International Journal of Pure and Applied Researches http://ijopaar.com; 2017 Vol. 4(1); pp. 01-08, ISSN: 2455-474X

Similarity analysis of protein sequences based on moment of inertia

K. Lakshmi, Mayukha Pal, and P. Manimaran*

C R Rao Advanced Institute of Mathematics, Statistics and Computer Science,
University of Hyderabad Campus, Hyderabad-50046.

Email: pa.manimaran@gmail.com

Abstract

Studying the protein structure is the basis for spying into its function. Tracing back the evolutionary relationship of proteins is a part of functional annotation. In this paper, we study the sequence similarity of Albumin proteins obtained from different organism. The protein sequences are considered as a rigid body with mass and transformed into vectors by the tensor of moment of inertia. From the calculated Eigen vectors, the Euclidean distance between any two sequences were obtained. The closeness between the organisms is inferred from the constructed phylogenetic tree.

Key words: Protein; Phylogeny; Moment of inertia; Similarity Analysis; Sequences.

1. Introduction

Graphical representation of biological sequences is important among the scientific community. As it allow the visualization of the hidden secrets in the sequence. The graphical representation clearly shows the similarity/differences between closely related sequences, allowing the quantification of the similarity. The first nucleotide graphical representation method was the H-curve proposed by Hamori and Ruskin (1983) followed by the chaos game representation (CGR) by Jeffery (1990). Following these several researchers have published different graphical representations of DNA sequences allowing effective sequence analysis (Jeffery, 1992; Nandy, 1994; Randić et al., 2003; Yau et al. 2003; Liao & Wang, 2004; Liao et al. 2005; Bielińska-Wa et al. 2007; Jafarzadeh & Iranmanesh, 2012; Wąż et al. 2014; Pal et al. 2015a, 2016b). However, the graphical representation of protein sequences emerged only recently as deciphering 20 amino acid proteins was more complicated compared to 4 bases DNA (Randić, 2004; Yau et al. 2008; Wen & Zhang, 2009; Liao et al. 2010; El-Maaty, 2010; He et al. 2012; El-Lakkani & El-Sherif, 2013; Xu et al. 2014; Ma et al. 2014; Li et al. 2014; El-Lakkani & Mahran, 2015). In 2006 a novel 2-D graphical representation using the unit 'magic' circle where the 20 amino acids are positioned equal distances traversing the circle is analogous to the scheme of Jeffrey's CGR representation was published (Randić et al. 2006). And in 2007 a 2-D graphical representation of proteins, where individual nucleic acids are represented as "spots" within a square frame distributed according to specific construction rules was proposed. Later, a cyclic order of 20 amino acids based on the order of 6-bit binary Gary code was introduced (He et al. 2012). Recently, in 2014 researches came up with the idea of using the moments of inertia as new descriptors for representing biological sequences (Wąż & Bielińska-Wąż, 2013a, 2014b). In the present work we have applied the method proposed by W. Hou and co-workers in 2016 where the amino acid is considered as material points with mass, located on a unit circumference on which a 3-D model represented the sequences. Based on the tensor of the moments of inertia was used to calculate the distance matrices of each protein sequence and the Euclidean distance between sequences measures the similarities (Hou et al. 2016). We have applied the method on albumin sequences from different organisms. This was employed to study the relationship among different albumin proteins with that of human. Albumin is one of the important plasma proteins with a typical

blood concentration of 5 g/100 ml. It has been extensively studied over several decades (Kragh-Hansen, 1981; Peters, 1985) for its physiological and pharmacological properties. Apart from maintaining the pH and osmotic pressure of plasma, the protein transports different solutes to the target organs. Albumin binds to a range of ligands starting from water, cations, fatty acids, hormones, bilirubin, thyroxin to pharmaceuticals. This incredible binding property makes albumin an important target in pharmacology (Fiume, 1988). The structure of human serum albumin, as well as serum albumin from other species, has been found to consist of three homologous domains probably derived through gene multiplication (Brown, 1976). Albumin belonging to the family of alpha fetoprotein and vitamin D binding proteins consists of nine double loops of disulfide bonds adjacent half cysteine residues (Carter & Ho, 1994). Proteins are widely used in the taxonomic and phylogenetic studies as it allow comparing the homologous proteins from different taxa. Albumin is one of the important targets in evolutionary studies as it was well studied, abundant, easily purified and stable (Gorman et al. 1971).

