



MOL2NET, International Conference Series on Multidisciplinary Sciences
*USINEWS-04: US-IN-EU Worldwide Science Workshop Series, UMN,
Duluth, USA, 2020*

2D Polar Co-ordinate Representation of Amino Acid Sequences With some applications to Ebola virus, SARS and SARS-CoV-2 (COVID-19)

Tathagata Dey ^{a,e}, Subhamoy Biswas ^{b,e}, Shreyans Chatterjee ^{c,e}, Smarajit Manna ^{d,e},
Ashesh Nandy ^e, Subhas C Basak ^{e,f}

^a Computer Science Department, Government College of Engineering and Textile Technology,
Serampore-712201, India

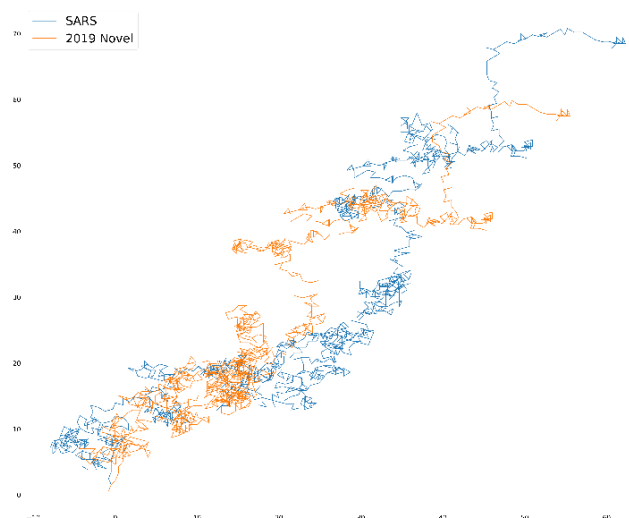
^b Electrical Engineering Department, Jadavpur University, Kolkata-700032, India

^c Microbiology Department, St. Xavier's College, Kolkata-700016, India

^d Jagadis Bose National Science Talent Search, Kolkata 700107, India

^e Centre for Interdisciplinary Research and Education, Kolkata-700068, India

^f Department of Chemistry and Biochemistry, University of Minnesota, Duluth, MN, USA;
sbasak@d.umn.edu

<p>Graphical Abstract</p> 	<p>Abstract.</p> <p>We consider a novel approach to mathematically define a graphing method to represent amino acid sequences of proteins in two-dimensional plane and characterize them numerically. The amino acids are represented by their relative magnitude of their hydrophobicity.</p>
--	---

Introduction

The structure and function of proteins arise from their amino acid (aa) sequences. There have been several approaches to quantify and mathematically characterize protein sequences arising from concepts of graphical representation and numerical characterization of DNA/RNA sequences [1] to novel representation of protein sequences in 2-dimensional plane [2]. Also, in some other approaches a

geometry-based alignment method has also been designed [3]. Our lab also proposed an approach in which a protein sequence was represented in a hypothetical 20D co-ordinate system [4]. Other algorithms including representation in 3D coordinate are also there. [5] [6] [7] [8] [9]

However, there are several issues in this graphical or numerical approach. Plotting amino acid sequences, using some algorithms and rules, can be more effective than plotting base sequences, as in the base sequence there can be deletion or addition of a base, which may lead to a frameshift mutation. In that case the whole genetic code reading frame will be changed, but there will be only a small change in the gene graph which can be hard to identify in a large sequence. While if we plot amino acids, because of its generally small sequence length in comparison with base sequence, one can notice whenever a change of amino acid occurs. It is pertinent to say, in case of a frameshift mutation, in the amino acid graph, from the point of mutation and onwards there will be changes which can be easier to separate from another graph of different time domain. Besides, due to degeneracy of the genetic code, sometimes many triplet codons may code for the same amino acid. So, there may be a change in base sequence graph but protein graph won't change, so it may interpret a closer estimation to the actual structure of the protein. While many multi-dimensional methods to facilitate numerical characterization of protein sequences have been proposed, a 2D representation of amino acid sequences in a Polar Coordinate plane may better help in visualizing sequence changes. Also, by analyzing the differences between two sequences by defining any descriptor or parameter, we can determine phylogenetic relationships.

