International Journal of Engineering &Technology, 7 (1) (2018) 181-184

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i1.9292
Research paper

Protein sequence comparison under a new complex representation of amino acids based on their physio-chemical properties

Jayanta Pal [1]*, SoumenGhosh[2], BansibadanMaji[3], Dilip Kumar Bhattacharya [4]

[1]*Department of Computer Science & Engineering, Narula Institute of Technology, Kolkata, India*
[2]*Department of Information Technology, Narula Institute of Technology, Kolkata, India*
[3]*Department of Electronics & Communication Engineering, National Institute of Technology, Durgapur, India*
[4]*Department of Pure Mathematics, Calcutta University, Kolkata, India*
**Corresponding author E-mail: jayantapal1978@yahoo.com*

Abstract

The paper first considers a new complex representation of amino acids of which the real parts and imaginary parts are taken respectively from hydrophilic properties and residue volumes of amino acids. Then it applies complex Fourier transform on the represented sequence of complex numbers to obtain the spectrum in the frequency domain. By using the method of 'Inter coefficient distances' on the spectrum obtained, it constructs phylogenetic trees of different Protein sequences. Finally on the basis of such phylogenetic trees pair wise comparison is made for such Protein sequences. The paper also obtains pair wise comparison of the same protein sequences following the same method but based on a known complex representation of amino acids, where the real and imaginary parts refer to hydrophobicity properties and residue volumes of the amino acids respectively. The results of the two methods are now compared with those of the same sequences obtained earlier by other methods. It is found that both the methods are workable, further the new complex representation is better compared to the earlier one. This shows that the hydrophilic property (polarity) is a better choice than hydrophobic property of amino acids especially in protein sequence comparison.

Keywords:*Complex Representation; DFT; Hydrophobicity Properties; Hydrophilicity (Polarity) Property; ICD; Phylogenetic Tree; Voss Representation.*

1. Introduction

Digital signal processing based DNA sequence analysis considers the mapping of DNA alphabets into digital signals. Perhaps the most widely used mapping scheme for this purpose is the Voss representation [1], however many other schemes have also been introduced, such as the tetrahedron [2], Z-curve [3], complex [4], [5], [6], quaternion [7], electron-ioninteraction potentials (EIIP) [8], real-numbers [5], [9], [10]. Binary representation of genome sequences refers to the Voss type of representation [1], where the nucleotides A, T, C, G are represented by the 4-component vectors 1, 0, 0, 0; 0, 1, 0, 0; 0,0,1,0 and 0,0,0,1 respectively. Based on such representation of nucleotides, DFT based analysis of genome sequences were made in [11] and [12]. Natural question is to see whether such Voss type of representation is possible for amino acid sequences and if so, whether DFT based analysis could also be made for comparison of Protein sequences based on such binary representations. Such generalization of Voss type of representation has been made recently in [13]. Its usefulness has also been shown in [14]. So it is to be seen whether it is possible to find protein sequence comparison based on such Voss type of representation of amino acids. Very recently the present author [15] applied DFT based analysis on such representation of Protein sequences to obtain their classifications. Next type of interesting representation of DNA sequences is the complex one. In [6] the author used the complex representation of the nucleotides by the numbers 1, -1, +i, -i. A DNA sequence is investigated by using a family of wavelets. The existence of a fractal shape, patterns and symmetries are shown. In [16] the author considers 4 arbitrary complex numbers given by 1 +i, 1 –i, -1 +i, -1 –i corresponding to 4 nucleotides A, T, C, G.DFT based analysis is not applied in DNA sequence comparison under this complex representation. But fractal analysis on the corresponding DNA walk is effective in this case.In this connection it may be mentioned that in [17], the authors used some complex representation of amino acids. The amino acids were represented by 20 complex numbers, whose real parts and imaginary parts were taken from the hydrophobicity properties [18] and from residue volumes [19] of amino acids. Again it is known that there is another important property of amino acids, viz., hydrophilicity (polarity) property. So it is worth considering a new complex representation of amino acids based on their polarity property and residue volumes. It may be noted that, complex representation of DNA sequences and that of Protein sequences differ significantly. The former is given by arbitrary complex numbers, but in the latter case, the complex numbers are derived from physio-chemical properties of amino acids. Any way in [17] complex representation of amino acids was used in connection with periodicity analysis of DNA sequences. First of all, 3-mer arrangements were made to the DNA sequences to convert them in sequences of amino acids. Then the above complex representations of amino acids were applied to represent the DNA sequences as a sequence of complex numbers. Lastly based on such complex representations of DNA sequences, periodicity analysis of DNA sequences was performed. But protein sequence comparison was not performed based on such complex representa-

Copyright © 2018Jayanta Pal et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

tion of amino acids.So there is a possibility to use such complex representation in Protein sequence comparison. Also it is logical to carry on protein sequence comparison with the new type of complex representation and to study relative advantage and disadvantage of protein sequence comparison based on the aforesaid two types of complex representations of amino acids. The objective of the present paper is to investigate in this direction.

