k-mer representations are also considered in different papers. Results of comparison show that they differ owing to the choice of k and also choice of distance measure. Recently it is shown that all k-mers can be well described by a probability vector of the same size, which may be taken as the descriptor [51]. By choice of k = 3 and distance measure as the information based similarity index, all genome sequences can be successfully classified.

Symbolic dynamic representation: In this case rule of symbolic dynamics is used in DNA sequence representation [52]. This method can visualize DNA sequences in three-dimensional coordinates with no loss of information in the transfer of data from a DNA sequence to its mathematical representation. It is used in the examination of similarities/dissimilarities among the coding sequences of the first exon of β -globin gene of different species.

REPRESENTATIONS FOR PROTEIN SEQUENCES, THEIR DESCRIPTORS AND DISTANCE MEASURES

Extensions of existing methods for genome sequences

(i) In the book chapter, the method of symbolic dynamics is extended from comparison of DNA sequences to protein sequences [53].

(ii) In the paper, 4 x 4 condensed matrix representation of genome sequences is extended to 20 x 20 condensed matrix representation of protein sequences [38,54]. The descriptors are the 20 x 20 matrix of the respective sequences. The distance measure is the alley index of matrices.

(iii) In this paper, 4D Binary representation of nucleotides is generalized to a 20D Binary representation of amino acids. The numerical sequences are made equal by addition of zeroes [55]. Then they are transformed to frequency domain by applying

Fourier transform. Addition of zeroes has no effect on the Fourier transform of the series. Then by applying ICD method in frequency domain, descriptors are obtained. Finally under Euclidean distance measures the protein sequences are compared. The method is equally applicable to sequences of equal and unequal lengths.

(iv) In the present paper, the concept of 12 component representations of codons is applied to extend the representation of an amino acid by a to 240 component vector. For those amino acids, which are expressed by a single codon, the representation is a crisp representation [56]. But those amino acids, which are expressed by multi-codons, the representation is a fuzzy representation. It is shown that each amino acid can be represented by a 240 component vector, of which 12 components have nonzero fuzzy values, rest are all zeroes. Based on lengths of respective 240 components, 20 amino acids under such representations are compared under usual Euclidean measures. Based on similarity of amino acids, it is shown that amino acids can be classified into six distinct groups. This is a pioneering work giving theoretical classified groups of amino acids of cardinality six.

(v) Probabilistic representation of DNA sequences [43] has been extended to probabilistic representation of protein sequences in [57]. But the same drawbacks are also present in this paper[43].

(vi) Extension of K-mer Method: So far as the extension of k-mer method in protein sequence comparison is concerned, it is noted that for protein sequences, the problem of k-mer representation is something difficult. In fact, for protein sequences, such distinct pair of amino acids are 202 = 400 in number, where as for distinct triplet of amino acids, these are 203 = 8000 in number. For higher values of k, number of such k-mers of amino acids is very large. So such method of representation is avoided for protein sequence comparison.

Representation based on physio-chemical properties of amino acids

(i) Based on two properties Volume and Polarity, a multiple sequence alignment program MAFFT was developed in But no attempt was made for the use of FFT in protein sequence comparison [58].

(ii) In paper, the complex representation based on the properties of hydrophobicity and residue volume is given. But no protein sequence comparison based on this representation is considered [59].

(iii) In paper, the complex representation of amino acids based on the properties of hydrophilicity and residue volumes is used [60]. The representation is not the same as the earlier one. In this paper the represented sequence is transferred to the frequency domain by Fourier transform. But the transformation is something special, as the original sequence under consideration is a complex sequence, not a real one. Anyway ICD method for such a transformation is modified accordingly and with suitable descriptor protein sequence is carried out using Euclidean norm as the distance measure.

Interestingly, the protein sequences are compared for both types of representations given in [59] and [60]. It is found that in the later case, the result is better. It proves that the property of hydro phillicity (polarity) is a better choice for protein sequence comparison.

(iv) Another property based representation and its use in protein sequence comparison is found in paper[61]. For analysis of protein sequences, they consider representations of amino acids under nine different properties. These are mW, hI, pk1, pk2, pI, S, cN,F(%) and vR.

3. Representations based on classified groups of amino acids with different cardinalities

Representations and comparison of protein sequences under different classified groups are considered in the following papers following different methods: These are given in paper corresponding to the representations i(a), ii(a), ii(b) and (iii) respectively [62-65].

In paper all the above methods of comparisons under different classifications of groups have been unified to a single method, which gives satisfactory, results in all the cases [66].

(d) Representation based on pair of classified groups of amino acids

Protein Sequence comparison on the basis of pair of classified groups of amino acids is found in paper. The 2D representation is really interesting [67]. But the limitations are that one of the classified group is not correct and that the methodology is not rigorous, it is nothing but a trial and error policy.

CONCLUSION

The method of comparison of Genome and protein sequences is still an ongoing process.

