HSPI The Open Access Publisher

JGMGT JOURNAL OF GENETIC MEDICINE AND GENE THERAPY
ISSN 2994-256X

Review Article

A Critical Review on Some Recent Developments in Comparison of Biological Sequences

DK Bhattacharya*

Ex. Professor & Head, Department of Pure Mathematics, Calcutta University, Kolkata, India

More Information

***Address for correspondence:** DK Bhattacharya, Ex. Professor & Head, Department of Pure Mathematics, Calcutta University, Kolkata, India, Email: dkb_math@yahoo.com

Submitted: April 01, 2024
Approved: April 24, 2024
Published: April 25, 2024

How to cite this article: Bhattacharya DK. A Critical Review on Some Recent Developments in Comparison of Biological Sequences. J Genet Med Gene Ther. 2024; 7: 008-014.

DOI: 10.29328/journal.jgmgt.1001010

Keywords: Biological groups of nucleotides; Physio-chemical properties of amino acids; Classified groups of amino acids; Standard secondary structures of protein

Abstract

The present review highlights some of the very important contributions to non-alignment ways of comparing biological sequences, which may be genome sequences of nucleotides, protein sequences of amino acids, or sequences of protein secondary structures. The discussion centers around specific methods applicable to the comparison of three types of sequences. The methods of comparison of genome sequences are based on three pairs of biological groups of nucleotides; the same for protein sequences are based on either physio-chemical property values of amino acids or on classified groups of amino acids of different cardinalities obtained from the physio-chemical properties; the same for sequences of secondary structures of proteins are based on their sequential expressions of structure elements of cardinality three and four. Comparison is made in the time domain and also in the frequency domain. Different taxa of known phylogeny are considered for comparison. It tries to find out the specific method of comparison, which can show the exact phylogeny of the taxa. If a new sequence appears in the database, it becomes essential to know its phylogeny. For this purpose, a phylogenetic tree is drawn on the sequences of the known taxa together with this new sequence using the best possible method. If the species having this new sequence belongs to the old taxa, there is nothing to worry about. Otherwise, the species with the new sequence has to be studied separately. This is the general reason for the construction of a phylogenetic tree in any form of biological sequence comparison.

Introduction

There are two types of methods of biological sequence comparison- one is alignment-based and the other one is alignment-free. The former type is now rarely used for its time complexity. Non-alignment (alignment-free) method of comparison of biological sequences means comparison based on the corresponding sequences of represented numerical values. Therefore the primary thing in the non-alignment method is the numerical representation of the sequence. This may be arbitrary or specific to the type of biological sequence, which is used for comparison.

For genome sequences, the representations may be real-valued (one or more dimensional), may be binary or non-binary, maybe complex-valued, and may even be quaternion-valued. Real-valued representations are considered in [1-19]; complex-valued representation is considered in [20-22]. The quaternion-valued representation is found in [23]. But in all these cases, the representations are found to be arbitrary. Out of them, one binary four-dimensional representation of nucleotides is found to be very useful. It is applied in genome sequence comparison in the frequency domain under the use of one-dimensional FFT and ICD methods of selection of descriptors [24]. It is also used in getting four component sequences corresponding to a given genome sequence and defining the descriptors given by different moment vectors of different degrees in the frequency domain [25,26]. However, descriptors given by such moment vectors have some limitations. It is found that the results of the comparison of genome sequences are sequence-specific. For some sequences moment vectors of certain degrees work well, but for other sequences, they fail to give correct results. The four-dimensional binary representation is also used for the representation of genomes on a unit 12-dimensional hypercube resulting in a comparison of genome sequences under the NTV metric defined therein [27,28]. The above 12-dimensional representation is found to have some shortcomings. These are removed in the similar 4×4=16 dimensional representation of genomes and their comparisons [29]. Recently, such a 16-dimensional representation of genome sequences has been used in obtaining a suitable form of metric divergence to apply in the comparison of genome sequences successfully [30].To

talk about the specific methods, which are applicable to the comparison of genome sequences only, it may be mentioned that four bases A, C, G, and T of primary DNA sequences may be classified in two different ways (Table 1).

Some of the representations based on chemical properties and strength of Hydrogen bonds are made in [31-36]. All these representations are found to be degenerate; they are not fit for geometric representations. The degeneracy is first removed in the representation of [37]. A graph-theoretic comparison based on the Bio-chemical properties of nucleotides is used in [38].

For Protein Sequence comparison, there are two types of representations-one being extensions of representations of nucleotides, the other one being specific to properties of amino acids.