2. Methodology

2.1. Complex representation of amino acids under hydrophobicity and residue volumes [17]

Amino Acid	Numerical Representation
Ala (A)	0.61 + 88.3i
Cys (C)	1.07 + 112.4i
Asp (D)	0.46 + 110.8i
Glu (E)	0.47 + 140.5i
Phe (F)	2.02 + 189.0i
Gly (G)	0.07 + 60.0i
His (H)	0.61 + 152.6i
IIe (I)	2.22 + 168.5i
Lys (K)	1.15 + 175.6i
Leu (L)	1.53 + 168.5i
Met (M)	1.18 + 162.2i
Tyr (Y)	1.88 + 193.0i
Trp (W)	2.65 + 227.0i
Val (V)	1.32 + 141.4i
Pro (P)	1.95 + 122.2i
Asn (N)	0.06 + 125.1i
Gln (Q)	148.7i
Arg (R)	0.60 + 181.2i
Ser (S)	0.05 + 88.7i
Thr(T)	0.05 118.2i

2.2. New complex representation of amino acids under hydrophlliicity and residue volumes:

Amino Acid	Numerical Representation
Ala (A)	1.8 + 88.3i
Cys (C)	-4.5+ 112.4i
Asp (D)	-3.5+ 110.8i
Glu (E)	-3.5+ 140.5i
Phe (F)	2.5+ 189.0i
Gly (G)	-3.5+ 60.0i
His (H)	-3.5+ 152.6i
IIe (I)	-0.4+ 168.5i
Lys (K)	-3.2+ 175.6i
Leu (L)	4.5+ 168.5i
Met (M)	3.8+ 162.2i
Tyr (Y)	-3.9+ 193.0i
Trp (W)	1.9+ 227.0i
Val (V)	2.8+ 141.4i
Pro (P)	-1.6+ 122.2i
Asn (N)	-0.8+ 125.1i
Gln (Q)	-0.7 + 148.7i
Arg (R)	-0.9 + 181.2i
Ser (S)	-1.3+ 88.7i
Thr(T)	4.2+ 118.2i

2.3. Special features of complex fourier transform

The Complex DFT
The forward complex DFT, written in polar form, is given by:

$$X[k] = \frac{1}{N}\sum_{k=0}^{N-1} x[n]e^{-j2k\pi n/N}$$

Both the time domain x [n], and the frequency domain X[k], are arrays of complex numbers, with k and n running from 0 to N-1.
First, the real Fourier transform converts a real time domain signal, x [n], into two real frequency domain signals, ReX[k] &ImX[k]. By using complex substitution, the frequency domain can be represented by a single complex array, X[k]. In the complex Fourier transform, both x [n] & X[k] are arrays of complex numbers.
Second, the real Fourier transform only deals with positive frequencies. That is, the frequency domain index, k, only runs from 0 to N/2. In comparison, the complex Fourier transform includes both positive and negative frequencies. This means k runs from 0 to N-1. The frequencies between 0 and N/2 are positive, while the frequencies between N/2 and N-1are negative. Remember, the frequency spectrum of a discrete signal is periodic, making the negative frequencies between N/2 and N-1 the same as between -N/2 and 0. The samples at 0 and N/2 straddle the line between positive and negative.
Third, in the real Fourier transform with substitution, a j was added to the sine wave terms, allowing the frequency spectrum to be represented by complex numbers. To convert back to ordinary sine and cosine waves, we can simply drop the j. This is the sloppiness that comes when one thing only represents another thing. In comparison, the complex DFT is a formal mathematical equation with j being an integral part. In this view, we cannot arbitrarily add or remove a j any more than we can add or remove any other variable in the equation.
Fourth, the real Fourier transform has a scaling factor of two in front, while the complex Fourier transform does not.