REFERENCES

1. Yao YH, Kong F, Dai Q, He PA. A sequence-segmented method applied to the similarity analysis of long protein sequence. MATCH: Commun in Mathematical Comput Chem. 2013 ;70(1): 431-7450.
2. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nature Struct Biol. 1999; 6(11): 1033-1038.
3. Yu ZG, Anh V, Lau KS. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J Theor Biol. 2004; 226(3):341-348.
4. Li C, Xing L, Wang X. 2-D graphical representation of protein sequences and its application to corona virus phylogeny. BMB Reports. 2007; 217-222.
5. Ghosh S , Pal J, Bhattachara DK. Classification of Amino Acids of a Protein on the basis of Fuzzy set theory. Int J Modern Sci Engineer Technol (IJMSET).
6. Gates MA. A simple way to look at DNA. J Theor Biol. 1986; 119(3):319-328.
7. Nandy A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Current Science.1994: 309-314.
8. Leong PM, Morgenthaler S. Random walk and gap plots of DNA sequences. Bioinformatics. 1995; 11(5):503-507.

9. Guo X, Randic M, Basak SC. A novel 2-D graphical representation of DNA sequences of low degeneracy. Chem Phys Lett. 2001; 350(1-2):106-112.
10. Yau SS, Wang J, Niknejad A, Lu C, Jin N, Ho YK. DNA sequence representation without degeneracy. Nuc Acids Res. 2003; 31(12): 3078-3080.
11. Liao B. A 2D graphical representation of DNA sequence. Chem Phys Lett. 2005; 401(1-3):196-199.
12. Liao B, Tan M, Ding K. Application of 2-D graphical representation of DNA sequence. Chem Phys Lett. 2005; 414(4-6): 296-300.
13. Song, J., Tang, H., A new 2-D graphical representation of DNA sequences and their numerical characterization, J biochem biophys methods, 2005; 63: 228-239.
14. Randić M, Vračko M, Lerš N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett. 2003; 368(1-2):1-6.
15. Randic M, Vracko M, Lers N, Plavsic D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem Phys Lett. 2003;371: 202-207.
16. Yao YH, Liao B, Wang TM. A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. J Mol Structure: THEOCHEM. 2005; 755(1-3):131-136.
17. Randić M, Vracko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inform Comput Sci. 2000; 40(5): 1235-1244.
18. Gates MA. A simple way to look at DNA. J Theor Biol. 1986; 119(3):319-328.
19. Nandy A, Nandy P. Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. Curr Sci. 1995: 75-85.
20. Leong PM, Morgenthaler S. Random walk and gap plots of DNA sequences. Bioinformatics. 1995; 11(5): 503-507.
21. Yao YH, Nan XY, Wang TM. A new 2D graphical representation—Classification curve and the analysis of similarity/dissimilarity of DNA sequences. J Mol Str: THEOCHEM. 2006; 764(1-3):101-108.
22. Das S, Pal J, Bhattacharya DK. Geometrical method of exhibiting similarity/dissimilarity under new 3D classification curves and establishing significance difference of different parameters of estimation. Int J Adv Res Comput Sci Soft Engg. 2015; 5(5): 279-287.
23. Randić M, Witzmann F, Vračko M, Basak SC. On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: application to peroxisome proliferators. Med Chem Res. 2001;10(7-8):456-479.
24. Randić M, Vracko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inform Comput Sci .2000; 40(5): 1235-1244.
25. Randić M, Vračko M, Lerš N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett. 2003;368 (1-2):1-6.
26. Randić M, Vračko M, Lerš N, Plavšić D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem Phy Lett. 2003;371: 202-207.
27. Qi ZH, Fan TR. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett. 2007; 442: 434-440.
28. Akhtar M, Epps J, Ambikairajah E. Signal processing in sequence analysis: advances in eukaryotic gene prediction. IEEE J Selected Topics in Signal Process. 2008; 2(3):310-321.
29. Chakravarthy N, Spanias A, Iasemidis LD, Tsakalis K. Autoregressive modeling and feature analysis of DNA sequences. EURASIP J Adv Signal Processing. 2004; 2004(1):952689.
30. Zhou H, Yan H. Autoregressive models for spectral analysis of short tandem repeats in DNA sequences. In: 2006 IEEE International Conference on Systems, Man and Cybernetics. 2006; 2: 1286-1290.
31. Anastassiou D. Genomic signal processing. IEEE signal processing magazine. 2001 ;18(4):8-20.
32. Cristea PD. Genetic Signal Representation and Analysis. In: SPIE Conference, BIOS ' 2002- International Biomedical Optics Symposium, Molecular Analysis and Informatics, San Jose.2002: 77-84.
33. Cattani C. Complex Representation of DNA Sequences. In: 2nd International Conference on Bioinformatics Research and Development-BIRD, Australia. 2008; 13: 528-537.
34. Brodzik AK, Peters O. Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences. In: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing IEEE,2005. 2005; 5: 373.
35. Chi R, Ding K. Novel 4D numerical representation of DNA sequences. Chem Phys Lett. 2005; 407: 63-67.
36. King BR, Aburdene M, Thompson A, Warres Z. Application of discrete Fourier inter-coefficient difference for assessing genetic sequence similarity. EURASIP J Bioinform Syst Biol. 2014; 2014(1):8.
37. Hoang T, Yin C, Zheng H, Yu C, He RL, Yau SS. A new method to cluster DNA sequences using Fourier power spectrum. J Theor Biol. 2015; 372:135-145.
38. Randić M. On characterization of DNA primary sequences by a condensed matrix. Chem Phys Lett. 2000; 317: 29-34.
39. Torres A, Nieto JJ. The fuzzy polynucleotide space: basic properties. Bioinformatics. 2003;19(5):587-592.
40. Nieto JJ, Torres A, Georgiou DN, Karakasidis TE. Fuzzy polynucleotide spaces and metrics. Bulletin Of Math Biol. 2006; 68(3):703-725.
41. Das S, De D, Dey A, Bhattacharya D. Some anomalies in the analysis of whole genome sequence on the basis of Fuzzy set theory. Int J Art intel Neural Networks. 2013; 3(2):38-41.
42. Das S, De D & Bhattacharya DK. Similarity and Dissimilarity of Whole Genomes using Intuitionistic Fuzzy Logic,Notes on Intuitionistic Fuzzy Sets. 2015; 2: 48-53.
43. Yu C, Deng M, Yau SS. DNA sequence comparison by a novel probabilistic method. Inform Sci. 2011; 181: 1484-1492.
44. Deng W, Luan Y. Analysis of similarity/dissimilarity of DNA sequences based on chaos game representation. In: Abstract and Applied Analysis Hindawi. 2013: 2013.
45. Qi ZH, Fan TR. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett. 2007; 442: 434-440.
46. Yu JF, Wang JH, Sun X. Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. MATCH Commun. Math. Comput. Chem. 2010; 63: 493-512.
47. Liu XQ, Dai Q, Xiu Z, Wang T. PNN-curve: A new 2D graphical representation of DNA sequences and its application. J Theor Biol. 2006; 243(4):555-561.
48. Randić M, Zupan J, Balaban AT. Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem Phy Lett. 2004 Oct 11; 397: 247-252.
49. Yu, J.F., Sun, X., Wang, J.H., TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. J Theor Biol. 2009; 261: 459-468.