In a new approach to graphical representation of protein sequences, we represented the 2-dimensional real co-ordinate in polar form ($\mathbb{R} \times \mathbb{R}$) and assigned a particular angle to each amino acid for analyzing their properties. Each aa (amino acid) is represented as a vector and not as a scalar. This enables us in the new representation to visually identify the separation of one protein sequence from another and also quantify intra-sequence and inter-sequence differences. We expect more features to be identified from this approach. Here we report primarily the methodology we have adopted.

AMINO ACID PROPERTIES:

Depending upon the structure an amino acid can be polar or non-polar. Polar amino acids tend to be on the surface most the time as the cellular milieu is mostly polar. And there are amino acids which along with being polar, show acidic or basic character. We give below the classification scheme of the 20 most common amino acids:

- Acidic Amino Acids (aspartate, glutamate) [D,E]
- Basic Acids (Arginine, Histidine, Lysine) [R,H,K]
- Non-Polar AAs [V,P,F,W,M,L,I,G,A]
- Polar Amphoteric AAs [C,S,Y,T,N,Q]

We also characterize all the amino acids based on their Hydrophobicity Index. We consider the data of hydrophobicity found in the world wide web. [10]

Amino Acid	Abbr.	At pH 2 ^A	At pH 7 ^B	Amino Acid	Abbr.	At pH 2 ^A	At pH 7 ^B
Leucine	L	100	100	Phenylalanine	F	92	97
Isoleucine	I	100	99	Tryptophan	W	84	97

Valine	V	79	76	Serine	S	-7	-5
Methionine	M	74	74	glutamine	Q	-18	-10
Cystine	C	52	63	Aspartic Acid	D	-18	-14
Tyrosine	T	49	49	Arginine	R	-26	-23
Alanine	A	47	41	Lysine	K	-37	-28
Threonine	Y	13	13	Asparagine	N	-41	-31
Glutamic Acid	E	8	8	Histidine	H	-42	-46
Glycine	G	0	0	Proline	P	-46	-55

Table 1. Hydrophobicity Index of amino acids

CO-ORDINATE ASSIGNMENT:

We assigned co-ordinates to each of the amino acids using their hydrophobicity index. Assigning the different types of amino acids at different sectors will allow to determine prototype of the graph by simple observation. As there are 20 amino acids, so we define each amino acid at a particular angle in the Cartesian plane. Upon change of amino acid, we move our vector forward. Simple analysis tells that we need to assign each amino acid at a gap of 18°.

So, assigning the amino acids is done in following manner:

The most hydrophobic amino acid is placed in +y axis and the most hydrophilic amino acid is placed at -y axis. Then accordingly we plot the other amino acids with hydrophilicity being vertically downwards.

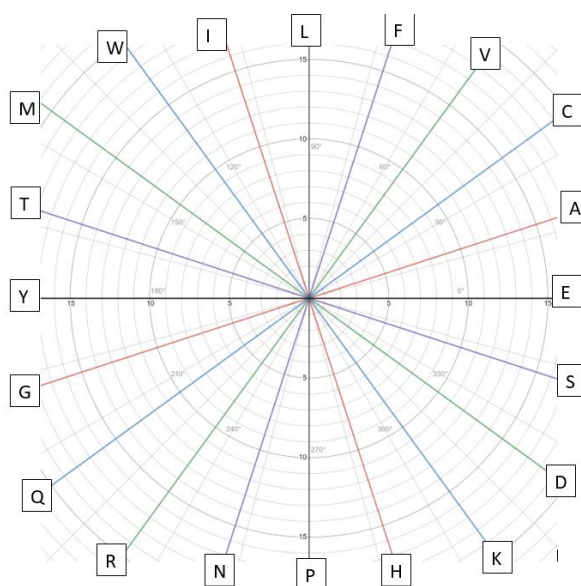


Fig. 1. Assignment of aa in polar co-ordinate

Amino Acid	Abbreviation	Angle(θ)
Alanine	A	18
Cystine	C	36
Aspartic Acid	D	324

Amino Acid	Abbreviation	Angle(θ)
Glutamic Acid	E	198
Phenylalanine	F	72
Glycine	G	36

Histidine	H	288	glutamine	Q	216
Isoleucine	I	108	Arginine	R	234
Lysine	K	306	Serine	S	342
Leucine	L	90	Tyrosine	T	162
Methionine	M	144	Valine	V	54
Asparagine	N	252	Tryptophan	W	126
Proline	P	270	Threonine	Y	180

Table 2. Angle table of amino acids

METHOD OF GRAPH INITIATION:

The graph starts from (0,0) and with each amino acid it moves 1 unit in the direction of that angle.

