Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Differentiation of Protein Sequence Comparison Based on Biological and Theoretical Classification of Amino Acids in Six Groups

Soumen Ghosh*
Department of Information Technology
Narula Institute of Technology
India

Subhram Das
Department of Computer Science & Engineering
Narula Institute of Technology
India

Jayanta Pal
Department of Computer Science & Engineering
Narula Institute of Technology
India

Dilip Kumar Bhattacharya
Department of Pure Mathematics
Calcutta University
India

Abstract— The paper tries to differentiate results of protein sequence comparisons based on classification of amino acids in six groups, which are obtained either biologically or theoretically. In fact, biological classification of such groups of amino acids is known from their side chain properties. But recently same type of classification of amino acids in six groups with different constituent members has been obtained from similarity properties of amino acids, where these properties are determined theoretically from fuzzy representations of the amino acids. Natural query is to see whether there is any similarity or dissimilarity of the results of comparisons of same protein sequences based on the above different groups of classification. The paper proves that the results of comparisons are almost uniform in both the cases. Moreover in the later case the results are finer than those of the former one.

Keywords— Amino acids, Protein sequences, Biological classification of amino acids, Theoretical classification of amino acids, Protein sequence comparison

I. INTRODUCTION

In DNA or RNA sequences triplet of nucleotides are called codons. Each of the codons represents an amino acid and instructs the cell machinery to produce the corresponding amino acid during the Translation phase of protein synthesis. Thus a protein is a linear chain of amino acids, which starts with a start codon ATG, which corresponds to the amino acid methionine, followed by a sequence of amino acids and ends with a stop codon. The amino acid sequence that makes a protein is called its primary structure. It is believed that the dynamical folding process and stable structure of a protein are determined by its primary structure. Thus the prediction of secondary and space structures of protein from its primary structure is a challenging problem. Now among the numerous available amino acids only 20 are generally found in living beings and every protein sequence is expressed by these 20 amino acids. Naturally protein sequence comparison also follows the same approach as is considered in genome sequence analysis. In details first of all, numerical representation of the protein sequences are obtained from the numerical values given to the individual amino acids, then graphical representation of the protein sequences are obtained; from these graphs descriptors are derived. These are finally used in comparing protein sequences. Such methods of comparison are found in [1, 2, 3]. As amino acids are 20 in number, it is better to classify them in different groups as far as possible and proceed for the analysis of protein sequence comparison on the basis of such classified groups, each with lesser number of amino acids than the whole set of 20 amino acids. This is another approach to protein sequence comparison. Hopefully such groups of amino acids are already known [4, 5, 6, 7] and they are obtained based on the biological properties of the amino acids. Such classifications are known to consist of two groups, three groups, four groups, five groups, six groups and seven groups. The analysis of protein sequences based on different groups of amino acids has already been tried except for group of three, group of six and group of seven. Interestingly a theoretically classified group of six has been derived very recently by Soumen Ghosh et. Al., [8]. Thus two types of classification of six groups of amino acids are now known, one is obtained biologically and other is obtained theoretically. Naturally it is required to see whether both the classified groups of six are useful in protein sequence comparison. If so, to find which one can give comparatively better results. This is the motivation behind writing this paper.

TABLE I LIST OF AMINO ACIDS WITH ABBREVIATIONS AND CORRESPONDING DNA CODONS

Amino Acid	SLC	DNA Codons		Amino Acid	SLC	DNA Codons
Isoleucine	I	ATT, ATC, ATA		Tyrosine	Y	TAT, TAC
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG		Tryptophan	W	TGG
Valine	V	GTT, GTC, GTA, GTG		Glutamine	Q	CAA, CAG

Phenylalanine	F	TTT, TTC		Asparagine	N	AAT, AAC
Methionine	M	ATG		Histidine	H	CAT, CAC
Cysteine	C	TGT, TGC		Glutamic acid	E	GAA, GAG
Alanine	A	GCT, GCC, GCA, GCG		Aspartic acid	D	GAT, GAC
Glycine	G	GGT, GGC, GGA, GGG		Lysine	K	AAA, AAG
Proline	P	CCT, CCC, CCA, CCG		Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Threonine	T	ACT, ACC, ACA, ACG		Stop codons	Stop	TAA, TAG, TGA
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC				

