International Journal of Engineering & Technology, 7 (2) (2018) 678-686

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i2.9546
Research paper

A sequential development towards a unified approach to protein sequence comparison based on classified groups of amino acids

Soumen Ghosh [1] *, Jayanta Pal [2], Bansibadan Maji [3], Dilip Kumar Bhattacharya [4]

[1] Information Technology, Narula Institute of Technology, Kolkata, West Bengal 700109, India
[2] Computer Science & Engineering, Narula Institute of Technology, Kolkata, West Bengal 700109, India
[3] Electronics & Communication Engineering, National Institute of Technology, Durgapur, West Bengal 713209, India
[4] Pure Mathematics, Calcutta University, Kolkata, West Bengal 700073, India
**Corresponding author E-mail: soumenghosh.kolkata@gmail.com*

Abstract

The methods of comparison of protein sequences based on different classified groups of amino acids add a significant contribution to the literature of protein sequence comparison. But the methods vary with choice of different classified groups. Therefore, the purpose of the paper is to develop a unified approach towards the analysis of protein sequence comparison based on classification of amino acids in different groups of different cardinality. The paper considers 4 group classification, 5 group classification and 6 group classifications of amino acids, and in each case it applies the unified method for comparing two types of protein sequences, viz., 9 proteins of ND5 category and 50 Corona virus Spike Proteins. The results agree with those, which were obtained earlier by other methods based on classified groups of amino acids. An-yway it is found that the present unified formula is relatively simpler and fundamentally different from the earlier ones. Further, it can be applied conveniently in comparison of protein sequences based on all different types of classified groups of amino acids.

Keywords: *Protein Sequence Comparison; Different Classified Groups of Amino Acids; ND5 Proteins; Corona Virus Spike Proteins;*

1. Introduction

It may be stated that just as a genome is always expressed by [4] nucleotides, a protein is expressed by 20 amino acids. Therefore, a protein sequence comparison usually follows the same approach as is considered in genome sequence analysis. In details, first of all, numerical representations 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 obtaining phylogenetic trees for comparing protein sequences [1, 2, 3]. Again amino acids are 20 in number, so 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. Such classified groups of amino acids are found in [4, 5, 6, 7] and they are listed below:

1.1. Classification I (4 groups): detailed HP model

Table 1: 4 Group Classifications (Hp Model)

Characteristic	Amino acids
Hydrophobic (H) (non-polar)	A, I, L, M, F, P, W, V
Negative polar class	D, E
Uncharged polar class	N , C, Q, G, S, T, Y
Positive polar class	R, H, K

Zu-Guo Yu, Vo Anh, Ka-Sing Lau 2004 [4] has studied this model through its Chaos game representation for multifractal and correlation analysis. But protein sequence comparison has not been taken up with these classified groups of amino acids.

1.2. Classification II (five groups): (I, A, G, E, K)

Wang and Wang 1999, 2000 [8-9] prove that the five letter code (I, A, G, E, K) is feasible for elucidating characteristics of real proteins with 20 kinds of amino acids. Following the methods of Wang and Wang 1999, 2000 [8-9], the 20 amino acids are classified in [5] groups as follows:

Table 2: 5 Group Classifications

Representative residues	Amino acids
I	C, M, F, I, L, V, W, Y
A	A, T, H
G	G, P
E	D, E
K	S, N, Q, R, K

For protein sequence comparison of this model Chun Li, Lili Xing, Xin Wang 2007 [5] consider the method of representation of protein sequences as is normally done with DNA sequences. The five letters are now associated with five horizontal lines. The representation is obtained in the form of a zigzag curve. Taking the alley Index of L/L matrix as the descriptor and using Euclidean distance as the distance measure, the authors obtain phylogenic tree of 56 corona virus spike proteins.

Copyright © 2018 Soumen Ghosh 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.

1.3. Classification III: (four groups)

Table 3: 4 Group Classifications (Hc Model)

Hydropathy characteristic	Abbreviation	Amino acids
Strongly Hydrophilic	POL	R, D, E, N, Q, K, H
Strongly Hydrophobic	HPO	L, I, V, A, M, F
Weakly Hydrophilic or weakly Hydrophobic (Ambiguous)	Ambi	S, T, Y, W
Special	None	C, G, P

As s, t, y, w can neither be called strongly hydrophilic, nor strongly hydrophobic, so they are ambiguous in classification. Hence they are given a separate group named ambiguous. W. Taylor [10] explains why Proline, Glycine and Cystine are put in a separate group. Actually Proline and Glycine do not belong to any hydropathy set because of their unique backbone properties, and Cystine is excluded from any set, because it has polarizable properties.
Yusen Zhang and Xian Yu 2010 [11] use it for sequence comparison of 9 ND5 proteins. The authors first represented the four classes by the letters R, L, S, C respectively, so as to express any protein sequence expressed in terms of R, L, S, C only. They explained the procedure by taking a sample sequence of length 20 given by merikelrdlmsqsrtreig and expressed it as lrrlrrlrrllsrsrsrrlc. Then they used the symbol $n_i^j, i = 1,2,3,4$ to represent the number of elements in respective classes; for i = 1, j = 1,2,3,4,5,6,7; for i =2, j = 1,2,3,4,5,6; for i=3, j=1,2,3,4; for i =4, j=1,2,3.They used n_{ij} to represent number of 2-Blocks like RR, RC etc. and Ni j to represent 3-blocks like RXR, where X is any one of R, L, S, C lying in between. Finally they obtained the 56 dimensional descriptor vector as v = [X_1, X_2, X_3, X_4], where