When the first amino acid is drawn, we move the co-ordinate system then to that point and then calculate for the next amino acid.

For, each turn of amino acids the move in the graph is represented as,

$$\vec{a} = r(\cos \theta \hat{i} + \sin \theta \hat{j})$$

Here, \vec{a} is a vector representing the presentation of an amino acid in the cartesian plane, θ being the angle of that amino acid and as for our system, we did not give any weight to count of amino acids, hence $r = 1$, i.e. that is 1 unit in every direction.

So, if the starting co-ordinate is (0,0) and then an amino acid with angle θ is found, the final co-ordinate should be $(\cos \theta, \sin \theta)$.

GRAPH PROPAGATION:

During each amino acid plotting, we assume the co-ordinate to be shifted to the final point and then plot.

Suppose we take our initial reference frame as S, with origin O at (0,0). Now suppose, the after a stretch of amino acids, we have arrived at a frame S' whose origin O' lies at (p, q) . Now by our convention we assume the co-ordinate system to shift to the point (p, q) .

As we move forward in reading the sequence, we plot the next amino acid, suppose with angle α . So, the next point with respect to the Frame of reference S' should be,

$$(\cos \alpha, \sin \alpha)$$

We suppose this to be, (p', q') .

As we see in the Figure 2., the final coordinate becomes $(p + p', q + q')$.

So, finally it is,

$$(p + \cos \alpha, q + \sin \alpha)$$

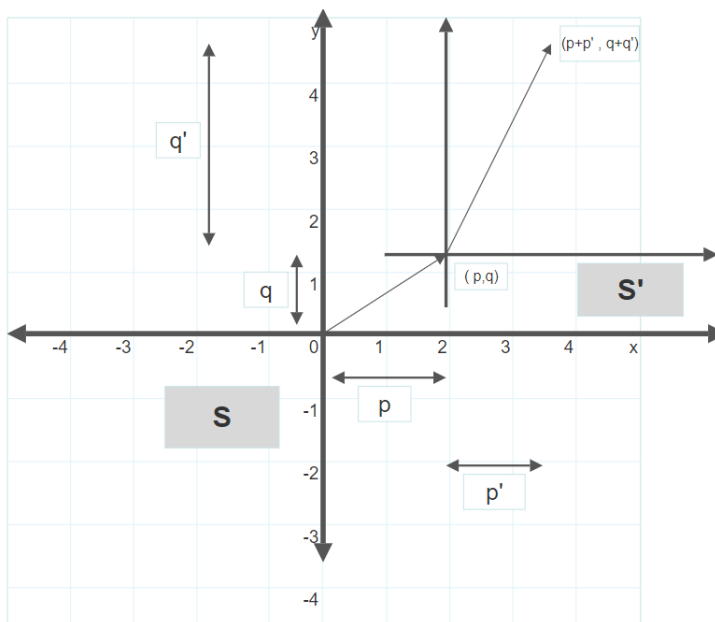


Fig. 2. Placing an amino acid method and coordinate analysis from relative frames of references.

If the next amino acid angle is β , it would be added up too as cos and sin components in x and y coordinate values in terms of the values $\cos \beta$ and $\sin \beta$. So, the final co-ordinate will become,

$$(p + \cos \alpha + \cos \beta, q + \sin \alpha + \sin \beta)$$

So, in this way, whenever a next amino acid is read, we add up its cos value to x co-ordinate and sin value to y co-ordinate.

So, the generalization of the above concept are as follows. Starting a sequence from (0,0) each amino acid is read and plotted with respective angle with the generalized formula given below. Final co-ordinate after placing n sequential amino acids is represented as,

$$\left(\sum_{i=1}^n \cos \theta_i, \sum_{i=1}^n \sin \theta_i \right)$$

NUMERICAL CHARACTERIZATION:

Plotting the sequence in 2D Cartesian Plane will give it a pattern and base for observation, but characterizing it with a signifying mathematical number will help in comparison.

We compute the weighted Centre of Mass of the graph plot and calculate the radius of graph which we name as Quotient Radius (q_R) of that protein.

Centre of Mass is defined as, (μ_x, μ_y) , where

$$\mu_x = \frac{\sum x_i}{N} \quad \mu_y = \frac{\sum y_i}{N}$$

$$q_R = \sqrt{\mu_x^2 + \mu_y^2}$$

Suppose we take an arbitrary sequence as **AWIHPTEFV**.