Some such extensions are the 20-dimensional binary representation of amino acids, which are used in protein sequence comparison. It is first used in obtaining theoretical classification of amino in a group of cardinality six [39]. Based on such binary representation of amino acids, representation of protein sequences on a 20×20 dimensional unit hypercube is obtained and protein sequence comparison is carried out effectively by using a modified form of NTV metric [40]. Next, the extended binary representation is used for protein sequence comparison in the frequency domain under one-dimensional FFT and modified ICD method [41]. The twenty-dimensional binary representation of amino acids has also been used recently in getting moment vectors in the frequency domain to use as descriptors of protein sequence comparison [42]. These descriptors are not sequence-specific as in the case of genome sequences. These descriptors are newly defined and are called minimal moment vectors. These are different from the standard moment vectors and the central moment vectors, which are used for genome sequence comparison. It has been possible to compare protein sequences efficiently using such minimal moment vectors of only degree two.

The second type of representation has two parts – one based on the physio-chemical property values of amino acids and the other based on classified groups of amino acids of different cardinalities based on the physio-chemical properties of the amino acids (Tables 2,3).

There are lots of papers on protein sequence comparison based on property values of amino acids, the number ranging from two to twelve [43-54]. Even complex-valued representation based on a pair of property values of amino acids is also known [55] and it is used successfully in protein sequence comparison [56]. But protein sequence comparison based on a single property value of amino acids is a very recent one [57]. This deals with meaningful non-binary representations, which are by no means arbitrary. Another very recent contribution to the non-binary representation of protein sequences is an interesting one [58]. It enables to giving of a non-binary representation of protein sequences on a 20×20 dimensional unit hypercube. Protein sequence comparison is successfully implemented by the modified NTV metric mentioned earlier.

Table 1: Classification of nucleotides.

Chemical Properties	Strength of the Hydrogen Bond
I. Purine group R = (A, G) and Pyrimidine group Y = (C, T). II. Amino group M = (A, C) and Keto group K = (G, T).	Weak H–bonds W = (A, T) and Strong H–bonds S = (G, C).

For protein sequence comparison based on classified groups, the following groups (Tables 4-9) are used:

Different classified groups of amino acids

The sequence comparison based on four classified groups of amino acids is found in [59], and the same for five classified groups is found in [60]. The methods are different in the two cases. A unified method is developed in [61], which works for four, five, and six group classifications.

For the purpose of comparison of Protein secondary structure sequences, it may be noted that Protein Secondary structure elements (SSE) of greatest interest include α-helices and β-strands. They are represented as H and E respectively. There is another kind of SSE called Coils and they are denoted by the letter C. A secondary structure sequence is a symbolic string comprising of the above three kinds of letters H, E, and C similar to .20 different letters representing amino acids in protein sequences. Recurrences of consecutive H in the symbolic representation of SSE mean those positions of the sequences previously occupied by different amino acids are now forming a single helix. The same is true for consecutive E and consecutive C. In the 1D summary (sequential form), SSE is represented graphically as α-helices depicted as waves and β-strands shown by wide arrowheads respectively. The remaining positions of strings of SSE are represented by C (coils) and they are shown simply by straight lines. Levitt and Chothia [62] first proposed the concept of structural classes of proteins in 1976. The proteins were grouped into one of the four classes, all-α, all-β, α + β, and α/β. The all-α and all-β proteins are defined to be composed of almost entirely α-helices and β-strands, almost in the sense of excluding the coils present in between. The α + β proteins consist of α-helices and β-strands, where α regions and β regions are largely separated and the β-strands are often anti-parallel; α/β proteins consist of alternative mixtures of α-helices and β-strands, where the β-strands are often parallel. These assignments basically characterize the overall secondary structures of proteins even in the up-to-date databases, and thus they have been generally accepted and widely used in the literature. Since 1976, the problem of protein classification has been tackled by many groups of researchers [63–69]. In most of the cases, the selection criterion is based on the helix/strand contents of proteins. The protein secondary structure sequence appears in PDB as a sequence of three

Table 2: Physical properties of Amino acids.

Sl.No	Amino acids		Relative Dis. RD	Sidechain Mass	Specific Volume	Residue Volume	Residue Wt	Mole Vol
1	Alanine	A	0.2227	15	0.64	43.5	71.08	31
2	Cysteine	C	1.0000	47	0.74	60.6	103.14	55
3	Methionine	M	0.1882	75	0.70	77.1	131.191	105
4	Proline	P	0.2513	41	0.63	60.8	97.12	32.5
5	Valine	V	0.1119	43	0.76	81	99.13	84
6	Phenylalanine	F	0.2370	91	0.86	91.3	147.17	132
7	Isoleucine	I	0.1569	57	0.90	107.5	113.16	111
8	Leucine	L	0.1872	57	0.90	107.5	113.16	111
9	Tryptophan	W	0.4496	130	0.75	105.1	186.21	170
10	Tyrosine	Y	0.1686	107	0.77	121.3	163.18	136
11	Aspartic acid	D	0.3924	59	0.71	123.6	115.09	54
12	Lysine	K	0.1739	72	0.68	144.1	128.17	119
13	Asparagine	N	0.2513	58	0.62	78	114.1	56
14	Arginine	R	0.0366	100	0.66	90.4	156.19	124
15	Serine	S	0.2815	31	0.60	74.1	87.08	32
16	Glutamic acid	E	0.1819	73	0.67	93.9	129.12	83
17	Glycine	G	0.3229	1	0.82	108.5	57.05	3
18	Histidine	H	0.0201	81	0.70	111.5	137.14	96
19	Gluatmine	Q	0.0366	72	0.67	99.3	128.13	85
20	Threonine	T	0.0000	45	-	72.5	101.11	61

Table 3: Chemical Properties of Amino acids.