Hypoalbuminemia is a medical condition with decreased albumin levels in blood. This may be caused due to inflammation, serious burns, vitamin deficiency, malnutrition, hyperthyroidism and many other factors. As a part of treatment intravenous injections of Albumin (Human) 20% is be administered. The protein is extracted from human and administered to other patients. Apart from hypoalbuminemia albumin therapy is also considered for hypovolemia, burns, surgery or trauma, cardiopulmonary bypass, acute respiratory distress syndrome, hemodialysis, and sequestration of protein-rich fluids (Mendez et al. 2005; Kato et al. 2010).

2. Materials and Methods

In our study, the amino acid sequences of Albumin protein were collected from NCBI website for 10 different organisms. In **Table-1**, we have provided the information regarding the protein sequences its corresponding accession no. and length. In this work, we apply the recently developed method by W. Hou and co-workers to find the evolutionary relationship between any two organisms (Hou et al. 2016). This approach makes use of the physio-chemical properties of amino acids to transform the sequences into vectors by the tensors of the moment of inertia. From the calculated Eigen values the Euclidean distance between any two organisms were computed which provides the sequence similarity.

Table-1: The information of the sequences used

Organism Name	**Accession No**	**Length**
Feliscatus (Cat)	CAA59279.1	608
Bostaurus(Cattle)	AAA51411.1	607
Canis lupus familiaris(Dog)	CAB64867.1	608
Homo sapiens (Human)	AAA98797.1	609
Macacamulatta (Monkey)	NP_001182578.1	608
Susscrofa (Pig)	AAT98610.1	607
Oryctolaguscuniculus(Rabbit)	NP_001075813.1	608
Rattusnorvegicus(Rat)	AAH85359.1	608
Gallus gallusdomesticus	NP_990592.2	615

(Chicken)		
Ovisaries (Sheep)	NP_001009376.1	607

Among the two physicochemical properties hydrophobicity and molecular mass are employed as descriptors to reflect the relationship between proteins. The 20 amino acids were divided into 2 groups based on hydrophobicity: hydrophobic amino acids H = {F, L, I, Y, W, M, V, A, P, C}; hydrophilic amino acids P = {S, N, K, D, R, T, H, Q, E, G}. Then, for a further classification, the amino acids are divided into four types [30] strong hydrophilic amino acids SP = {S, N, K, D, R}; weak hydrophilic amino acids WP = {T, H, Q, E, G}; strong hydrophobic amino acids SH = {F, L, I, Y, W}; weak hydrophobic amino acids = {M, V, A, P, C}.

The four types are of amino acids are arranged in four different quadrants, along the circumference of the circle with unit radius. The hydrophobic amino acids are placed in the first and second quadrant while hydrophilic ones are placed in the third and fourth quadrant. Within each quadrant the ordering is done alphabetically according to their name abbreviations. The 20 points on the circumference of the circle have the coordinates given by $x_i = \cos(2\pi i/20)$, $y_i = \sin(2\pi i/20)$, i = 0, 1, 2 . . . 19. The z-axis coordinate of amino acid is determined by their relative residue weight. According to the weight, 20 amino acids are ranked according to **Table-2**. The z-axis coordinates for smaller molecular mass, are labeled by −1 and other amino acids are labeled by 1. Based on the proposed method we drew the 3-D representation of the albumin protein sequences from 10 organisms including: Feliscatus (Cat), BosTaurus (Cattle),Canis lupus familiaris (Dog), Homosapiens (Human), Macacamulatta (Monkey), Susscrofa (Pig), Oryctolaguscuniculus (Rabbit), Rattusnorvegicus (Rat) Gallus gallusdomesticus (Chicken) and Ovisaries (Sheep).The sequences were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/protein/) the accessions and the sequence length are mentioned in Table 1. The moments of inertia was applied on the 3-D model, considering each amino acid as a material point and mass m=1. The points are distributed as the 3-D Cartesian coordinates. The coordinates of the center of mass of the 3-D graph in the Cartesian coordinate system are defined as

$$\left.\begin{aligned} \mu x &= \frac{\sum_i m_i x_i}{\sum m_i} \\ \mu y &= \frac{\sum_i m_i y_i}{\sum m_i} \\ \mu z &= \frac{\sum_i m_i z_i}{\sum m_i} \end{aligned}\right\} \quad (1)$$