So, according to Table 1., the angles are as follows,

($18^\circ, 108^\circ, 54^\circ, 0^\circ, 144^\circ, 270^\circ, 126^\circ, 162^\circ$)

AA	x	y
A	0.951056516295153	0.30901699437494
W	0.642039521920206	1.26007351067010
I	1.229824774212679	2.06909050504504
H	2.229824774212679	2.06909050504504
P	1.420807779837732	2.65687575733752
T	1.420807779837731	1.65687575733752
F	0.833022527545258	2.4658927517124
V	-0.118033988749894	2.774909746087417
μ	8.609349685111548	15.261825527610075

Table 2. Short sequence co-ordinate values

$$\mu_x = 8.609349685111548$$

$$\mu_y = 15.261825527610075$$

$$n = 8$$

$$q_R = \sqrt{(8.609349685111548/8)^2 + (15.261825527610075/8)^2}$$

$$q_R = 2.1903346649104676$$

And the graph will be:

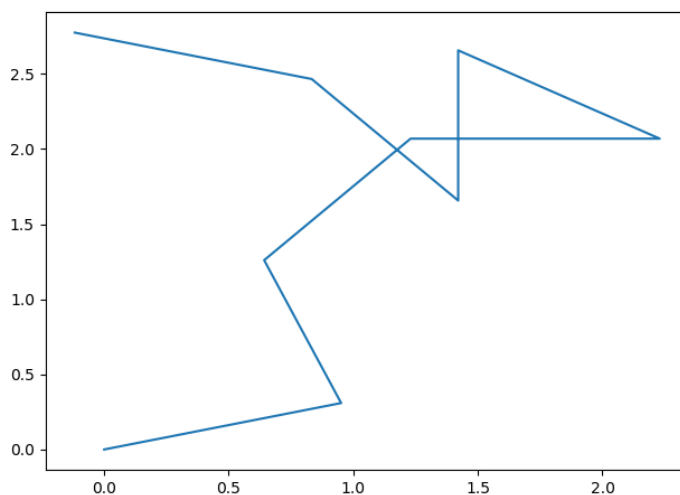


Fig. 3. Short sequence plotting.

PLOTTING BASE AND AMINO ACID OF SAME SEQUENCE:

Figure 4 and 5 represent the base and amino acid sequences of Ebola Virus of years 2014 and 2018. For each of a year's sequence, we computed the Polar Graph and previously used g_R [1]. We plotted the two sequences together superimposing each other and tried to interpret the difference in the two sequences.

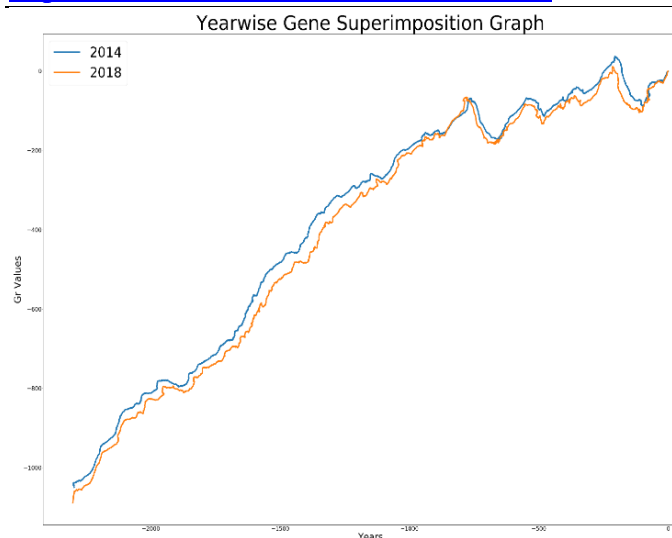


Fig. 4. Superimposition graph of base plotting of 2014 and 2018 Ebola Virus Sequences.