2.4. Method of 'inter coefficient distances' under complex fourier transform

First we remember the method of 'Inter coefficient distances' (ICD)as applied to real Fourier transform.
20 real values are associated with 20 amino acids of the protein sequences and thereby a real valued sequence is obtained. Now FFT is applied on this real sequence to get the corresponding spectrum. From this spectrum the amplitudes of the spectrum are calculated. From the values of these amplitudes, a sequence of values is calculated based on the absolute differences of the succeeding term from the preceding one. Thus from the sequence of length n, say, first of all (n/2) +1 = m+1, say, number of amplitudes of the spectrum are determined and finally a sequence, 'a' consisting of m number of amplitudes based on the differences is obtained. This sequence 'a' is now normalized by dividing each element with $\|a\|$ and thus the final normalized sequence of length m is determined. This final sequence of real numbers of length m is taken as the descriptor for sequence comparison.
For complex sequences the method is slightly different. When represented sequence of complex numbers are subjected to complex Fourier transform, both the time domain x [n], and the frequency domain X[k], are arrays of complex numbers, with k and n running from 0 to N-1. So now the N number of amplitudes of the spectrum is obtained. Hence a sequence of real numbers 'a' of length N-1 is obtained by taking the difference of each succeeding term from the preceding one. Now this sequence is normalized by dividing each term by $\|a\|$.This final sequence of real numbers of length N-1 is taken as the descriptor for sequence comparison.
Methodology consists of precisely the following steps:

1) To give details of the data sets of different protein sequences.
2) To represent the given protein sequences as sequence of complex numbers by using II.
3) To apply the method of 'Inter coefficient distances' to get the descriptor vectors.
4) To obtain the distance matrices based on such descriptors.
5) To draw the corresponding phylogenetic trees based on representation II.
6) To represent the given protein sequences as sequences of complex numbers by using I.
7) To repeat steps 2 and 3 to get the corresponding phylogenetic trees based on representation I.
8) To compare the phylogenetic trees obtained in steps 4 and 6.
9) To compare both the phylogenetic trees with phylogenetic trees of the same protein sequences obtained earlier by other methods.

10) To discuss relative advantages and disadvantages of the methods.

3. Results and discussion

In this paper two different complex representation is considered; one using hydrophobicity and residue volumes [17] and the new one is under hydrophilicity and residue volumes. The strength of the new representation is verified from the results of comparison, which are almost the same as their known biological similarity. Phylogenetic tree obtained by the use of complex representation of amino acids under hydrophobicity and residue volumes of ND4, ND5 and ND6 are shown in figure 1, figure 3 and figure 5 respectively. Phylogenetic tree obtained by New Complex representation of amino acids under hydrophilicity and residue volumes of ND4, ND5 and ND6 are shown in figure 2, figure 4 and figure 6 respectively.

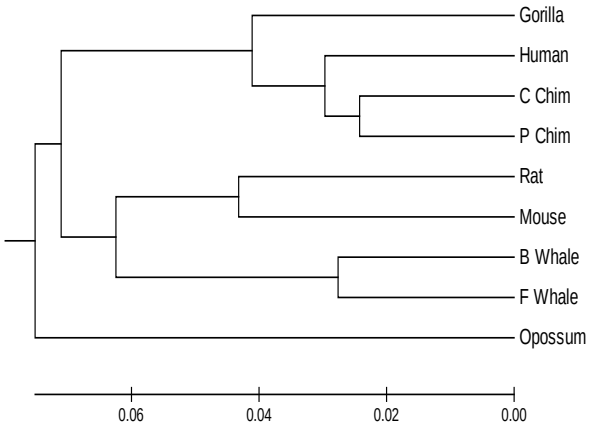


Fig. 1: Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids Under Hydrophobicity and Residue Volumes of ND4.

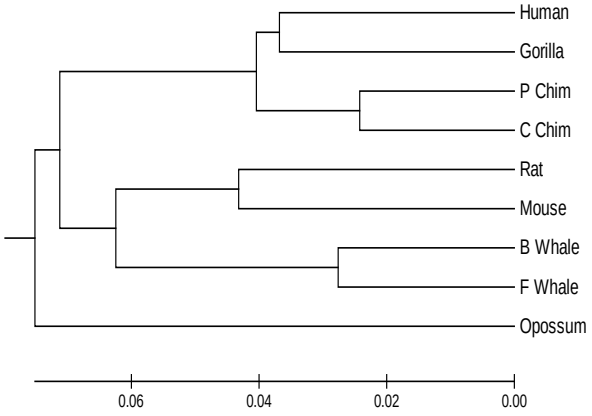


Fig. 2: Phylogenetic Tree Obtained by the Use of New Complex Representation of Amino Acids Under Hydrophilicity and Residue Volumes of ND4.