Sl. No	Amino acids		pKa- $COOH^{17}$	pKa- NH^{+17}_{3}	Hydropathy Index h	Hydro-phobicity	Hydro-Phillicity	Isoelectric Point Pi	Polar requirement
1	Alanine	A	2.34	9.69	1.8	-0.7	1.8	6.01	7.0
2	Cysteine	C	1.71	10.78	2.5	1.8	-4.5	5.07	4.8
3	Methionine	M	2.28	9.21	1.9	-0.7	-3.5	5.74	5.3
4	Proline	P	1.99	10.60	-1.6	-0.8	-3.5	6.48	6.6
5	Valine	V	2.32	9.62	4.2	-1.6	2.5	5.97	5.6
6	Phenylalanine	F	1.83	9.13	2.8	-4.2	-3.5	5.48	5.0
7	Isoleucine	I	2.36	9.68	4.5	3.8	-3.5	6.02	4.9
8	Leucine	L	2.36	9.60	3.8	4.5	-3.5	5.98	4.9
9	Tryptophan	W	2.38	9.39	-0.9	1.9	-0.4	5.89	5.2
10	Tyrosine	Y	2.20	9.11	-1.3	2.8	3.2	5.66	20.5
11	Aspartic acid	D	2.09	9.82	-3.5	-1.3	4.5	2.77	13.0
12	Lysine	K	2.18	8.95	-3.9	-0.09	3.9	9.74	10.1
13	Asparagine	N	2.02	8.80	-3.5	-3.5	1.9	5.41	10.0
14	Arginine	R	2.17	9.04	-4.5	-3.5	2.8	10.76	9.1
15	Serine	S	2.21	9.15	-0.8	-3.5	-1.6	5.68	7.5
16	Glutamic acid	E	2.19	9.67	-3.5	-3.5	-0.8	3.22	12.5
17	Glycine	G	2.34	9.60	-0.4	-3.9	-0.7	5.97	7.9
18	Histidine	H	1.82	9.17	-3.2	-4.5	-0.9	7.59	8.4
19	Glutamine	Q	2.17	9.13	-3.5	-3.2	-1.3	5.65	8.6
20	Threonine	T	2.63	10.43	-0.7	2.5	4.2	5.87	6.6

Table 4: Three-group classification of amino acids.

Characteristic	Amino Acids
Dextrorotatory	E, A, I, K, V
Levorotatory	N, C, H, L, M, F, P, S
Irrotational	G, Y, R, D, Q

Table 5: Four-group classification of amino acids (detailed HP model).

Characteristic	Amino Acids
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

Table 6: Four-group classification of amino acids (HC Model).

Hydropathy characteristic	Abbreviation	Amino Acids
Strongly Hydrophilic	POL	R, N, D, Q, E, 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

Table 7: Five-group classification of amino acids.

Representative residues	Amino acids
I	C, M, F, I, L, V, W, Y
A	A, T, H
G	G, P
E	D, E
K	S, N, Q, R, K

letters H, E, and C. On closed observation, it is also possible to observe up and down helices and strands. So in addition to geometric representations of helix and strand by wave and wide arrows respectively, there is another form of the geometric representation of SSE, called TOPS (topology of protein structure) diagram; this is one of the most popular protein structure topological descriptors. TOPS considers sequences of the secondary structure elements (SSEs), along with relationships like spatial adjacency within the fold and approximate orientation, neglecting details like lengths and structures of loops, and the lengths of the secondary structure

Table 8: Six-group classification of amino acids (Biological).

Characteristic	Amino Acids
Side chain is aliphatic	G, A, V, L, I
Side chain is an organic acid	D, E
Side chain contains a sulfur	C, M
Side chain is an alcohol	S, T, Y
Side chain is aromatic	F, W, Y
Side chain is an organic base	R, K, H

Table 9: Six-group classification of amino acids (Theoretical).