$$X_1 = \frac{\sum_{j=1}^{n_i} n_i^j}{\text{length (S)}}, i = 1,2,3,4; X_2 = \frac{n_{ij}}{\sum_{j=1}^{4} n_{ij}}, i, j = 1,2,3,4;$$

$$X_3 = \frac{N_{ij}}{\sum_{j=1}^{4} N_{ij}}, i, j = 1,2,3,4; X_4 = \frac{n_i^j}{\sum_{j=1}^{n_i} n_i^j}, i = 1,2,3,4 \quad (1)$$

X_1 has 4 components, X_2 has 16 components, X_3 has 16 components, and X_4 has 20 components. This method is slightly modified in [12] to apply for sequence comparison of proteins classified in the following six groups.

1.4. Classification IV: six groups (biologically obtained)

Table 4: 6 Group Classification (Biologically Obtained)

Characteristic	Amino acids
Side chain is aliphatic	G, A, V, L, I
Side chain is an organic acid	D, E, N, Q
Side chain coniatns a sulphur	M, C
Side chain is an alcohol	S, T, Y
Side chain is an organic base	R, K, H
Side chain is aromatic	F, W, P

1.5. Classification V: six groups (theoretical) [13]

Table 5: 6 Group Classification (Theoretically Obtained)

Representative residues	Amino acids
I	I
L	L, R
A	V, A, G, P, T
E	F, C, Y, Q, N, H, E, D, K
M	M, W
S	S

1.6. Modified formula

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 sequence of length N, at first all members of G_i are replaced by their representatives R_i, i= 1, 2... m, to make a new sequence of m variables R_i. Let the frequencies of number of occurrences of amino acids of $G_1, G_2\ldots, G_m$ be respectively given by $f_{11}, f_{12},..,f_{1n_1}, f_{21}, f_{22},..,f_{2n_2},.....,f_{m1}, f_{m2},..,f_{mn_m}$. In the newly represented protein sequence where all members of G_i are replaced by their representatives R_i, i= 1,2,..,m, let the frequencies of number of occurrences of R_i, i = 1,2,..,m be g_i, i = 1,2,..,m. Similarly let the frequencies of number of occurrences of Ri taken two at a time and the frequencies of number of occurrences of R_i taken two at a time with any amino acid lying in between the combination be respectively given by $h_1, h_2, .., h_{m^2}$ and $w_1, w_2, .., w_{m^2}$ .Then the descriptor for protein sequence comparison is taken as a $20+m+m^2+m^2$

Component vector

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

Although the formula is a general one applicable to any m-group classification, still it has been applied to 6 group classifications only. Further, it has been applied for ND5 proteins only, but 56 corona virus spike proteins were not compared under this method. Thus the method of protein sequence comparison based on one type of classified 4 group (classification I) is not available. Again comparison of sequences of 9 ND5 proteins under 4 group (classification III) has been made in [13] and the same for 56 corona virus spike proteins under 5 group (classification II) has been taken up in [14]. Also six group (classifications IV (A) and IV (B) were tried with modified formula for comparison of only 9 ND5 proteins [12]. The formula was not tested with 50 corona virus proteins. Anyway the methods are not same in all cases. They differ from choice of classified groups –obviously there is no way of comparing the results for the same sequence obtained under different methods of analysis. So it is necessary to obtain a unified approach to protein sequence comparison independent of the type of classified groups of amino acids and independent of the choice of protein sequences. The motivation of the present paper is to try to obtain a unified formula to protein sequence comparison based on classified groups of amino acids of all categories. But even the modified formula (2) does not work in all categories. For example when formula (2) was applied on ND5 protein sequences based on classification III it gives the following unsatisfactory result (figure 1):

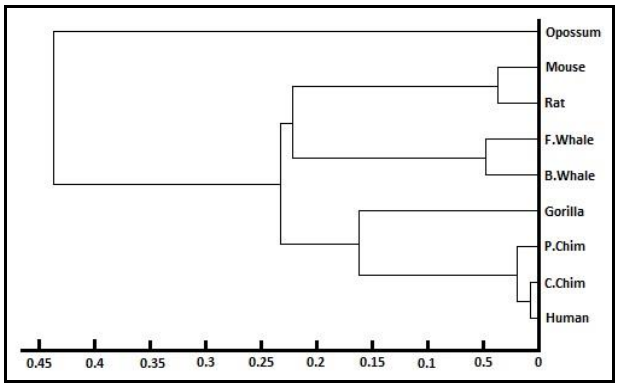


Fig. 1: Phylogenetic Tree of ND5 Proteins Based on Classification III.

So further modification of the formula of (2) is necessary in order to get a uniform one applicable to all categories of classifications of amino acids. The present paper tries to obtain such a unified formula.