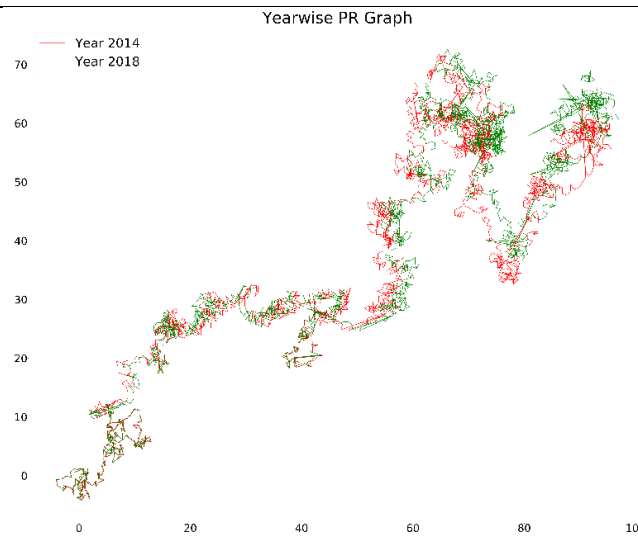


Fig. 5. Superimposition graph of amino acid plotting of 2014 and 2018 Ebola Virus Sequences.

Comparing the two graphs we see that,

- There are around 19000 bases in the Fig. 3., whereas only 2000 amino acids in Fig. 4. So, clearly, the inference can be drawn more clearly in this case, from the changes in the sequence.
- In Figure 4 the sequence seem to be all along of same pattern, just the 2018 sequence, a bit lagging behind. Whereas in Polar plot of Figure 5 it can be figured out that there are a lot more changes in the sequence and hence, the amino acids differ a lot more.
- In (fig. 4.) we can figure out the abundance of polar or non-polar amino acids, regionwise.

In another study we plot the seven types of Corona Virus, namely, 229E alpha, HKU1 beta, NL63 alpha, OC43 beta, MERS, SARS and SARS-CoV-2 (COVID-19) full genomes (Figure 6).

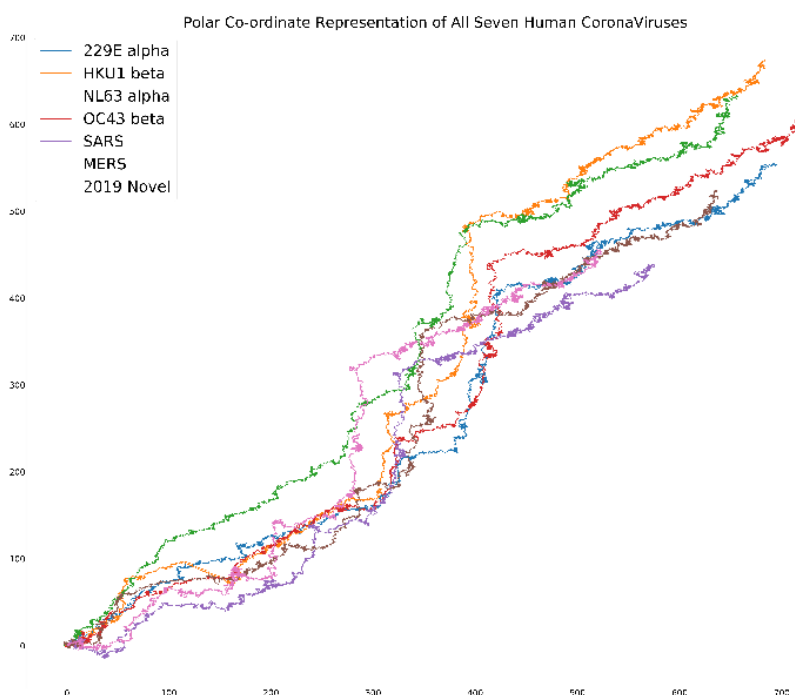


Fig. 6. Superimposed graph of Seven Corona Virus Strains.

This graph gives us some idea about the similarities between SARS and COVID-19. So we plotted the two of them separately.