Sl	Amino acids	Representative residues
1	I	I
2	L, R	L
3	V, A, G, P, T	A
4	F, C, Y, Q, N, H, E, D, K	E
5	M, W	M
6	S	S

elements themselves. Some of the attempts to obtain such TOPS representation are given in [70-74]. In 1977, Stemberg and Thornton [70] presented the TOPS cartoons, which were originally drawn manually. This gives graphical representations of SSEs. Flores, et al. [71]. TOPS diagrams provide a concise way of describing the structural topology of protein. This includes their sequences of SSE together with some information about α-helices and β-strands, which are represented respectively by triangles and circles. Secondary structure elements are considered to have a direction of 'up' (out of the plane of the diagram) or 'down' (into the plane of the diagram). As each SSE in a TOPS diagram is associated with direction up or down, so in all four letters are needed to express the TOPS string; they are {h, H, e, E}, where E stands for 'up' strand, e for 'down' strand, H for 'up' helix, and h for 'down' helix. In [74] a computer system is developed to compare protein SSE represented by the TOPS diagram. It may be noted that the comparison of proteins based on their representations as sequences of four secondary structures follows the same line as that of genome sequences. The reason is that symbolically there is no difference in the two cases. Genomes are sequences of four nucleotides whereas proteins are sequences of four secondary structures. Therefore most of the attempts at protein structure comparison consider sequences of four secondary structures H, h, E, and e. In [75] the authors compared similarities/ dissimilarities of the SSE of protein by using a universal similarity metric. They used different types of data sets – (i) Random data set of 40 proteins (ii) Chew–Kedem dataset of 36 proteins [76] (iii) Skolnick dataset of 39 proteins [77] and (iv) Leluk–Konieczny–Roterman dataset [78]. Structural similarity between proteins in four different datasets was investigated under the USM (universal symmetric metric). The sample represented alpha, beta, alpha-beta, tim-barrel, globins, and serpine protein types. The use of the proposed metric shows a correct measurement of the similarity and classification of the proteins in the four datasets. Gilbert, et al. [79,80] explored the alignment-free comparison of the topology of protein structure diagrams. In [81] the authors compared TOPs strings based on LZ complexity; the same method was used earlier in connection with a comparison of DNA sequences [82]. The results of [81] when compared with those of [80] are found to be better in some cases. Alignment-free LZ complexity method is also compared with the alignment-based Clustal W method. It is found that the results of Clustal W are not satisfactory in some cases. In [83] the authors used information discrepancy measures to compare protein secondary structures. In [84] the authors tried to study the comparison of SSE based on TOPS representation using the first difference and second difference of the three-dimensional represented values. It is a new attempt. However, the results were not satisfactory. In [85] the authors applied a novel method of comparison of SSE from a data set comprising 20 different structures belonging to standard four classes and were able to show that the classification was justified theoretically. They also considered 36 protein structures from the Chew–Kedem database and their method was successful in classifying the structures very well. But both the data sets when analyzed by the ClustalW method show incorrect classification in some cases.

So far as the comparison of SSE expressed by three structures H, E, and C is concerned, possibly there are very few attempts to date. In [86] transition probability matrix and structural characteristic vectors of proteins are constructed. The phylogenetic trees of 20 proteins from four different classes α,β,α+,α/β and TOPS strings of the 36 protein chains in the Chew–Kedem dataset are constructed. The result shows that this new approach to measuring the similarities between protein secondary structures is computationally efficient. As the above four classes have nothing to do with the phylogeny classes of the Taxa, the above 20 proteins having three structures are to be further examined for their phylogeny classes under methods suitable for the purpose. There is another interesting area of research. A comparison is made on the degree of prediction of the secondary structure sequences of PDB with the same sequence obtained by some other methods. 2D graphical representation of SSE is considered based on triplets H, E, and C [87]. First, a truncated portion of 2pgdI is taken from PDB. This is the reality secondary structure sequence of the protein 2pgdI. Next similar truncated sequences corresponding to the predicted secondary structure sequences by the NN prediction method of Rost and Sander [88] and the PHD method of Kneller, et al. [68] are taken up one by one. From the 2D represented curves, M/M matrices are formed in all cases and finally, their leading eigenvalues are taken as descriptors. Now lesser is the difference between the two leading eigenvalues, and better is the similarity between the two sequences. It is found that the similarity between the Reality (the reality secondary structure sequence) and the PHD (the secondary structure sequence predicted by the PHD method) is better than the one between the Reality and the NN prediction (the secondary structure sequence predicted by the NN prediction method). Taking the reality secondary structure sequence as the reference, it may be concluded that, the PHD prediction

method gives better results than the NN prediction method. A graphic representation of SSE given by triplets H, E, and C is also considered in [89]. This graphic representation is called the S curve. The S curve is the unique representation of a given secondary structure sequence in the sense that the sequence and the S curve can be uniquely determined from each other. Therefore, the S curve contains all the information that the secondary structure sequence contains. In this paper three sequences of oxo-acid-lyase 1csc are compared, one corresponds to the reality sequence given by PDB, one is predicted by the NN prediction method of [88] and the last one is predicted by the PHD prediction method of [68]. The comparison is made visually from their corresponding S curves. Again the same conclusion is reached that the PHD prediction method is better than the NN prediction method.