2. Unified approach to protein sequence comparison

Unified Formula for construction of descriptor vector for comparison of protein sequences

With the same notations as used in formula (2), the descriptor for protein sequence comparison is taken as a $20+m+m^2+m^2$ component vector given by

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

The formula is similar but not identical with that of (2).

A sample example to demonstrate the method:

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 10 amino acids (SYPHYVKSIV). 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 get the first twenty components {1, 0, 2, 0, 0, 0, 0, 1, 0, 2, 2, 0, 0, 0, 1, 0, 0, 1, 0}.

Next the sequence of the amino acids of the example is converted to f d c d d c d f a c.

Now six components are the number of occurrences of each group i.e. {1, 0, 3, 4, 0, 2}.

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 = 0, 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 = 2, ce = 0, cf = 1, da = 0, db = 0, dc = 2, dd = 1, de = 0, df = 1, ea = 0, eb = 0, ec = 0, ed = 0, ee = 0, ef = 0, fa = 1, fb = 0, fc = 0, fd = 1, fe = 0 and ff = 0. So the 36 components are

{0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 2, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0}.

Next m^2 components are the number of occurrences of two pair combinations of six groups after taking two at a time with any group lying in between the combination. Here number of occurrences of axa = 0, axb = 0, axc = 0, axd = 0, axe = 0, axf = 0, bxa = 0, bxb = 0, bxc = 0, bxd = 0, bxe = 0, bxf = 0, cxa = 0, cxb = 0, cxc = 0, cxd = 1, cxe = 0, cxf = 1, dxa = 1, 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 = 0, fxe = 0 and fxf =0. So the 36 components are {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0}. Thus the 98component descriptor is

(1, 0, 2, 0, 0, 0, 0, 0, 1, 0, 2, 2, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 3, 4, 0, 2, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 2, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0 , 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0)

3. Details of protein sequences considered for comparison

Table 6: Detail List of [9] ND5 Species

Sl. No.	Species	ID/Accession	Database	Length
Seq 1	Human (Homo sapiens)	AP-000649	NCBI	603
Seq 2	Gorilla(Gorilla gorilla)	NP-008222	NCBI	603
Seq 3	Common chimpanzee (Pan troglodytes)	NP-008196	NCBI	603
Seq 4	Pigmy chimpanzee (Pan paniscus)	NP-008209	NCBI	603
Seq 5	Fin whale (Balenoptera physalus)	NP-006899	NCBI	606
Seq 6	Blue whale (Balenoptera musculus)	NP-007066	NCBI	606
Seq 7	Rat (Rattus norvegicus)	AP-004902	NCBI	610
Seq 8	Mouse (Mus musculus)	NP-904338	NCBI	607
Seq 9	Opossum (Didelphis virginiana	NP-007105	NCBI	602

Table 7: Detail List of 50 Corona Virus Spike Proteins

No.	Accession number	Name	Abbreviation	Length
1	CAB91145	Transmissible gastroenteritis virus, genomic RNA	TGEVG	1447
2	NP_058424	Transmissible gastroenteritis virus	TGEV	1447
3	AAK38656	Porcine epidemic diarrhea virus strain CV777	PEDVC	1383
4	NP_598310	Porcine epidemic diarrhea virus	PEDV	1383
5	NP_937950	Human corona virus OC43	HCoV-OC43	1361
6	AAK83356	Bovine corona virus isolate BCoV-ENT	BCoVE	1363
7	AAL57308	Bovine corona virus isolate BCoV-LUN	BCoVL	1363
8	AAA66399	Bovine corona virus strain Mebus	BCoVM	1363
9	AAL40400	Bovine corona virus strain Quebec	BCoVQ	1363
10	NP_150077	Bovine corona virus	BCoV	1363
11	AAB86819	Mouse hepatitis virus strain MHV-A59C12 mutant	MHVA	1324
12	YP_209233	Murine hepatitis virus strain JHM	MHVJHM	1376
13	AAF69334	Mouse hepatitis virus strain Penn 97-1	MHVP	1321
14	AAF69344	Mouse hepatitis virus strain ML-10	MHVM	1324
15	NP_045300	Mouse hepatitis virus	MHV	1324
16	AAP92675	Avain infectious bronchitis virus isolate BJ	IBVBJ	1169
17	AAS00080	Avain infectious bronchitis virus strain Ca199	IBVC	1169
18	NP_040831	Avain infectious bronchitis virus	IBV	1162
19	AAS10463	SARS corona virus GD03T0013	GD03T0013	1255
20	AAU93318	SARS corona virus PC4-127	PC4-127	1255
21	AAV49720	SARS corona virus PC4-137	PC4-137	1255
22	AAU93319	SARS corona virus PC4-205	PC4-205	1255
23	AAU04646	SARS corona virus civet007	civet007	1255
24	AAU04649	SARS corona virus civet010	civet010	1255
25	AAU04664	SARS corona virus civet020	civet020	1255
26	AAV91631	SARS corona virus A022	A022	1255
27	AAV49730	SARS corona virus B039	B039	1255
28	AAP51227	SARS corona virus GD01	GD01	1255
29	AAS00003	SARS corona virus GZ02	GZ02	1255
30	AAP30030	SARS corona virus BJ01	BJ01	1255
31	AAP13567	SARS corona virus CUHK-W1	CUHK-W1	1255
32	AAP50485	SARS corona virus FRA	FRA	1255