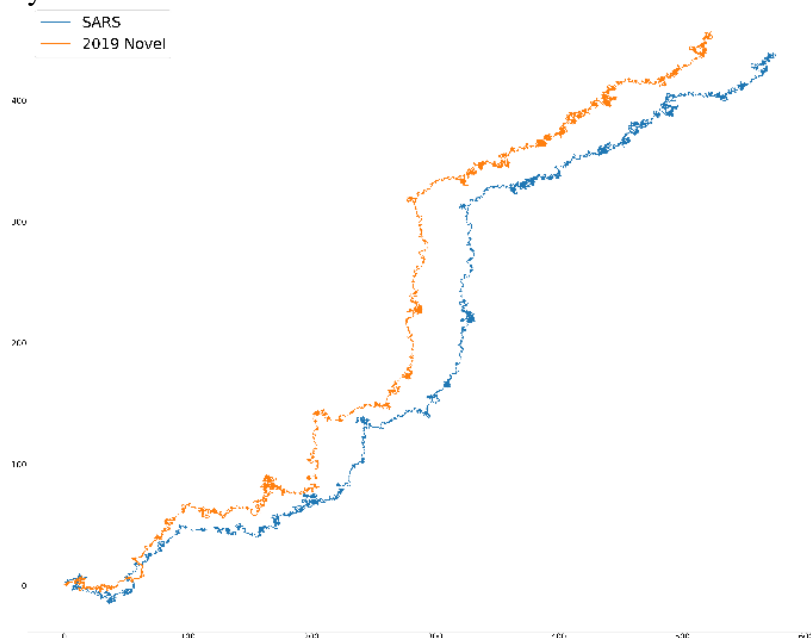


Fig. 7. Superimposition of SARS and Novel Corona Virus Full Genomes.

The parallel nature between them shows the abundance of similar amino acids in the two sequences. To highlight their differences, we plotted their Spike Glycoproteins.

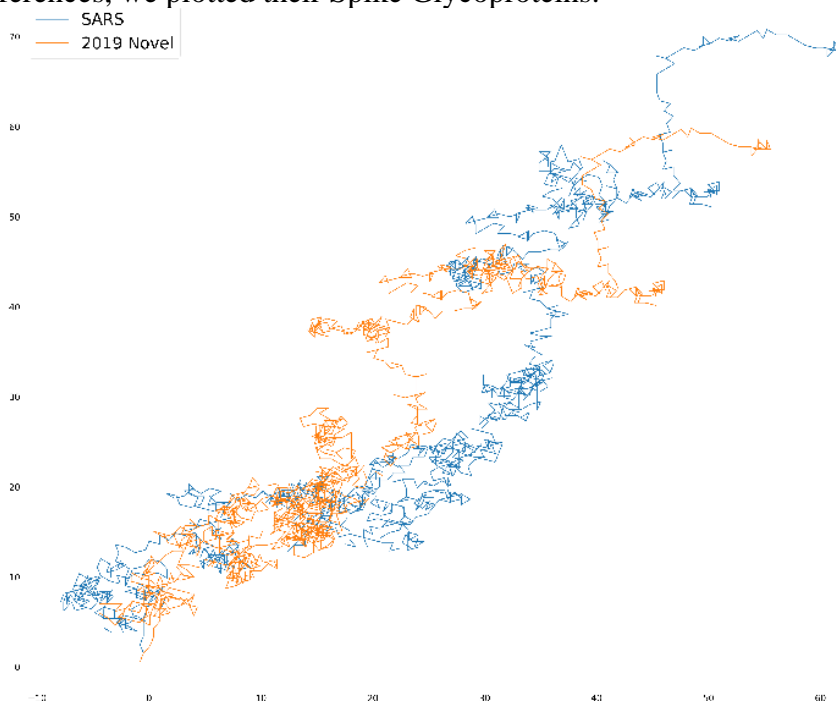


Fig. 8. Superimposition of SARS and Novel Corona Virus Surface Glycoproteins.

This shows the changes that happened in the surface glycoproteins, and also the shifting of clusters and the ending similarities.

PROTEIN FAMILIES:

In another study, we collected various proteins in our sample space. Our goal is to see whether proteins of different families have different representation. To serve this purpose, we collected the following,

- Human Keratin Protein ([CAA73943](#))

- Human Globin Proteins, as Alpha Globin ([ABD95910](#)), Beta Globin ([AAA88054](#)), Gamma Globin ([CAA23771](#)) and Delta Globin ([CAA23763](#)).
- Human Angiotensin I converting enzyme 2 ([ACT66268](#)).
- Human Transferase [partial] ([SBU87545](#)).
- Human Collagen Protein ([BAA04809](#)).
- Human Mitochondrial Protein ([NC012920](#)).^[11]

