PROTEIN SIMILARITY / DISSIMILARITY STUDY USING MOMENT VECTOR BY NON-HOMOLOGOUS METHOD

D. Vijayalakshmi
Assistant Professor, Department of Mathematics
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya,Kanchipuram, Tamilnadu

S. Hemalatha
M. Phil Scholar ,Department of Mathematics
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya,Kanchipuram, Tamilnadu

Abstract: Similarity / Dissimilarity of protein is measured using moment vector. The moment vector is constructed in a novel way by using physic-chemical properties of amino acids. Moment vector upto four dimensions are constructed. Euclidean distance between the moment vectors of each pair of protein measures the similarity between proteins.

Keywords: Protein Sequence, Sequence Similarity, Moment Vector.

1. Introduction: Development of sequencing techniques have increased the number of biological sequences in the data bases. Extracting the potential information from these sequences proves to be a challenging task. Extracting the information is essential as they determine the barer for the physicology and anatomy study of organisms. This study can be done by homologous method and non-homologous method.

Some non-homologous method Yao [1] studied the protein sequence based on pk_a value of COOH and NH_3^+ of 20 amino acids graphically. Xiao and Wu[2] represented protein in a 2-D graph based on physicochemical properties Liao et al [3] calculated the similarity or dissimilarity using distance formula and also represented protein sequence in a 2-D graph. Randic [4] represented protein in 2-D graph using the physicochemical property to study the similarity. Feng & Zhang [5] represented protein as Z_p-curve using hydrophobicity of amino acids. Bai and Wang [6] studied protein by 2-D graph using nucleotide triplet codons. Based on 3 physicochemical properties Abo el Maaty et al [7] represented protein in a 3-D graphical method. Yao [8] developed a graphical method to study about protein based on the Pk_a values of COOH and NH_3^+ of a amino acids. Also Xiao and Wu [9] studied protein in 2 dimensional graph using the physicochemical properties. In [10], Liao et al measures the similarity and dissimilarities between proteins using distance formula and also developed a new 2-D graphical representation. In this paper we present a novel moment vector using the primary structure of protein sequence. The moment vector of 4 dimension is obtained for each protein similarity/dissimilarity between protein sequences are measured using Euclidean distance between these moment vector of proteins.

2. Moment Vector: The primary structure of protein – the amino acid sequence plays a vital role in this part. PI_a and Kh values of the amino acids are considered as x and y co-ordinates of the amino acids. Then the moment vector of 1 – dimension is calculated using the formula.

$M_1 = X$

$X= \frac{\Sigma(xi-yi)^1}{n^1}$ where n is the number of amino acids in the sequence.

The moment vector (X, Y) of 2-dimension is calculated using the formula

$M_2 = (X, Y)$, $X= \frac{\Sigma(xi-yi)^1}{n^1}$, $Y=\frac{\Sigma(xi-yi)^2}{n^2}$

The moment vector (X,Y,Z) of 3-dimension is calculated using the formula

M_3=(X,Y,Z) $X=\frac{\sum(xi-yi)^1}{n^1}$, $Y=\frac{\sum(xi-yi)^2}{n^2}$, $Z=\frac{\sum(xi-yi)^3}{n^3}$

M_4=(X,Y,Z,S) $X=\frac{\sum(xi-yi)^1}{n^1}$, $Y=\frac{\sum(xi-yi)^2}{n^2}$, $Z=\frac{\sum(xi-yi)^3}{n^3}$, $S=\frac{\sum(xi-yi)^4}{n^4}$

The Euclidean distance between moment vectors of each pair of proteins are calculated. This Euclidean distance measures the similarity between protein.

Data Used:

Protein ID	Protein Structure	Protein Description
4fc1		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
3ue7		TTCCPSIVARSNXNACRLPGTPEALCATYTGCIIIPGATCPGDYAN
3nir		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
2eyb		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
2eyc		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
2eyd		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
1yv8		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
1yva		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN
1cxr		TTCCPSIVARSNFNVCRLPGTSEAICATYTGCIIIPGATCPGDYAN
1cnr		TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN

Moment Vector:

	Protein-1 (4FC1)	Protein-2 (3UE7)	Euclidean distance
One dimension	X= 5.510652	X=5.524783	0.014130435
Two dimension	X= 5.510652	X=5.524783	0.022164611
	Y=0.872683696	Y=0.855607372	
Three dimension	X= 5.510652	X=5.524783	0.024536339
	Y=0.872683696	Y=0.855607372	
	Z=0.170296821	Z=0.159772471	
Four dimension	X= 5.510652	X=5.524783	0.024968447
	Y=0.872683696	Y=0.855607372	
	Z=0.170296821	Z=0.159772471	
	S=0.039524883	Z=0.034899799	

	Protein-1 (4FC1)	Protein-3 (3NIR)	Euclidean distance
One dimension	X= 5.510652	X= 5.510652	0
Two dimension	X= 5.510652	X= 5.510652	0
	Y=0.872683696	Y=0.872683696	
Three dimension	X= 5.510652	X= 5.510652	0
	Y=0.872683696	Y=0.872683696	
	Z=0.170296821	Z=0.170296821	
Four dimension	X= 5.510652	X= 5.510652	0
	Y=0.872683696	Y=0.872683696	
	Z=0.170296821	Z=0.170296821	
	S=0.039524883	S=0.039524883	