Discussion

For comparison of general Biological sequences, the method of analysis is performed in the time domain and also in the frequency domain. Apart from general methods of comparison, there are methods specific to genome sequences, as the nucleotides possess some Bio-chemical properties. Similar methods specific to protein sequences exist, as the amino acids have some interesting physio-chemical properties. For sequences of protein secondary structures, there are no specific methods. In fact, such sequences may be considered as sequences of three or five different symbols characterized by three or five secondary structures. They are not comparable with the 4-letter representation of genome sequences and the 20-letter representation of protein sequences. Separate methodologies are to be developed for comparison of sequences of protein secondary structures. However, for TOP's representation of secondary structures, methods of comparison of genome sequences are equally applicable as it considers only four letters just like four nucleotides A, G, C, and T.

Future scope

Comparison of genome sequences under di- nucleotide and tri-nucleotide representation may be an interesting area of research. This is not fully explored. The application of FFT in genome and protein sequences may be further studied. This might give some new results. Protein structure comparison using three types of secondary structures H, E, and C may be an important issue. A protein structure comparison using five elements H, h, E, e, and C may be an interesting one. The application of FFT in sequences of secondary structures is a new area of research. This may be studied further in detail.

Conclusion

Method of comparison of Biological sequences is an ongoing area of research. Such methods of comparison for sequences of protein secondary structures are less compared to those applicable for genome and protein sequences.

References

1. Gates MA. A simple way to look at DNA. J Theor Biol. 1986 Apr 7;119(3): 319-28. doi: 10.1016/s0022-5193(86)80144-8. PMID: 3016414.
2. Nandy A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Current Science. 1994; 309-314.
3. Leong PM, Morgenthaler S. Random walk and gap plots of DNA sequences. Bioinformatics. 1995; 11(5): 503-507.
4. Guo X, Randic M, Basak SC. A novel 2-D graphical representation of DNA sequences of low degeneracy. Chemical Physics Letters. 2001; 350(1-2):106-112.
5. Yau SS, Wang J, Niknejad A, Lu C, Jin N, Ho YK. DNA sequence representation without degeneracy. Nucleic Acids Res. 2003 Jun 15;31(12):3078-80. doi: 10.1093/nar/gkg432. PMID: 12799435; PMCID: PMC162336.
6. Liao B. A 2D graphical representation of DNA sequence. Chemical Physics Letters. 2005; 401(1-3):196-199.
7. Song J, Tang H. A new 2-D graphical representation of DNA sequences and their numerical characterization. J Biochem Biophys Methods. 2005 Jun 30;63(3):228-39. doi: 10.1016/j.jbbm.2005.04.004. PMID: 15939477.
8. Randić M, Vračko M, Lerš N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2003; 368(1-2):1-6.
9. Randić M, Vračko M, Lerš N, Plavšić D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chemical Physics Letters. 2003; 371(1-2): 202-207.
10. Yao YH, Liao B, Wang TM. A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. Journal of molecular structure: Theochem. 2005; 755(1-3):131-136.
11. Randić M, Vracko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inf Comput Sci. 2000 Sep-Oct;40(5):1235-44. doi: 10.1021/ci000034q. PMID: 11045819.
12. Nandy A, Nandy P. Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. Current Science. 1995; 75-85.
13. Yao YH, Nan XY, Wang TM. A new 2D graphical representation—Classification curve and the analysis of similarity/dissimilarity of DNA sequences. Journal of Molecular Structure: Theochem. 2006; 764(1-3): 101-108.
14. 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. Intl J Adv Res Comp Sci Softw Engg. 2015; 5:279-287.
15. 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. Medicinal Chemistry Research. 2001; 10(7-8):456-479.
16. Qi ZH, Fan TR. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters. 2007; 442(4-6): 434- 440.
17. Akhtar M, Epps J, Ambikairajah E. Signal processing in sequence analysis: advances in eukaryotic gene prediction. IEEE Journal of Selected Topics in Signal Processing. 2008; 2(3): 310-321.
18. Chakravarthy N, Spanias A, Iasemidis LD, Tsakalis K. Autoregressive modeling and feature analysis of DNA sequences. EURASIP Journal on Advances in Signal Processing. 2004; 2004(1):1-16.
19. Chi R, Ding K. Novel 4D numerical representation of DNA sequences, Chemical Physics Letters. 2005; 407:63-67.
20. Anastassiou D. Genomic Signal Processing. IEEE Signal Processing Magazine. 2001; 18:8-20.

21. Cristea PD. Genetic Signal Representation and Analysis, SPIE Conference, BIOS'2002- International Biomedical Optics Symposium, Molecular Analysis and Informatics, San Jose USA, B.O.4623-10, 2002; 77-84.

22. Cattani C. Complex Representation of DNA Sequences, 2nd International Conference on Bioinformatics Research and Development-BIRD. 2008; 13: 528-537.

23. Brodzik AK, Peters O. Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences. Proc IEEE ICASSP. 2005; 5: 373-376.

24. 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. doi: 10.1186/1687-4153-2014-8. Epub 2014 May 28. PMID: 24991213; PMCID: PMC4077688.