33	AAP41037	SARS corona virus TOR2	TOR2	1255
34	AAQ01597	SARS corona virus Taiwan TC1	TaiwanTC1	1255
35	AAQ01609	SARS corona virus Taiwan TC2	TaiwanTC2	1255
36	AAP13441	SARS corona virus Urbani	Urbani	1255
37	AAQ94060	SARS corona virus AS	AS	1255
38	AAP30713	SARS corona virus CUHK-Su10	CUHK-Su10	1255
39	AAP33697	SARS corona virus Frankfurt 1	Frankfurt1	1255
40	AAP94737	SARS corona virus CUHK-AG01	CUHK-AG01	1255
41	AAP94748	SARS corona virus CUHK-AG02	CUHK-AG02	1255
42	AAP37017	SARS corona virus TW1	TW1	1255
43	AAR87523	SARS corona virus TW2	TW2	1255
44	BAC81348	SARS corona virus TWH genomic RNA	TWH	1255
45	BAC81362	SARS corona virus TWJ genomic RNA	TWJ	1255
46	AAP72986	SARS corona virus HSR 1	HSR1	1255
47	AAR23250	SARS corona virus Sin01-11	Sino1-11	1255
48	AAR23258	SARS corona virus Sin03-11	Sino3-11	1255
49	AAR14803	SARS corona virus PUMC01	PUMC01	1255
50	AAR14807	SARS corona virus PUMC02	PUMC02	1255

4. Results

Classification I: ([4] Groups)
ND5 proteins

The results obtained by unified method

Table 8: Distance Matrix of ND5 Proteins Based on Classification

	Human	Gorilla	C. Chim	P. Chim	F. Whale	B. Whale	Rat	Mouse	Opossum
Human	0.0000								
Gorilla	0.0329	0.0000							
C. Chim	0.0203	0.0425	0.0000						
P. Chim	0.0203	0.0338	0.0196	0.0000					
F. Whale	0.0606	0.0546	0.0696	0.0611	0.0000				
B. Whale	0.0569	0.0527	0.0639	0.0556	0.0167	0.0000			
Rat	0.0585	0.0597	0.0614	0.0563	0.0654	0.0635	0.0000		
Mouse	0.0711	0.0659	0.0765	0.0682	0.0537	0.0528	0.0407	0.0000	
Opossum	0.0843	0.0723	0.0912	0.0811	0.0696	0.0711	0.0570	0.0500	0.0000

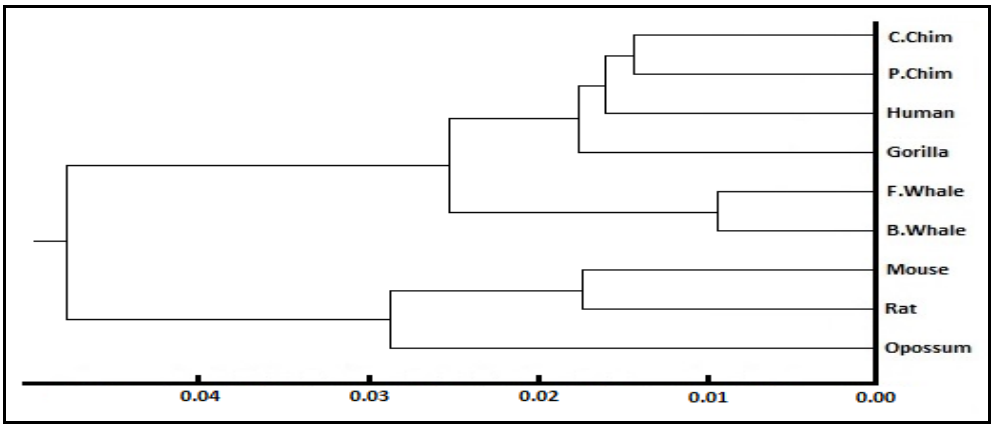


Fig. 2: Phylogenetic Tree of ND5 Proteins Based on Classification I

50 Corona Virus spike proteins

Table 9: Part of Distance Matrix of 50 Corona Virus Spike Proteins Based on Classification I

	BCoVM	MHVA	MHVP	MHVM	PEDVC	BCoVE	BCoVQ	BCoVL	Urbani	CUHK-W1
BCoVM	0.0000									
MHVA	0.0321	0.0000								
MHVP	0.0318	0.0127	0.0000							
MHVM	0.0326	0.0041	0.0117	0.0000						
PEDVC	0.0550	0.0548	0.0538	0.0566	0.0000					
BCoVE	0.0173	0.0362	0.0340	0.0361	0.0634	0.0000				
BCoVQ	0.0052	0.0300	0.0301	0.0306	0.0534	0.0209	0.0000			
BCoVL	0.0140	0.0331	0.0318	0.0333	0.0602	0.0070	0.0169	0.0000		
Urbani	0.0794	0.0711	0.0744	0.0721	0.0657	0.0892	0.0753	0.0849	0.0000	
CUHK-W1	0.0795	0.0711	0.0744	0.0721	0.0657	0.0892	0.0754	0.0848	0.0039	0.0000