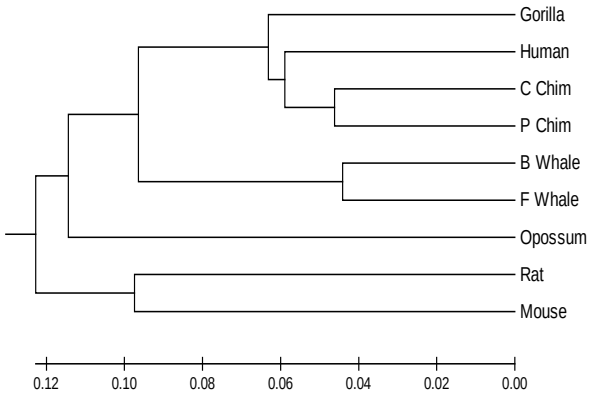


Fig. 3: Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids Under Hydrophobicity and Residue Volumes of ND5.

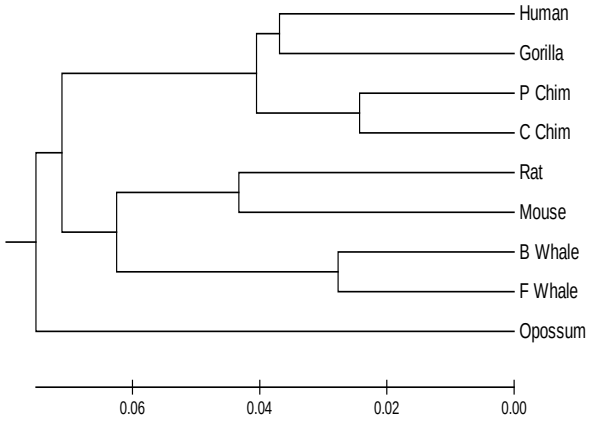


Fig. 4: Phylogenetic Tree Obtained by the Use of New Complex Representation of Amino Acids Under Hydrophilicity and Residue Volumes of ND5.

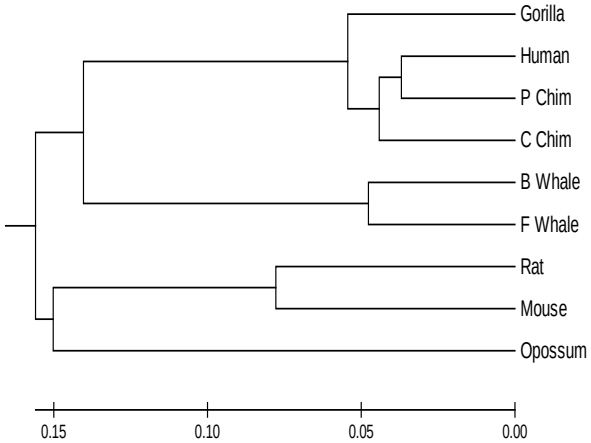


Fig. 5: Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids Under Hydrophobicity and Residue Volumes of ND6.

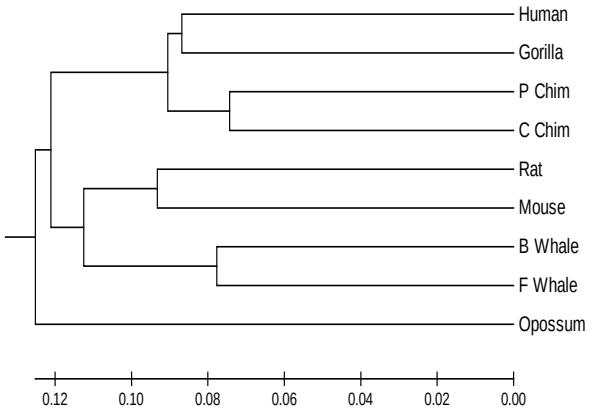


Fig. 6: Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids Under Hydrophilicity and Residue Volumes of ND6.

4. Conclusion

It has been observed that the Phylogenetic tree produced by the new complex representation of amino acids under hydrophilicity and residue volumes are more consistent and almost the same as their known biological similarity. We can conclude that polarity is the best physio-chemical property for representing amino acids numerically for comparison of Protein Sequence.