Where xi, yi, zi are the coordinates of material point mi. The tensor of the moments of inertia is defined by the matrix

$$\hat{I} = \begin{bmatrix} Ixy & -Iyy & -Ixz \\ -Iyx & Iyy & -Iyz \\ -Izx & -Izy & Izz \end{bmatrix} \quad (2)$$

Table-2: Amino acids residue weight and z-axis coordinates

Amino acid	**Symbol**	**Residue wt.**	**Z**
Alanine	A	71.80	-1
Cysteine	C	103.14	-1
Methionine	M	131.19	1

Proline	P	97.12	-1
Valine	V	99.13	-1
Phenylalanine	F	147.17	1
Isoleucine	I	113.16	-1
Leucine	L	113.16	-1
Tryptophan	W	186.21	1
Tyrosine	Y	163.18	1
Aspartic Acid	D	115.09	1
Lysine	K	128.17	1
Asparagine	N	114.10	-1
Arginine	R	156.19	1
Serine	S	87.08	-1
Glutamic acid	E	129.12	1
Glycine	G	57.05	-1
Histidine	H	137.14	1
Glutamine	Q	128.13	1
Threonine	T	101.11	-1

The elements of the matrix are defined as

$$\left.\begin{aligned} I_{xx} &= \sum_i m_i((y_i^\mu)^2 + (z_i^\mu)^2), \\ I_{yy} &= \sum_i m_i((x_i^\mu)^2 + (z_i^\mu)^2), \\ I_{zz} &= \sum_i m_i((x_i^\mu)^2 + (y_i^\mu)^2), \\ I_{xy} &= I_{yx} = \sum_i m_i x_i^\mu y_i^\mu, \\ I_{yz} &= I_{zy} = \sum_i m_i y_i^\mu z_i^\mu, \\ I_{xz} &= I_{zx} = \sum_i m_i y_i^\mu z_i^\mu, \end{aligned}\right\} \quad (3)$$

Where $x_i^\mu, y_i^\mu, z_i^\mu$ are the coordinates of m_i in the Cartesian coordinate system for which the origin has been selected at the center of mass. We calculate the eigen values of matrix î

Which is labeled by $\lambda_1,\lambda_2,\lambda_3$. Let us define the vector $\vec{v}^{(s)} = (\lambda_1,\lambda_2,\lambda_3)$ to represent the protein sequence S, we obtain the similarity of two sequence (S^1, S^2) from the Euclidean distance

$$D(S^1, S^2) = \left|\left|\vec{v}(S1) - \vec{v}(S2)\right|\right|_2 \quad (4)$$

Lower the value of D the greater is the similarity between the sequences.

3. Results and Discussion

A rooted phylogenetic tree was constructed from the distance matrix. Figure-1 shows the phylogenetic relationship of albumin sequences among different organisms. This consists of ten organisms and be classified into two major clusters. The first cluster consists of all the mammalian albumin and the second cluster consist of chicken albumin. The clusters are differentiated based on the branch colours. The two clusters clearly indicate the difference in the albumin structure between the class aves and mammalia. Among the mammalian cluster the following pairs have common ancestors:

1) rat & mouse,

2) human & monkey,

3) cat & dog,

4) sheep & cattle.

The branch lengths represent the evolutionary lineages changing over time. Based on this we can infer that sheep and cattle albumin are closer to each other and hence the lower evolutionary rate among the proteins and both the organism fall under the same Order, Artiodactyla (even-toed ungulate). Horse on the other hand falls under a different order Perissodactyla (odd-toed ungulate) all the three falls under the same Clade, Euungulata. Followed, is the rat and mouse pair both categorized under the same subfamily, Murinae. Next closely related is the cat and dog forming same order, Carnivora with different suborders Feliformia and Caniformia respectively. Monkey and human albumin has evolved more compared to the other pairs with common ancestors. A previous study on evolution of trappin genes in mammals also reveals similar evolutionary patterns for the mammals mentioned in the present study (Kato et al. 2010).