25. Zhao B, Duan V, Yau SS. A novel clustering method via nucleotide-based Fourier power spectrum analysis. J Theor Biol. 2011 Jun 21;279(1):83-9. doi: 10.1016/j.jtbi.2011.03.029. Epub 2011 Apr 2. PMID: 21443881; PMCID: PMC7094093.

26. Hoang T, Yin C, Zheng H, Yu C, Lucy He R, Yau SS. A new method to cluster DNA sequences using Fourier power spectrum. J Theor Biol. 2015 May 7;372:135-45. doi: 10.1016/j.jtbi.2015.02.026. Epub 2015 Mar 5. PMID: 25747773; PMCID: PMC7094126.

27. Nieto JJ, Torres A, Georgiou DN, Karakasidis TE. Fuzzy polynucleotide spaces and metrics. Bull Math Biol. 2006 Apr;68(3):703-25. doi: 10.1007/s11538-005-9020-5. PMID: 16794951.

28. Torres A, Nieto JJ. The fuzzy polynucleotide space: basic properties. Bioinformatics. 2003 Mar 22;19(5):587-92. doi: 10.1093/bioinformatics/btg032. PMID: 12651716.

29. Ghosh S, Pal J, Maji B, Bhattacharya DK. A method of genome sequence comparison based on a new form of fuzzy polynucleotide space. 7th International Conference on Emerging Applications of Information Technology (EAIT 2022). DOI: 10.1007/978-981-19-5191-6_11.

30. Ghosh S, Pal J, Maji B, Cattani C, Bhattacharya DK. Choice of Metric Divergence in Genome Sequence Comparison. Protein J. 2024 Mar 16. doi: 10.1007/s10930-024-10189-x. Epub ahead of print. PMID: 38492188.

31. Raychaudhury C, Nandy A. Indexing scheme and similarity measures for macromolecular sequences. J Chem Inf Comput Sci. 1999 Mar-Apr;39(2):243-7. doi: 10.1021/ci980077v. PMID: 10192941.

32. Randić M. On characterization of DNA primary sequences by a condensed matrix. Chemical Physics Letters. 2000; 317(1-2):29-34.

33. He PA, Wang J. Characteristic sequences for DNA primary sequence. J Chem Inf Comput Sci. 2002 Sep-Oct;42(5):1080-5. doi: 10.1021/ci010131z. PMID: 12376994.

34. Guo X, Randic M, Basak SC. A novel 2-D graphical representation of DNA sequences of low degeneracy. Chemical Physics Letters. 2001; 350(1-2):106-112.

35. Liu Y, Guo X, Xu J, Pan L, Wang S. Some notes on 2-D graphical representation of DNA sequence. J Chem Inf Comput Sci. 2002 May-Jun;42(3):529-33. doi: 10.1021/ci010017g. PMID: 12086510.

36. Yao Y, Nan XY, Wang T. A new 2D graphical representation—Classification curve and the analysis of similarity/dissimilarity of DNA sequences, Journal of Molecular Structure: THEOCHEM. 2006; 764(1–3):101-108.

37. 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, International Journal of Advanced Research in Computer Science and Software Engineering. 2015; 5: 279-287.

38. Das S, Das A, Mondal B, Dey N, Bhattacharya DK, Tibarewala DN. Genome sequence comparison under a new form of tri-nucleotide representation based on bio-chemical properties of nucleotides. Gene. 2020 Mar 10;730:144257. doi: 10.1016/j.gene.2019.144257. Epub 2019 Nov 21. PMID: 31759983.

39. Ghosh S, Pal J, Bhattacharya DK. Classification of Amino Acids of a Protein on the basis of Fuzzy set theory. International Journal of Modern Sciences and Engineering Technology. 2014; 1(6): 30-35.

40. Ghosh S, Pal J, Maji B, Bhattacharya DK. Protein Sequence Comparison on Fuzzy Matrix Amino Acid Space. IEEE Sponsored International Conference on Technological Advancements and Innovations (ICTAI - 2021). 2021; 10-12 Nov. 2021, DOI: 10.1109/ICTAI53825.2021.9673411.

41. Pal J, Ghosh S, Maji B, Bhattacharya DK. Use of FFT in protein sequence comparison under their binary representations. Computational Molecular Bioscience. 2016; 6(02): 33.

42. Pal J, Ghosh S, Maji B, Bhattacharya DK. MMV method: a new approach to compare protein sequences under binary representation. J Biomol Struct Dyn. 2024 Feb 20:1-7. doi: 10.1080/07391102.2024.2317982. Epub ahead of print. PMID: 38375605.

43. Yu L, Zhang Y, Gutman I, Shi Y, Dehmer M. Protein Sequence Comparison Based on Physicochemical Properties and the Position-Feature Energy Matrix. Sci Rep. 2017 Apr 10;7:46237. doi: 10.1038/srep46237. Erratum in: Sci Rep. 2017 May 04;7:46787. PMID: 28393857; PMCID: PMC5385872.