50. Das S, Das A, Mondal B, Dey N, Bhattacharya DK, Tibarewala DN et al. Genome sequence comparison under a new form of tri-nucleotide representation based on bio-chemical properties of nucleotides. Gene.2020: 730:144257.
51. Das S, Deb T, Dey N, Ashour AS, Bhattacharya DK, Tibarewala DN. Optimal choice of k-mer in composition vector method for genome sequence comparison. Genomics. 2018; 110: 263-273.
52. Wang S, Tian F, Feng W, Liu X. Applications of representation method for DNA sequences based on symbolic dynamics. J Mol Str: THEOCHEM. 2009; 909: 33-42.
53. Pal J, Dey A, Ghosh S, Bhattacharya DK, Mukherjee T. Analysis of similarity between Protein Sequences through the study of Symbolic Dynamics. In: Computational Advancement in Communication Circuits and Systems Springer. 2015: 197-214.
54. Ghosh S, Pal J, Maji B, Bhattacharya DK. Condensed Matrix Descriptor for Protein Sequence Comparison. Int J Anal Mass Spectrom Chromat. 2016; 4:1-3.
55. Pal J, Ghosh S, Maji B, Bhattacharya DK. Use of fft in protein sequence comparison under their binary representations. Comput Mol Biosci. 2016; 6:33.
56. Ghosh S, Pal J, Bhattacharya DK. Classi-fication of Amino Acids of a Protein on the Basis of Fuzzy Set Theory. Int J Mod Sci Technol. 2014;1:30-35.
57. Gupta MK, Niyogi R, Misra M. Their Similarity Analysis with Probabilistic Method. MASTCH Commun. Math Chem.2014; 72: 519-532.
58. Katoh K, Misawa K, Kuma KI, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002; 30: 3059-3066.
59. Yin C, Yau SS. Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes. In: 2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. 2008: 223-227.
60. Pal J, Maji B, Bhattacharya DK. Protein sequence comparison under a new complex representation of amino acids based on their physio-chemical properties. Int J Engg Technol. 7; 2018: 181-184.
61. Ping P, Zhu X, Wang L. Similarities/dissimilarities analysis of protein sequences based on PCA-FFT. J biol syst.2017; 25: 29-45.
62. Yao YH, Kong F, Dai Q, He PA. A sequence-segmented method applied to the similarity analysis of long protein sequence. MATCH: Commun Math Comput Chem. 2013; 70:431-450.
63. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nature struct biol.1999; 6:1033-1038.
64. Yu ZG, Anh V, Lau KS. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J Theor Biol. 2004; 226: 341-348.
65. Li C, Xing L, Wang X. 2-D graphical representation of protein sequences and its application to corona virus phylogeny. BMB reports. 2007: 217-222.
66. Ghosh S, Pal J, Maji B, Bhattacharya DK. A sequential development towards a unified approach to protein sequence comparison based on classified groups of amino acids - Int J Engg Technol. 2018; 7: 678-686.
67. Yao YH, Kong F, Dai Q, He PA. A sequence-segmented method applied to the similarity analysis of long protein sequence. MATCH-Commun Math Co. 2013; 70: 431-450.