II. METHODOLOGY

Let there be m classified groups G_1, G_2,.., G_m containing n_1, n_2,.., n_m number of amino acids respectively, where $n_1+n_2+..+n_m = 20$. In a protein time series of length N, let the frequencies of number of occurrences of amino acids of G_1, G_2, .. , G_m be respectively given by

$$f_{11}, f_{12},.., f_{1n_1}; f_{21}, f_{22},.., f_{2n_2}; f_{31}, f_{32},.., f_{3n_3};...; f_{m1}, f_{m2},.., f_{mn_m}$$

In the newly represented protein sequence where all members of Gi are replaced by Gi, i= 1,2,..,m, let the frequencies of number of occurrences of Gi, i = 1,2,..,m be gi, i = 1,2,..,m. Similarly let the frequencies of number of occurrences of Gi taken two at a time and the frequencies of number of occurrences of G_i taken two at a time with any amino acid lying in between the combination be respectively given by

$$h_1, h_2,....., h_{m^2}; w_1, w_2,..., w_{m^3}$$

Then the descriptor for protein sequence comparison is taken as a $20+m+m^2+m^2$ component vector given by

$$\left(\frac{f_{11}, f_{12},.., f_{1n_1}; f_{21}, f_{22},.., f_{2n_2}; f_{31}, f_{32},.., f_{3n_3};...; f_{m1}, f_{m2},.., f_{mn_m}}{N}; \frac{g_1, g_2,....., g_m}{N}; \frac{h_1, h_2,....., h_{m^2}}{\sum_{i=1}^{m^2} h_i}; \frac{w_1, w_2,..., w_{m^3}}{\sum_{i=1}^{m^2} w_i} \right)$$

Now two protein sequences are compared by finding the distance matrix based on such descriptors and obtaining corresponding phylogenetic trees.

Illustration of the methodology

Let us illustrate the procedure for six classified groups (*A, B, C, D, E, F*) by a simple artificial example of a protein sequence of 20 amino acids (SYPHYVKSIVASTFIISLFP). Here group *A* represents the amino acids of I, *B* represents L,R, *C* represents V,A,G,P,T, *D* represents F,C,Y,Q,N,H,E,D,K, *E* represents M,W and *F* represents S. Here m=6. After calculating the number of occurrences of each amino acids we got the first twenty components {3,1,2,2,0,0,1,0,2,1,4,2,0,0,0,1,0,0,1,0}. The sequence of the amino acids of the example is converted to *FDCDDCDFACCFCDAAFBDC*. Next six components are the number of occurrences of each groups i.e. {3,1,6,6,0,4}. Next m^2 i.e. 36 components are the number of occurrences of two pair combinations of six groups. Here number of occurrences of *AA* = 1, *AB* = 0, *AC* = 1, *AD* = 0, *AE* = 0, *AF* = 1, *BA* = 0, *BB* = 0, *BC* = 0, *BD* = 1, *BE* = 0, *BF* = 0, *CA* = 0, *CB* = 0, *CC* = 1, *CD* = 3, *CE* = 0, *CF* = 1, *DA* = 1, *DB* = 0, *DC* = 3, *DD* = 1, *DE* = 0, *DF* = 1, *EA* = 0, *EB* = 0, *EC* = 0, *ED* = 0, *EE* = 0, *EF* = 0, *FA* = 1, *FB* = 1, *FC* = 1, *FD* = 1, *FE* = 0 and *FF* = 0. So the 36 components are {1,0,1,0,0,1,0,0,0,1,0,0,0,0,1,3,0,1,1,0,3,1,0,1,0,0,0,0,0,0,1,1,1,1,0,0}. Next m^2 components are the number of occurrences of two pair combinations of six groups after taken two at a time with any groups lying in between the combination. Here number of occurrences of *AXA* = 3, *AXB* = 0, *AXC* = 0, *AXD* = 0, *AXE* = 0, *AXF* = 0, *BXA* = 0, *BXB* = 0, *BXC* = 1, *BXD* = 0, *BXE* = 0, *BXF* = 0, *CXA* = 1, *CXB* = 0, *CXC* = 1, *CXD* = 1, *CXE* = 0, *CXF* = 2, *DXA* = 2, *DXB* = 0, *DXC* = 1, *DXD* = 2, *DXE* = 0, *DXF* = 0, *EXA* = 0, *EXB* = 0, *EXC* = 0, *EXD* = 0, *EXE* = 0, *EXF* = 0, *FXA* = 0, *FXB* = 0, *FXC* = 2, *FXD* = 2,*FXE* = 0 and *FXF* =0. So the 36 components are {3,0,0,0,0,0,0,0,1,0,0,0,1,0,1,1,0,2,2,0,1,2,0,0,0,0,0,0,0,0,0,0,2,2,0,0}.