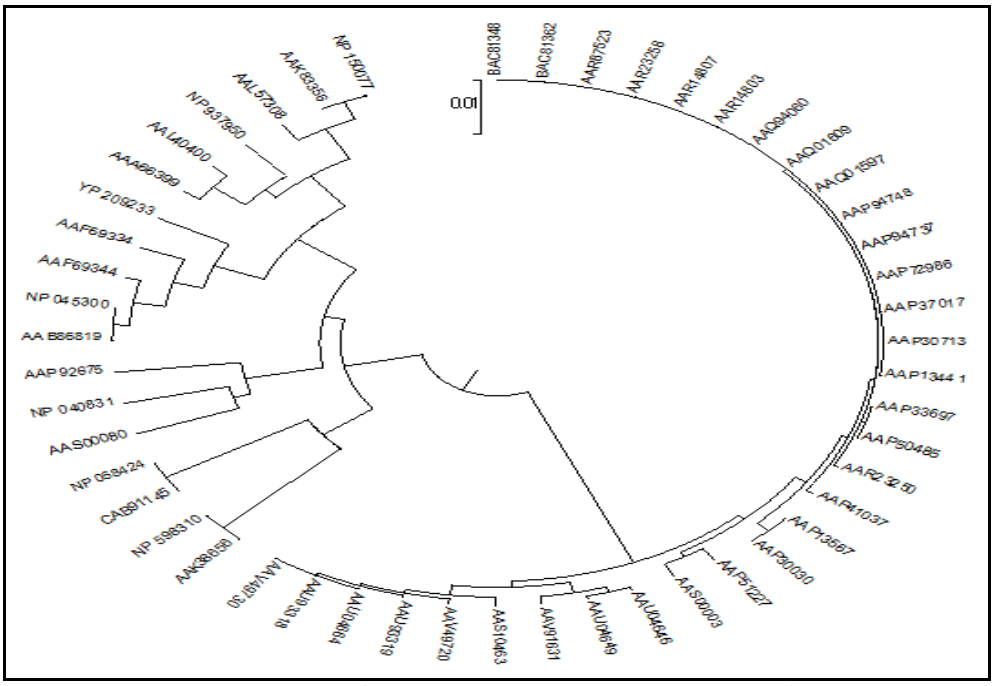


Fig. 3: Phylogenetic Tree of 50 Corona Virus Spike Proteins Based on Classification I.

Classification II: (5 Groups)
ND5 proteins

Table 10: Distance Matrix of ND5 Proteins Based on Classification II

	Human	Gorilla	C. Chim	P. Chim	F. Whale	B. Whale	Rat	Mouse	Opossum
Human	0.0000								
Gorilla	0.0387	0.0000							
C. Chim	0.0266	0.0335	0.0000						
P. Chim	0.0354	0.0326	0.0238	0.0000					
F. Whale	0.0553	0.0456	0.0494	0.0401	0.0000				
B. Whale	0.0567	0.0481	0.0489	0.0387	0.0175	0.0000			
Rat	0.1020	0.0857	0.0925	0.0791	0.0857	0.0823	0.0000		
Mouse	0.1036	0.0858	0.0945	0.0819	0.0846	0.0805	0.0344	0.0000	
Opossum	0.1244	0.1074	0.1159	0.1000	0.1044	0.1011	0.0532	0.0555	0.0000

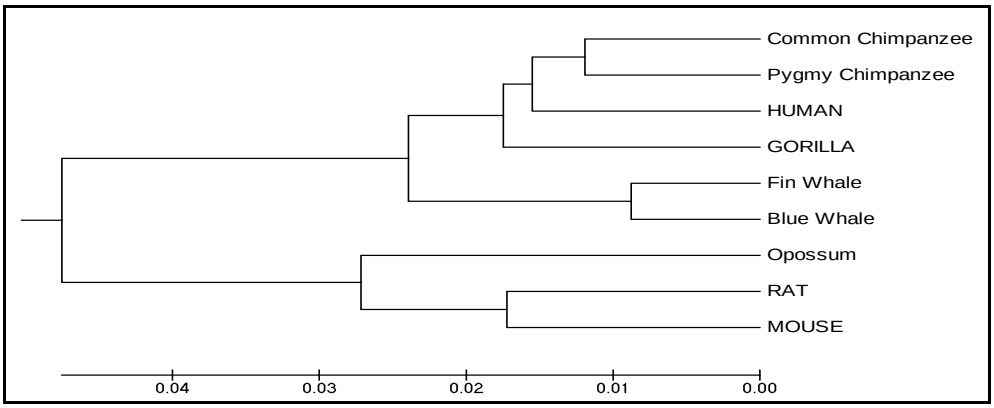


Fig. 4: Phylogenetic Tree of ND5 Proteins Based on Classification II.