	Protein-2 (3UE7)	Protein-3 (3NIR)	Euclidean Distance
One dimension	X=5.524783	X= 5.510652	0.014130435
Two dimension	X=5.524783	X= 5.510652	0.022164611
	Y=0.855607372	Y=0.872683696	
Three dimension	X=5.524783	X= 5.510652	0.024536339
	Y=0.855607372	Y=0.872683696	
	Z=0.159772471	Z=0.170296821	
Four dimension	X=5.524783	X= 5.510652	0.024968447
	Y=0.855607372	Y=0.872683696	
	Z=0.159772471	Z=0.170296821	
	Z=0.034899799	S=0.039524883	

Result:

Table 1: Result using M1

	PRO 2	PRO 3	PRO 4	PRO 5	PRO 6	PRO 7	PRO 8	PRO 9	PRO 10
PRO 1	0.014130435	0	0	0	0	0	0	0.045217391	0
PRO 2		0.014130435	0.014130435	0.014130435	0.014130435	0.014130435	0.014130435	0.059347826	0.014130435
PRO 3			0	0	0	0	0	0.045217391	0
PRO 4				0	0	0	0	0.045217391	0
PRO 5					0	0	0	0.045217391	0
PRO 6						0	0	0.045217391	0
PRO 7							0	0.045217391	0
PRO 8								0.045217391	0
PRO 9									0.045217391

Table 2: Result using M2

	PRO 2	PRO 3	PRO 4	PRO 5	PRO 6	PRO 7	PRO 8	PRO 9	PRO 10
PRO 1	0.022134611	0	0	0	0	0	0	0.04649459	0
PRO 2		0.022134611	0.022134611	0.022134611	0.022134611	0.022134611	0.022134611	0.05967638	0.022134611
PRO 3			0	0	0	0	0	0.04649459	0
PRO 4				0	0	0	0	0.04649459	0
PRO 5					0	0	0	0.04649459	0
PRO 6						0	0	0.04649459	0
PRO 7							0	0.04649459	0
PRO 8								0.04649459	0
PRO 9									0.04649459

Table 3: Result using M3

	PRO 2	PRO 3	PRO 4	PRO 5	PRO 6	PRO 7	PRO 8	PRO 9	PRO 10
PRO 1	0.024536339	0	0	0	0	0	0	0.046553565	0
PRO 2		0.024536339	0.024536339	0.024536339	0.024536339	0.024536339	0.024536339	0.060234645	0.024536339
PRO 3			0	0	0	0	0	0.046553565	0
PRO 4				0	0	0	0	0.046553565	0
PRO 5					0	0	0	0.046553565	0
PRO 6						0	0	0.046553565	0
PRO 7							0	0.046553565	0
PRO 8								0.046553565	0
PRO 9									0.046553565

Table 4: Result using M4

	PRO 2	PRO 3	PRO 4	PRO 5	PRO 6	PRO 7	PRO 8	PRO 9	PRO 10
PRO 1	0.024968447	0	0	0	0	0	0	0.046556042	0
PRO 2		0.024968447	0.024968447	0.024968447	0.024968447	0.024968447	0.024968447	0.060377089	0.024968447
PRO 3			0	0	0	0	0	0.046556042	0
PRO 4				0	0	0	0	0.046556042	0
PRO 5					0	0	0	0.046556042	0
PRO 6						0	0	0.046556042	0
PRO 7							0	0.046556042	0
PRO 8								0.046556042	0
PRO 9									0.046556042

Conclusion: The similarity / dissimilarity of protein is studied using non-homologous moment vector. This is a novel method of constructing moment vector using physicochemical properties of amino acids. This method yields a result close to the results obtained by Blast sequence site. This also shows that as the dimensions of moment vector increase the accuracy of result obtained also increase.

References:

1. Yao YH, Dai Q, Li C, He PA Nan XY, Zhang YZ. Analysis of similarity / Dissimilarity of protein sequences. Proteins 2008;73:864-871.

2. Wu ZC, Xiao X, Chou KC. 2D-MH. A web server for generating graphic representation of protein sequence based on the physicochemical properties of their constituent amino acids. J.Theor, Biol.2010 267:29-34
3. Liao B, Liao B Y, sum XM, Zeng QG. A novel method for similarity analysis and protein subcellular localization prediction. Bioinformatics 2010; 20 : 2678 – 2683.
4. Randic M. 2D – graphical representation of protein based on physicochemical properties of amino acids, chemical Physics Letter 2007; 444 176 – 180
5. Z.P.Feng and C.T. Zhang, "A graphic representation of protein sequence and predicting the subcellulae locations of prokaryotic proteins. "International journal of Biochemistry and cell biology, Vol 34, No 3 . pp. 298 – 207, 2002
6. F. Bai and T. Wang "A 2-D graphical representation of protein sequence based on nucleotide triplet codons". Chemical Physics letters . vol 413 no: 4-6. pp 458-462, 2005
7. M.I. Abo el Maaty M.M.Abo Elkhier and M.A.Abd Elwahaab, "3D graphical representation of protein sequences and their statistical characterization". Physical A : Statistical Mechanics and its applications vol. 389 no 21 PP 4668 – 4676 2010.
8. Yao YH, Dai Q, Lic, He PA, Nan XY, Zhang YZ. Analysis of similarity / dissimilarity of protein sequences. Proteins 2008; 73: 864 – 871.
9. WuZC, 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 boil 2010 ; 267 : 29 – 34.
10. Liao B, Liao BY, Sun XM, Zeng QG. A novel method for similarity analysis and protein subcellular localization prediction. Bioinformatics 2010. 26 : 2678 – 2683.

***