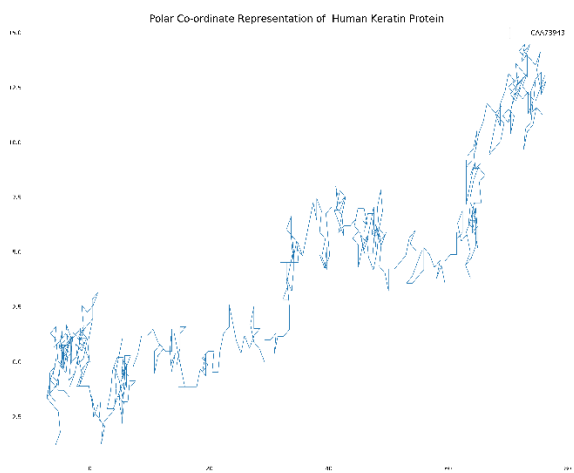


Fig. 9. Polar Coordinate Graph of Human Keratin Protein

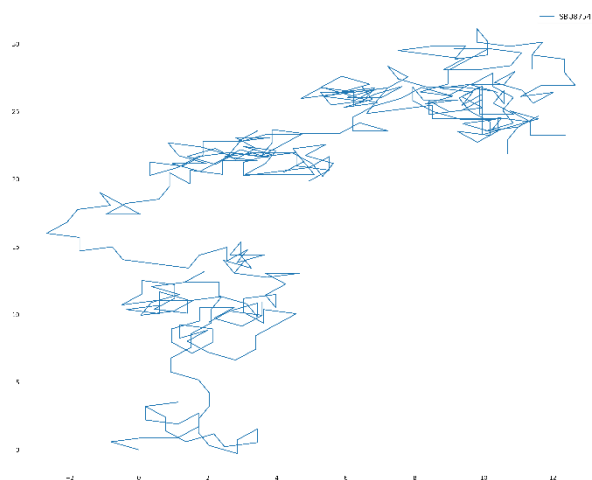


Fig. 12. Human Transferase Protein

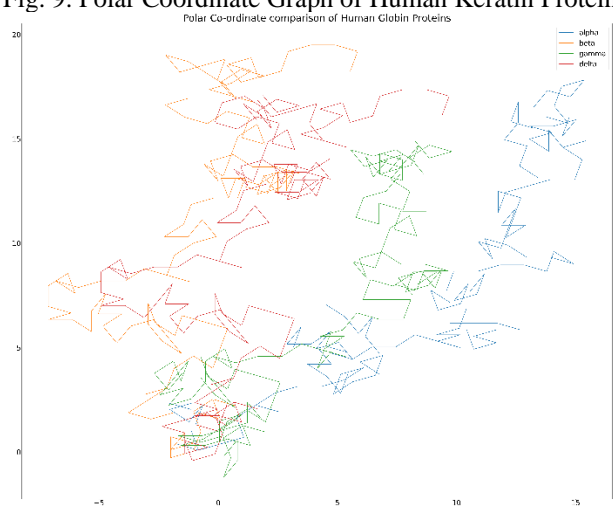


Fig. 10. Superimposition of Human Alpha, Beta, Gamma and Delta Globins.

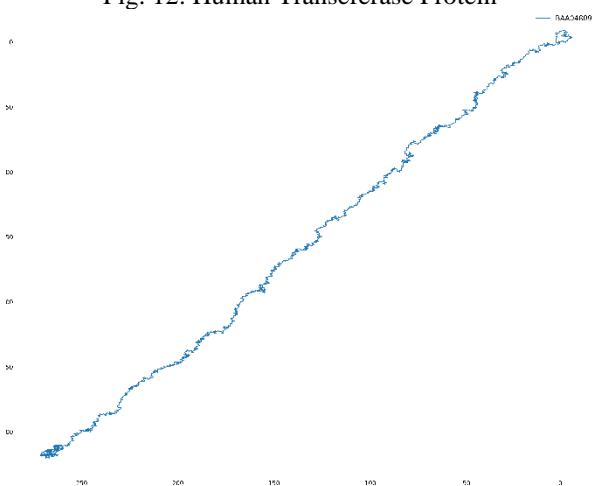


Fig. 13. Human Collagen Protein.