References

[1] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," Phy. Rev. Lett., vol. 68, no. 25, pp. 3805-3808, June 1992.https://doi.org/10.1103/PhysRevLett.68.3805.

[2] B. D. Silverman, and R. Linsker, "A measure of DNA periodicity," J. Theor. Biol., vol. 118, pp. 295-300, 1986.https://doi.org/10.1016/S0022-5193(86)80060-1.

[3] R. Zhang, and C. T. Zhang, "Z curves, an intuitive tool for visualizing and analyzing the DNA sequences," J. Biomol. Struct. Dyn, vol. 11, no. 4, pp. 767-782, February 1994.https://doi.org/10.1080/07391102.1994.10508031.

[4] D. Anastassiou, "Genomic signal processing," IEEE Signal Proc.Mag., vol. 18, no. 4, pp. 8-20, July 2001.https://doi.org/10.1109/79.939833.

[5] P. D. Cristea, "Genetic signal representation and analysis," in Proc. SPIE Conference, International Biomedical Optics Symposium (BIOS'02), vol. 4623, pp. 77-84, 2002.

[6] Complex Representation of DNA Sequences by Carlo Cattani- M. Elloumi et al. (Eds.): BIRD 2008, CCIS 13, pp. 528–537, 2008._c Springer-Verlag Berlin Heidelberg 2008.

[7] A. K. Brodzik, and 0. Peters, "Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences,"in Proc. IEEE ICASSP, vol. 5, pp. 373-376, 2005.

[8] J. Ning, C. N. Moore, and J. C. Nelson, "Preliminary wavelet analysis of genomic sequences," in Proc. IEEE Bioinformatics Conf (CSB), pp. 509-510, August 2003.https://doi.org/10.1109/CSB.2003.1227391.

[9] G. L. Rosen, "Signal processing for biologically-inspired gradient source localization and DNA sequence analysis," PhD thesis, Georgia Institute of Technology, Aug. 2006.

[10] N. Chakravarthy, A. Spanias, L. D. lasemidis, and K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," EURASIP JASP, vol. 1, pp. 13-28, 2004.

[11] King, B.R., Aburdene, M., Thompson, A. and Warres, Z. (2014) Application of Discrete Fourier Inter-Coefficient Difference for Assessing Genetic Sequence Similarity. EURASIP Journal on Bioinformatics and Systems Biology, 2014, 8.https://doi.org/10.1186/1687-4153-2014-8.

[12] Tung Hoang, Changchuan Yin, HuiZheng, Chenglong YU, Rong Lucy He, Stephen S, T. Tay - A new method to cluster DNA sequences using Fourier power spectrum- Journal of Theoretical Biology- 372 (2015), 135-145.https://doi.org/10.1016/j.jtbi.2015.02.026.

[13] Ghosh, S., Pal, J. and Bhattacharya, D.K. (2014) Classification of Amino Acids of a Protein on the Basis of Fuzzy Set Theory. International Journal of Modern Sciences and Engineering Technology, 1, 30-35.

[14] Ghosh, S., Pal, J. S. Das and Bhattacharya, D.K (2015)-Biological and Theoretical Classifications of Amino Acids in Six Groups. International Journal of Computer Science and Software Engineering, 5, 695-698.

[15] Pal, J., Ghosh, S., Maji, B. and Bhattacharya, D.K. (2016) Use of FFT in Protein Sequence Comparison under Their Binary Representations. Computational Molecular Bioscience, 6, 33-40. https://doi.org/10.4236/cmb.2016.62003.

[16] D. Anastassiou, Frequency-domain analysis of bimolecular sequences, Bioinformatics, vol.16, no.4, pp. 1073-1081, 2000. https://doi.org/10.1093/bioinformatics/16.12.1073.

[17] Changchuan Yin and Stephen S. –T. Yau, Numerical representation of DNA sequences Based on Genetic Code Context and its applications in Periodicity Analysis Genomes- 978-1—1779-7/08/$25.00@2008 IEEE

[18] P. Argos, J.K.M.Rao and P.A.Hargrave, structural prediction of membrane bound proteins, Eur.J.Biochevol.128, pp. 565-575, 1982.https://doi.org/10.1111/j.1432-1033.1982.tb07002.x.

[19] D. E. Godsack and R. C. Chalifoux, Contribution of the free energy of mixing hydrophobic side chains to the stability of the tertiary structure, Journal of Theoretical Biology vol. 39, pp. 645-651, 1973.https://doi.org/10.1016/0022-5193(73)90075-1.