III. DETAILS OF PROTEIN SEQUENCES FOR COMPARISON

TABLE II DATABASE SOURCE OF NINE ND5 PROTEINS

Sl. No.	Species	ID/ACCESSION	Database	Length
Seq1	Human (Homo sapiens)	AP-000649	NCBI	603
Seq2	Gorilla(Gorilla gorilla)	NP-008222	NCBI	603
Seq3	Common Chimpanzee (Pan troglodytes)	NP-008196	NCBI	603
Seq4	Pigmy Chimpanzee (Pan paniscus)	NP-008209	NCBI	603
Seq5	Fin Whale (Balenoptera physalus)	NP-006899	NCBI	606
Seq6	Blue Whale (Balenoptera musculus)	NP-007066	NCBI	606
Seq7	Rat (Rattus norvegicus)	AP-004902	NCBI	610
Seq8	Mouse (Mus musculus)	NP-904338	NCBI	607
Seq9	Opossum (Didelphis virginiana	NP-007105	NCBI	602

IV. RESULTS AND DISCUSSION

Classification I: (Biologically obtained)

1. Side chain is aliphatic -- G,A,V,L,I
2. Side chain is an organic acid -- D,E,N,Q
3. Side chain is containing a sulphur -- M,C
4. Side chain is an alcohol -- S,T,Y
5. Side chain is an organic base -- R,K,h
6. Side chain is aromatic -- F,W,P

TABLE III DISTANCE MATRIX OF NINE ND5 PROTEINS BASED ON CLASSIFICATION I

	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9
Seq1	0.000000	0.038881	0.022651	0.025726	0.054693	0.056831	0.078440	0.085455	0.110946
Seq2		0.000000	0.043499	0.045360	0.055971	0.062829	0.085765	0.093135	0.112980
Seq3			0.000000	0.023608	0.059424	0.061174	0.081110	0.087900	0.113435
Seq4				0.000000	0.052717	0.052793	0.071391	0.077975	0.101867
Seq5					0.000000	0.020646	0.068914	0.071017	0.090568
Seq6						0.000000	0.069933	0.069965	0.091609
Seq7							0.000000	0.049956	0.064287
Seq8								0.000000	0.071735
Seq9									0.000000

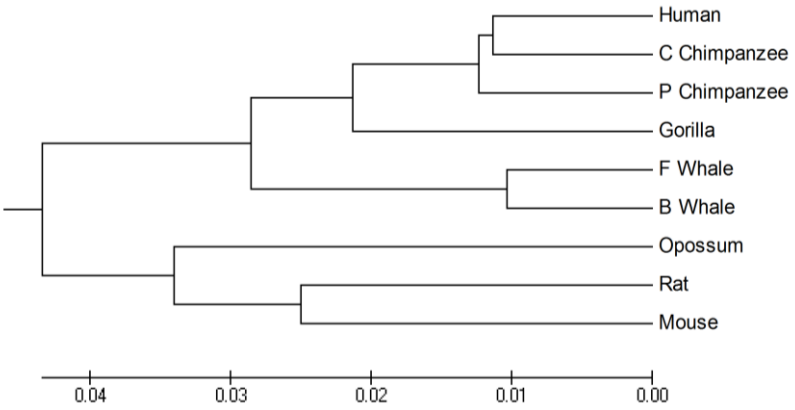

Fig. 1 Phylogenetic tree of Nine ND5 Proteins based on Classification I

Classification II: (Theoretically obtained)

1. Group 1 -- I
2. Group 2 -- L,R
3. Group 3 -- V,A,G,P,T
4. Group 4 -- F,C,Y,Q,N,H,E,D,K
5. Group 5 -- M,W
6. Group 6 -- S