44. Wu ZC, Xiao X, Chou KC. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol. 2010 Nov 7;267(1):29-34. doi: 10.1016/j.jtbi.2010.08.007. Epub 2010 Aug 7. PMID: 20696175.

45. Randić M. 2-D graphical representation of proteins based on physicochemical properties of amino acids. Chem Phys Lett. 2007; 440: 291– 295, DOI: 10.1016/j.cplett.2007.04.03

46. Zhang Y, Zhan Y, Xu C. A novel method of 2D graphical representation for proteins and its application. MATCH Commun Math Comput Chem. 2016; 75: 431- 446.

47. Qi ZH, Jin MZ, Li SL, Feng J. A protein mapping method based on physio-chemical properties and dimension reduction. Comput Biol Med. 2015; 57:1-7. DOI: 10.1016/j.compbiomed.2014.11.012

48. Yao YH, Dai Q, Li L, Nan XY, He PA, Zhang YZ. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation. J Comput Chem. 2010 Apr 15;31(5):1045-52. doi: 10.1002/jcc.21391. PMID: 19777597.

49. Yu C, Cheng SY, He RL, Yau SST. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene. 2011; 486:110.

50. Zhang YP, Ruan JS, He PA. Analyzes of the similarities of protein sequences based on the pseudo amino acid composition. Chem Phys Lett. 2013; 590: 239– 244, DOI: 10.1016/j.cplett.2013.10.076

51. Ma T, Liu Y, Dai Q, Yao Y, He PA. A graphical representation of protein based on a novel iterated function system. Phys A. 2014; 403:21- 28. DOI: 10.1016/j.physa.2014.01.067

52. Ping P, Zhu X, Wang L Similarities/dissimilarities analysis of protein sequences based on PCA-FFT. J Biol Syst. 2017; 25:29- 45. DOI: 10.1142/s0218339017500024

53. Mahmoodi-Reihani M, Abbasitabar F, Zare-Shahabadi V. A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties. Phys A. 2018; 510: 477– 485. DOI: 10.1016/j.physa.2018.07.011

54. Mahmoodi-Reihani M, Abbasitabar F, Zare-Shahabadi V. In Silico Rational Design and Virtual Screening of Bioactive Peptides Based on QSAR Modeling. ACS Omega. 2020 Mar 10;5(11):5951-5958. doi: 10.1021/acsomega.9b04302. PMID: 32226875; PMCID: PMC7097998.

55. Yin C, Yau SST. Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. 2008; 223-227.

56. Pal J, Ghosh S, Maji B, Bhattacharya DK. Protein sequence comparison under a new complex representation of amino acids based on their physio-chemical properties. Int J Eng Technol. 2018; 7:181-184.

57. Pal J, Ghosh S, Maji B, Bhattacharya DK. Mathematical Approach to Protein Sequence Comparison Based on Physiochemical Properties. ACS Omega. 2022 Oct 17;7(43):39446-39455. doi: 10.1021/acsomega.2c06103. PMID: 36340165; PMCID: PMC9631895.

58. Ghosh S, Pal J, Cattani C, Maji B, Bhattacharya DK. Protein sequence comparison based on representation on a finite dimensional unit hypercube, Journal of Biomolecular Structure and Dynamics. 2023; DOI: 10.1080/07391102.2023.2268719

59. Zhang Y, Yu X. Analysis of Protein Sequence similarity- 978-1-4244-6439-5/19/$26.00(c) IEEE. 2010.

60. Li C, Xing L, Wang X. 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. BMB Rep. 2008 Mar 31;41(3):217-22. doi: 10.5483/bmbrep.2008.41.3.217. PMID: 18377725.

61. Soumen G, Pal J, Maji B, Bhattacharya DK. A sequential development towards a unified approach to protein sequence comparison based on classified groups of amino acids. International Journal of Engineering & Technology. 2018; 7(2): 678-686.

62. Levitt M, Chothia C. Structural patterns in globular proteins. Nature. 1976 Jun 17;261(5561):552-8. doi: 10.1038/261552a0. PMID: 934293.

63. Nishikawa K, Kubota Y, Ooi T. Classification of Proteins into Groups Based on Amino Acid Composition and Other Characters, II. Grouping into Four Types -*The Journal of Biochemistry. 1993;* 94: 997–1007.

64. Sheridan RP, Dixon JS, Venkataraghavan R, Kuntz ID, Scott KP. Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. Biopolymers. 1985 Oct;24(10):1995-2023. doi: 10.1002/bip.360241011. PMID: 4074850.

65. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem. 1986 Jan;99(1):153-62. doi: 10.1093/oxfordjournals.jbchem.a135454. PMID: 3957893.

66. Klein P, Delisi C. Prediction of protein structural class from the amino acid sequence. Biopolymers. 1986 Sep;25(9):1659-72. doi: 10.1002/bip.360250909. PMID: 3768479.