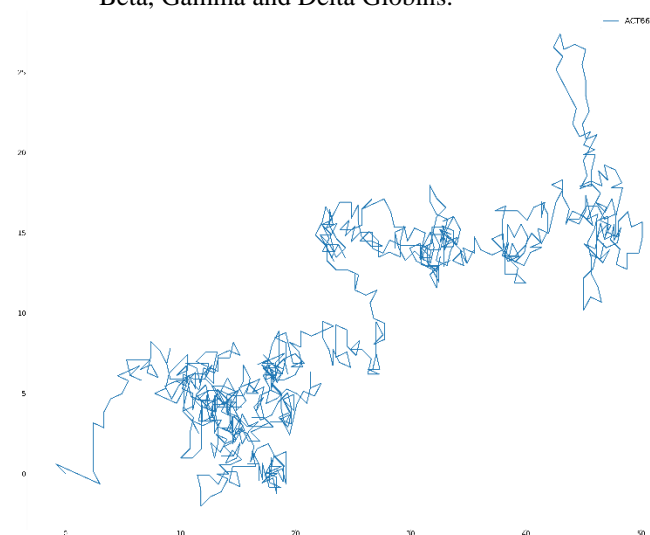


Fig. 11. Human Angiotensin I converting Enzyme 2

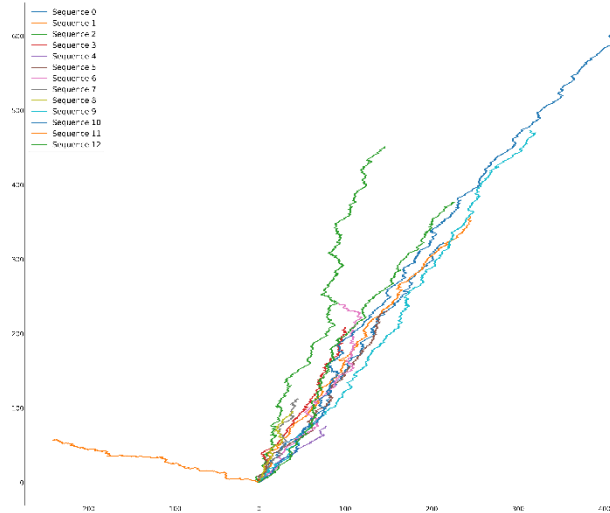


Fig. 14. Human Mitochondrial Protein.

Clearly, we see the comparison of six different protein families and their polar co-ordinate plots. It is pertinent to say, the characteristic types of graph - they all are unambiguous and easily separable from each other. So, we come to the conclusion that this method also provides different characteristic graphs for different protein families.

DISCUSSION:

Here we used our polar plot method to represent graphically various types of proteins including the Ebola Virus, MERS, SARS and SARS-CoV-2 (COVID-19). The preliminary results reported here indicate that our new method is a useful tool for the characterization of different types of proteins.

We are carrying out further studies in order to find relationships between sequence and surface exposure of different parts of protein from their graph. Such results will be published subsequently.

REFERENCES:

- [1] A Nandy, M Harle, S C Basak, Mathematical descriptors of DNA sequences: development and applications, ARKIVOC Vol. **9**, 211-238, 2006.
- [2] Randic M, Zupan J, Balaban AT, Vikić-Topić D, Plavšić D. Graphical Representation of Proteins. *Chem. Rev.*, 2011, 111 (2), 790–862.
- [3] Milan Randić. On a geometry-based approach to protein sequence alignment. *Journal of Mathematical Chemistry*, vol. 43, no. 2, February 2008.
- [4] A Nandy, A Ghosh and P Nandy, Numerical Characterization of Protein Sequences and Application to Voltage-Gated Sodium Channel Alpha Subunit Phylogeny, *In Silico Biology* **9**, 77-87, 2009.
- [5] M. Randić, K. Mehulić, D. Vukićević, T. Pisanski, D. Vikić-Topić, D. Plavšić. Graphical representation of proteins as four-colour maps and their numerical characterization. *Journal of Molecular Graphics and Modelling* 27 (2009) 637–641
- [6] P. He, J. Wei, Y. Yao, Z. Tie. A novel graphical representation of proteins and its application. *Physica A* 391 (2012) 93–99.
- [7] M. M. Reihani, F. Abbasitabar, V. Z. Shahabad. A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties. *Physica A* 510 (2018) 477–485.
- [8] ZH Qi, J Feng, XQ Qi, L Li. Application of 2D graphic representation of protein sequence based on Huffman tree method. *Computers in Biology and Medicine* 42 (2012) 556–563.
- [9] J Wen, YY Zhang. A 2D graphical representation of protein sequence and its numerical characterization. *Chemical Physics Letters* 476 (2009) 281–286.

LINKS:

[10] <https://www.sigmaaldrich.com/life-science/metabolomics/learning-center/amino-acid-reference-chart.html>

[11] <https://www.ncbi.nlm.nih.gov/>