Current results on [50] corona virus spike proteins

Table 11: Part of Distance Matrix of 50 Corona Virus Spike Proteins Based on Classification II

	BCoVM	MHVA	MHVP	MHVM	PEDVC	BCoVE	BCoVQ	BCoVL	Urbani	CUHK-W1
BCoVM	0.0000									
MHVA	0.0434	0.0000								
MHVP	0.0443	0.0159	0.0000							
MHVM	0.0447	0.0045	0.0148	0.0000						
PEDVC	0.0611	0.0611	0.0595	0.0627	0.0000					
BCoVE	0.0144	0.0391	0.0398	0.0402	0.0581	0.0000				
BCoVQ	0.0040	0.0432	0.0439	0.0445	0.0615	0.0148	0.0000			
BCoVL	0.0161	0.0377	0.0390	0.0389	0.0573	0.0055	0.0163	0.0000		
Urbani	0.0670	0.0559	0.0545	0.0568	0.0653	0.0630	0.0666	0.0604	0.0000	
CUHK-W1	0.0679	0.0566	0.0552	0.0576	0.0653	0.0637	0.0675	0.0612	0.0037	0.0000

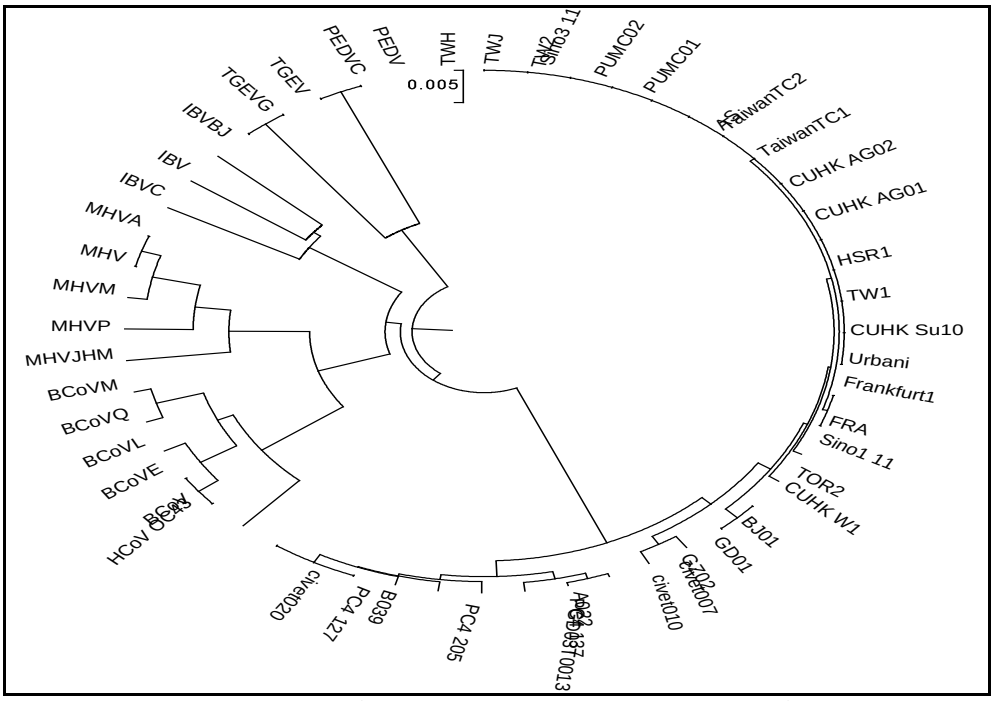


Fig. 5: Phylogenetic Tree of 50 Corona Virus Spike Proteins Based on Classification II.

Classification III: ([4] Groups)
ND5 proteins

Table 12: Distance Matrix of ND5 Based on Classification III

	Human	Gorilla	C. Chim	P. Chim	F. Whale	B. Whale	Rat	Mouse	Opossum
Human	0.0000								
Gorilla	0.0412	0.0000							
C. Chim	0.0220	0.0463	0.0000						
P. Chim	0.0229	0.0346	0.0215	0.0000					
F. Whale	0.0508	0.0459	0.0620	0.0527	0.0000				
B. Whale	0.0478	0.0449	0.0576	0.0489	0.0180	0.0000			
Rat	0.0683	0.0630	0.0707	0.0631	0.0616	0.0633	0.0000		
Mouse	0.0747	0.0740	0.0802	0.0732	0.0627	0.0673	0.0417	0.0000	
Opossum	0.0965	0.0784	0.1016	0.0892	0.0724	0.0762	0.0587	0.0602	0.0000