67. Sun XD, Huang RB. Prediction of protein structural classes using support vector machines. Amino Acids. 2006 Jun;30(4):469-75. doi: 10.1007/s00726-005-0239-0. Epub 2006 Apr 20. PMID: 16622605.

68. Kneller DG, Cohen FE, Langridge R. Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol. 1990 Jul 5;214(1):171-82. doi: 10.1016/0022-2836(90)90154-E. PMID: 2370661.

69. Mao B, Chou KC, Zhang CT. Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins. Protein Eng. 1994 Mar;7(3):319-30. doi: 10.1093/protein/7.3.319. PMID: 8177880.

70. Sternberg MJ, Thornton JM. On the conformation of proteins: an analysis of beta-pleated sheets. J Mol Biol. 1977 Feb 25;110(2):285-96. doi: 10.1016/s0022-2836(77)80073-9. PMID: 845953.

71. Flores TP, Moss DM, Thornton JM Solution phase bio panning method using engineered decoy proteins. Protein Engineering. 1994; 7:31-37.

72. Westhead DR, Hutton DC, Thornton JM. **Trends Biochem Sci.** 1998; Jan;**23**(1):35-6. doi: 10.1016/s0968-0004(97)01161-4

73. Westhead DR, Slidel TW, Flores TP, Thornton JM. Protein structural topology: Automated analysis and diagrammatic representation. Protein Sci. 1999 Apr;8(4):897-904. doi: 10.1110/ps.8.4.897. PMID: 10211836; PMCID: PMC2144300.

74. Gilbert D, Westhead D, Viksna J, Thornton J. A computer system to perform structure comparison using TOPS representations of protein structure- Computers and Chemistry. 2001; 26:23-30.

75. Krasnogor N, Pelta DA. Measuring the similarity of protein structures by means of the universal similarity metric- Bioinformatics. 2004; 20:1015–1021.

76. Chew LP, Kedem K. Finding the consensus shape for a protein family. Algorithmica. 2003; 38(1):115–129.

77. Krasnogor N. Self-generating metaheuristics in bioinformatics: The proteins structure comparison case. J Genet Program Evolv Mach. 2003; 5.

78. Leluk J, Konieczny L, Roterman I. Search for structural similarity in proteins. Bioinformatics. 2003 Jan;19(1):117-24. doi: 10.1093/bioinformatics/19.1.117. PMID: 12499301.

79. Gilbert D, Rosselló F, Valiente G, Veeramalai M. Alignment-free comparison of TOPS strings. London Algorithmics and Stringology. 2006; 8:177-197.

80. Gilbert DR, Rossello F, Valiente G, Veeramalai M. Alignment-free comparison of TOPS strings, London Algorithmics and Stringology, Texts. In Algorithmics. 2007; 8: 177-197. Eds. J Daykin, M Mohamed, K Steinhofel. College Publications

81. Liu L, Wang T. Comparison of TOPS strings based on LZ complexity. J Theor Biol. 2008 Mar 7;251(1):159-66. doi: 10.1016/j.jtbi.2007.11.016. Epub 2007 Nov 21. PMID: 18166201.

82. Li B, Li YB, He HB. LZ complexity distance of DNA sequences and its application in phylogenetic tree reconstruction. Genomics Proteomics Bioinformatics. 2005 Nov;3(4):206-12. doi: 10.1016/s1672-0229(05)03028-7. PMID: 16689687; PMCID: PMC5172548.

83. Zhang S, Yang L, Wang T. Use of information discrepancy measure to compare protein secondary structures. Journal of Molecular Structure: THEOCHEM. 2009; 909(1-3):102-106.

84. Guo Y, Wang TM. A new method to analyze the similarity of protein structure using TOPS representations. J Biomol Struct Dyn. 2008 Dec;26(3):367-74. doi: 10.1080/07391102.2008.10507251. PMID: 18808202.

85. Pal D, Dey S, Ghosh P, Bhattacharya DK, Das S, Maji B. A Unique Approach for Protein Secondary Structure Comparison under TOPS Representation - To appear in Journal of Bio molecular Structure and Dynamics- Taylor & Francis. 2004.

86. Zhang S, Yang L, Wang T. Use of information discrepancy measure to compare protein secondary structures. Journal of Molecular Structure: THEOCHEM. 2009; 909(1-3):102-106.

87. Liu L, Wang T. 2D representation of protein secondary structure sequences and its applications. J Comput Chem. 2006 Aug;27(11):1119-24. doi: 10.1002/jcc.20430. PMID: 16721724.

88. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 1993 Jul 20;232(2):584-99. doi: 10.1006/jmbi.1993.1413. PMID: 8345525.

89. Zhang CT, Zhang R. S curve, a graphic representation of protein secondary structure sequence and its applications. Biopolymers. 2000 Jun;53(7):539-49. doi: 10.1002/(SICI)1097-0282(200006)53:7<539::AID-BIP2>3.0.CO;2-2. PMID: 10766950.