Albumin is one of the widely studied proteins for its exclusive binding property. The protein is considered as drug delivery targets. Several evolutionary studies have been conducted on the human serum albumin (HSA) and bovine serum albumin (BSA). This study aims at reviewing the phylogenetic relationship of albumin evolution. A newly proposed method of converting the protein sequences into vectors of moments of inertia and computing the distance matrix was employed. The results from the study were compared with the previous studies to test to authenticity of the method. The comparison gave positive results (Kato et al. 2010). The present method employed in phylogeny generation has proved effective with better computational speed. This proves the method to be reliable for several bioinformatics sequence analysis.

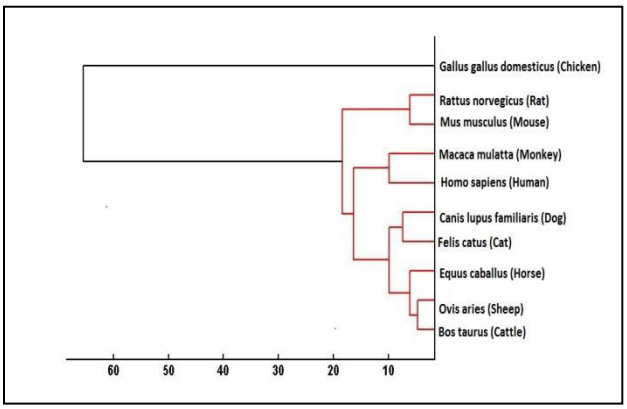

Figure-1: Phylogeny of albumin proteins of eight species

4. Conclusion

As mentioned above albumin is an important plasma protein as it maintains the osmotic pressure of blood. For various conditions external administration of albumin is used as a treatment. Currently used albumin injections are derived from human plasma donors. Due to its wide usage, in future the need for albumin injections may increase and we may require more human donors which may be difficult. As a first step to overcome this we need to find alternative sources of albumin. In this study we have performed phylogenetic analysis on albumin sequences from different closely related organisms. This has revealed the evolutionary pattern for protein sequences. Based on this study further research can be done to identify the closest and the feasible human albumin alternate.

References

1. Bielińska-Wa, D., Clark, T., Wa, P., Nowak, W. & Nandy, A. (2007). 2D-dynamic representation of DNA sequences, Chemical Physics Letters, 442(1), 140-144.
2. Brown, J. R. (1976). Structural origins of mammalian albumin, Federation proceedings, 35(10), 2141-2144.
3. Carter, D. C. & Ho, J. X. (1994). Structure of serum albumin, Advances in protein chemistry, 45, 153-203.
4. El-Lakkani, A. & El-Sherif, S. (2013). Similarity analysis of protein sequences based on 2D and 3D amino acid adjacency matrices, Chemical Physics Letters, 590, 192-195.
5. El-Lakkani, A. & Mahran, H. (2015). An efficient numerical method for protein sequences similarity analysis based on a new two-dimensional graphical representation, SAR and QSAR in Environmental Research, 26(2), 125-137.
6. El-Maaty, M. I. A., Abo-Elkhier, M. M. & Elwahaab, M. A. A. (2010). 3D graphical representation of protein sequences and their statistical characterization, Physica A, 389(21), 4668-4676.
7. Fiume, L., Busi, C., Mattioli, A. & Spinosa, G. (1988). Targeting of antiviral drugs bound to protein carriers. Critical reviews in therapeutic drug carrier systems, 4(4), 265-284.
8. Gorman, G. C., Wilson, A. C. & Nakanishi, M. (1971). A biochemical approach towards the study of reptilian phylogeny: evolution of serum albumin and lactic dehydrogenase, Systematic zoology, 20(2), 167-185.