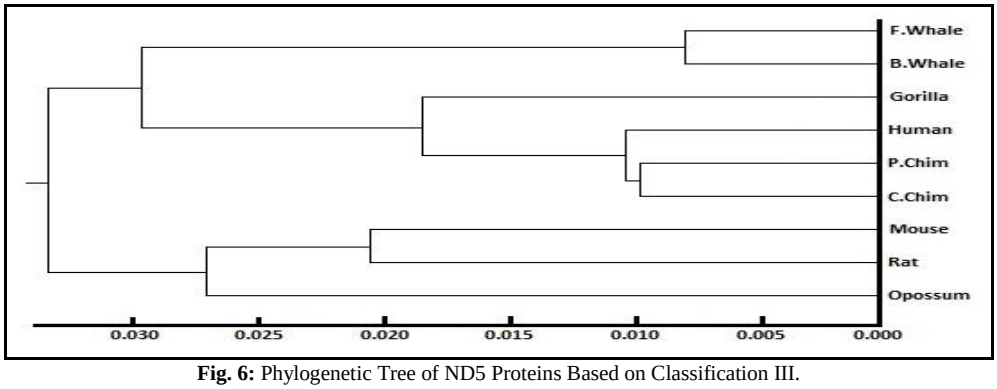


Fig. 6: Phylogenetic Tree of ND5 Proteins Based on Classification III.

Classification IV: ([6] Groups)

Table 13: Distance Matrix of ND5 Based on Classification IV

	Human	Gorilla	C. Chim	P. Chim	F. Whale	B. Whale	Rat	Mouse	Opossum
Human	0.0000								
Gorilla	0.0399	0.0000							
C. Chim	0.0236	0.0450	0.0000						
P. Chim	0.0268	0.0462	0.0256	0.0000					
F. Whale	0.0573	0.0600	0.0628	0.0537	0.0000				
B. Whale	0.0598	0.0662	0.0650	0.0543	0.0205	0.0000			
Rat	0.0793	0.0861	0.0816	0.0713	0.0685	0.0705	0.0000		
Mouse	0.0884	0.0949	0.0905	0.0800	0.0722	0.0725	0.0529	0.0000	
Opossum	0.1179	0.1212	0.1197	0.1075	0.0951	0.0967	0.0691	0.0767	0.0000

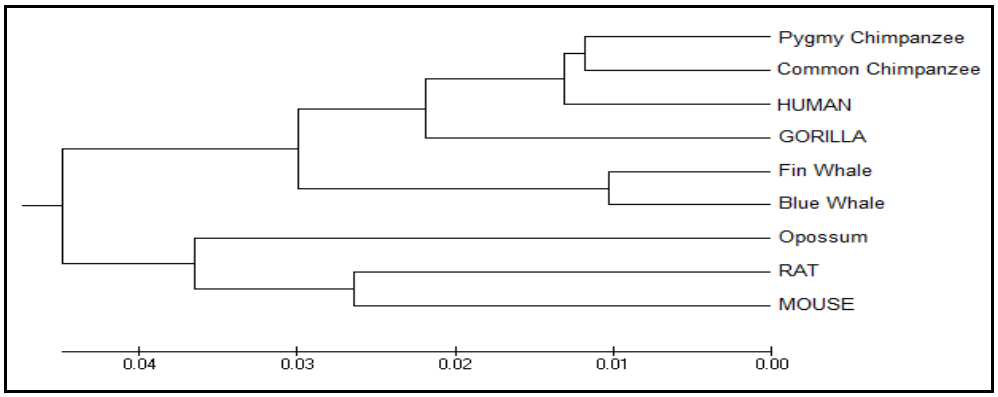


Fig. 7: Phylogenetic Tree of ND5 Proteins Based on Classification IV.

Table 14: Part of Distance Matrix of 50 Corona Virus Spike Proteins Based on Classification IV.

	BCoVM	MHVA	MHVP	MHVM	PEDVC	BCoVE	BCoVQ	BCoVL	Urbani	CUHK-W1
BCoVM	0.0000									
MHVA	0.0347	0.0000								
MHVP	0.0351	0.0128	0.0000							
MHVM	0.0358	0.0055	0.0124	0.0000						
PEDVC	0.0729	0.0746	0.0722	0.0771	0.0000					
BCoVE	0.0139	0.0371	0.0370	0.0382	0.0716	0.0000				
BCoVQ	0.0058	0.0336	0.0343	0.0347	0.0736	0.0164	0.0000			
BCoVL	0.0135	0.0362	0.0367	0.0372	0.0713	0.0069	0.0148	0.0000		
Urbani	0.0675	0.0637	0.0644	0.0630	0.0951	0.0687	0.0666	0.0675	0.0000	
CUHK-W1	0.0684	0.0650	0.0656	0.0641	0.0963	0.0693	0.0676	0.0681	0.0039	0.0000