TABLE IV DISTANCE MATRIX OF NINE ND5 PROTEINS BASED ON CLASSIFICATION II

	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9
Seq1	0.000000	0.039520	0.021531	0.032377	0.061099	0.065155	0.115234	0.116073	0.140694
Seq2		0.000000	0.036468	0.041304	0.065758	0.068626	0.102938	0.110177	0.132224
Seq3			0.000000	0.023375	0.059550	0.063663	0.107149	0.111600	0.135364
Seq4				0.000000	0.056999	0.061551	0.096271	0.104466	0.124526
Seq5					0.000000	0.016528	0.107035	0.107391	0.123551

Seq6						0.000000	0.108860	0.108149	0.126832
Seq7							0.000000	0.053303	0.071680
Seq8								0.000000	0.075308
Seq9									0.000000

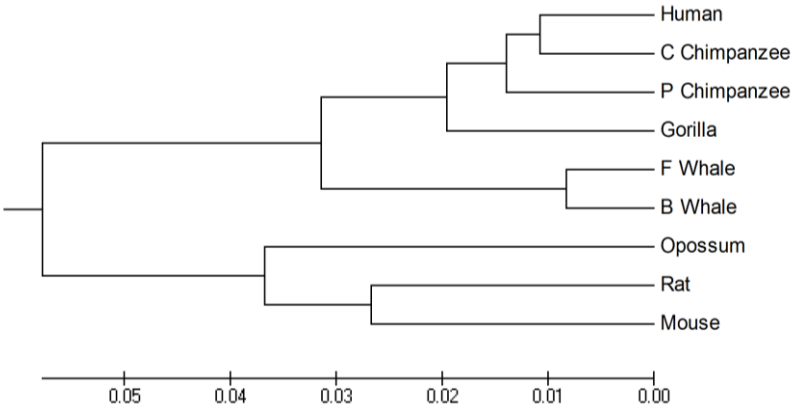

Fig. 2 Phylogenetic tree of Nine ND5 Proteins based on Classification II

The phylogenetic tree obtained from classified six groups of amino acids either biologically or theoretically show uniform results on comparison. However it may be more critical to remark that the phylogenetic trees differ in scales of representation. The scales of comparison are found to be smaller in case of theoretical one as compared to those of the biological one.

V. CONCLUSIONS

The results of comparison based on classification of amino acids in six groups of amino acids show that the methodology is equally applicable to both the cases. Moreover the results of comparison are comparatively finer in the later case than the earlier one. Thus theoretical classification may be claimed to be better than the biological one.

REFERENCES

[1] Milan Randic et. Al., "Novel 2-D graphical representation of proteins", Chemical Physics Letters, Elsevier, doi:10.1016/j.cplett.2005.11.091.

[2] Manoj Kumar Gupta, Rajdeep Niyogi, Manoj Misra, "A 2D Graphical Representation of Protein Sequence and Their Similarity Analysis with Probabilistic Method", MATCH Commun. Math. Comput. Chem. 72 (2014) 519-532, ISSN 0340 – 6253. (4 groups)

[3] Milan Randic, Jure Zupan,, Alexandru, T. Balban, ' Unique graphical representation of protein sequences based on nucleotide triplet codons, Chemical Physics Letters 397 (2004) 247-252

[4] Zu-Guo, Vo Anh, Ka-Sing Lau, "Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analysis", Journal of Theoretical Biology, Elsevier, dio: 10.1016/j.jtbi.2003.09.009. (5 groups)

[5] Chun Li, Lili Xing, Xin Wang, "2-D graphical representation of protein sequences and its application to coronavirus phylogeny", BMB reports, July 2007, Page 217-222.

[6] Yusen Zhang, Xiangtian Yu, "Analysis of protein sequence similarity"978-1-4244-6439-5/10/2010 IEEE, pp. 1255-1258.

[7] Yu-hua Yao, Fen Kong, Qi Dai, Ping-an He, " A Sequence segmented method applied to the Simiarity analysis of Long Protein Sequnce", MATCH Commun. Math.Comput. Chem. 70 (2013) 431-450

[8] Soumen Ghosh et. al., "Classification of Amino Acids of a Protein on the basis of Fuzzy set theory", International Journal of Modern Sciences and Engineering Technology, ISSN 2349-3755, Volume 1, Issue 6, 2014, pp.30-35.