9. Hamori, E. & Ruskin, J. (1983). H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, The Journal of Biological Chemistry, 258(2), 1318-1327.
10. He, P. A., Li, D., Zhang, Y., Wang, X. & Yao, Y. (2012). A 3D graphical representation of protein sequences based on the Gray code, Journal of Theoretical Biology, 304, 81-87.
11. He, P. A., Wei, J., Yao, Y. & Tie, Z. (2012). A novel graphical representation of proteins and its application, Physica A, 391(1), 93-99.
12. Hou, W., Pan, Q. & He, M. (2016). A new graphical representation of protein sequences and its applications, Physica A, 444, 996-1002.
13. Jafarzadeh, N. & Iranmanesh, A. (2012). A novel graphical and numerical representation for analyzing DNA sequences based on codons, MATCH Communications in Mathematical and in Computer Chemistry, 68(2), 611-620.
14. Jeffrey, H. J. (1990). Chaos game representation of gene structure, Nucleic Acids Research, 18(8), 2163-2170.
15. Jeffrey, H. J. (1992). Chaos game visualization of sequences, Computers & Graphics, 16(1): 25-33.
16. Kato, A., Rooney, A. P., Furutani, Y. & Hirose, S. (2010). Evolution of trappin genes in mammals, BMC evolutionary biology, 10(1), 31.
17. Kragh-Hansen, U. (1981). Molecular aspects of ligand binding to serum albumin, Pharmacological reviews, 33(1), 17-53.
18. Li, Y., Liu, Q., Zheng, X. & He, P. A. (2014). UC-Curve: A highly compact 2D graphical representation of protein sequences, International Journal of Quantum Chemistry, 114(6), 409-415.
19. Liao, B. & Wang, T.M. (2004). New 2D graphical representation of DNA sequences, Journal of Computational Chemistry, 25(11), 1364-1368.
20. Liao, B., Liao, B., Sun, X. & Zeng, Q. (2010). A novel method for similarity analysis and protein sub-cellular localization prediction, Bioinformatics, 26(21), 2678-2683.
21. Liao, B., Zhang, Y., Ding, K. & Wang, T.M. (2005). Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, Journal of Molecular Structure: THEOCHEM, 717(1),199-203.
22. Ma, T., Liu, Y., Dai, Q., Yao, Y. & He, P. A. (2014). A graphical representation of protein based on a novel iterated function system, Physica A, 403, 21-28.
23. Mendez, C. M., Mc Clain, C. J. & Marsano, L. S. (2005). Albumin therapy in clinical practice, Nutrition in clinical practice, 20(3), 314-320.
24. Nandy, A. (1994). A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, Current Science, 309-314.
25. Pal M., Satish, B., Srinivas, K., Madhusudana Rao P., & Manimaran, P. (2015). Multifractal detrended cross-correlation analysis of coding and non-coding DNA sequences through chaos-game representation, Physica A, 436, 596-603.
26. Pal M., Satya Kiran, V., Madhusudana Rao P., & Manimaran P. (2016). Multifractal detrended cross-correlation analysis of genome sequences using chaos-game representation, Physica A, 456, 288-293.
27. Peters, T. (1985). Serum albumin, Advances in protein chemistry, 37, 161-245.

28. Randić, M., Vračko, M., Lerš, N. & Plavšić, D. (2003). Novel 2-D graphical representation of DNA sequences and their numerical characterization, Chemical Physics Letters, 368(1), 1-6.
29. Randić, M. (2004). 2-D graphical representation of proteins based on virtual genetic code, SAR and QSAR in Environmental Research, 15(3), 147-157.
30. Randić, M., Butina, D. & Zupan, J. (2006). Novel 2-D graphical representation of proteins, Chemical Physics Letters, 419(4), 528-532.
31. Waz, P. & Bielinska-Waz, D. (2013). Moments of inertia of spectra and distribution moments as molecular descriptors, MATCH Communications in Mathematical and in Computer Chemistry, 70(3), 851-865.
32. Wąż, P. & Bielińska-Wąż, D. (2014). 3D-dynamic representation of DNA sequences, Journal of Molecular Modeling, 20(3), 2141.
33. Wąż, P., Bielińska-Wąż, D. & Nandy, A. (2014). Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences, Journal of Mathematical Chemistry, 52(1), 132-140.
34. Wen, J. & Zhang, Y. (2009). A 2D graphical representation of protein sequence and its numerical characterization, Chemical Physics Letters, 476(4), 281-286.
35. Xu, S. C., Li, Z., Zhang, S. P. & Hu, J. L. (2014). Primary structure similarity analysis of proteins sequences by a new graphical representation, SAR and QSAR in Environmental Research, 25(10), 791-803.
36. Yao, Y., Yan, S., Xu, H., Han, J., Nan, X., He, P. A. & Dai, Q. (2014). Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation, Evolutionary Bioinformatics, 10, 87-96.
37. Yau, S. S. T., Wang, J., Niknejad, A., Lu, C., Jin, N. & Ho, Y. K. (2003). DNA sequence representation without degeneracy, Nucleic Acids Research, 31(12), 3078-3080.
38. Yau, S. S. T., Yu, C. & He, R. (2008). A protein map and its application, DNA and Cell Biology, 27(5), 241-250.