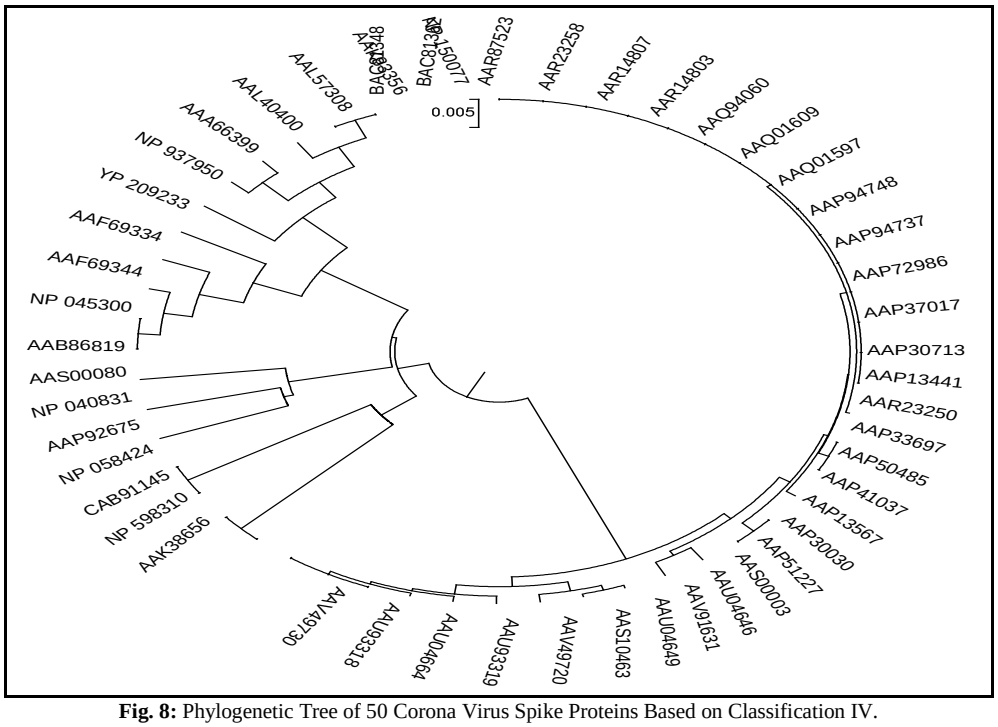


Fig. 8: Phylogenetic Tree of 50 Corona Virus Spike Proteins Based on Classification IV.

Classification V: ([6] Groups)

Table 15: Distance Matrix of ND5 Based on Classification V.

	Human	Gorilla	C. Chim	P. Chim	F. Whale	B. Whale	Rat	Mouse	Opossum
Human	0.0000								
Gorilla	0.0404	0.0000							
C. Chim	0.0217	0.0372	0.0000						
P. Chim	0.0327	0.0430	0.0238	0.0000					
F. Whale	0.0614	0.0668	0.0600	0.0575	0.0000				
B. Whale	0.0654	0.0696	0.0640	0.0620	0.0164	0.0000			
Rat	0.1133	0.1021	0.1056	0.0945	0.1052	0.1073	0.0000		
Mouse	0.1132	0.1080	0.1087	0.1018	0.1033	0.1050	0.0528	0.0000	
Opossum	0.1407	0.1326	0.1354	0.1250	0.1226	0.1265	0.0733	0.0754	0.0000

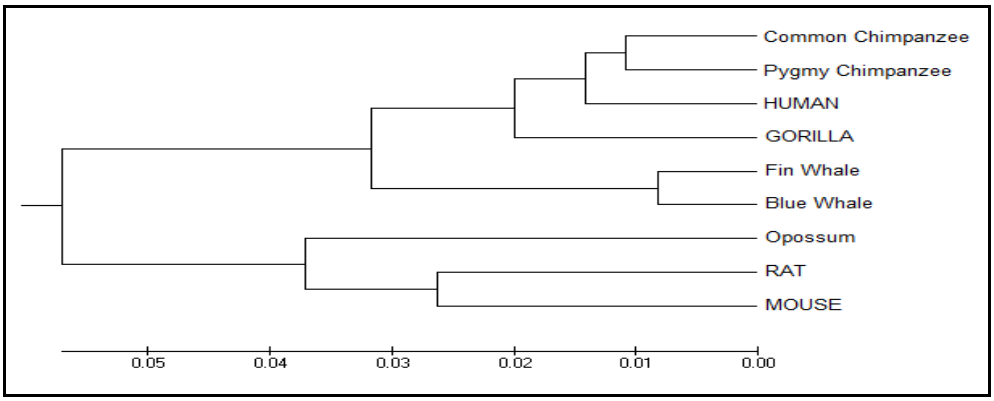


Fig. 9: Phylogenetic Tree of ND5 Proteins Based on Classification V.

Table 16: Part of Distance Matrix of 50 Corona Virus Spike Proteins Based on Classification V.

	BCoVM	MHVA	MHVP	MHVM	PEDVC	BCoVE	BCoVQ	BCoVL	Urbani	CUHK-W1
BCoVM	0.0000									
MHVA	0.0377	0.0000								
MHVP	0.0372	0.0175	0.0000							
MHVM	0.0384	0.0053	0.0162	0.0000						
PEDVC	0.0862	0.0782	0.0794	0.0808	0.0000					
BCoVE	0.0119	0.0359	0.0342	0.0364	0.0843	0.0000				
BCoVQ	0.0055	0.0376	0.0370	0.0381	0.0890	0.0129	0.0000			
BCoVL	0.0134	0.0353	0.0344	0.0361	0.0822	0.0052	0.0145	0.0000		
Urbani	0.0525	0.0548	0.0524	0.0544	0.0877	0.0502	0.0525	0.0493	0.0000	
CUHK-W1	0.0536	0.0562	0.0535	0.0557	0.0891	0.0512	0.0535	0.0504	0.0037	0.0000

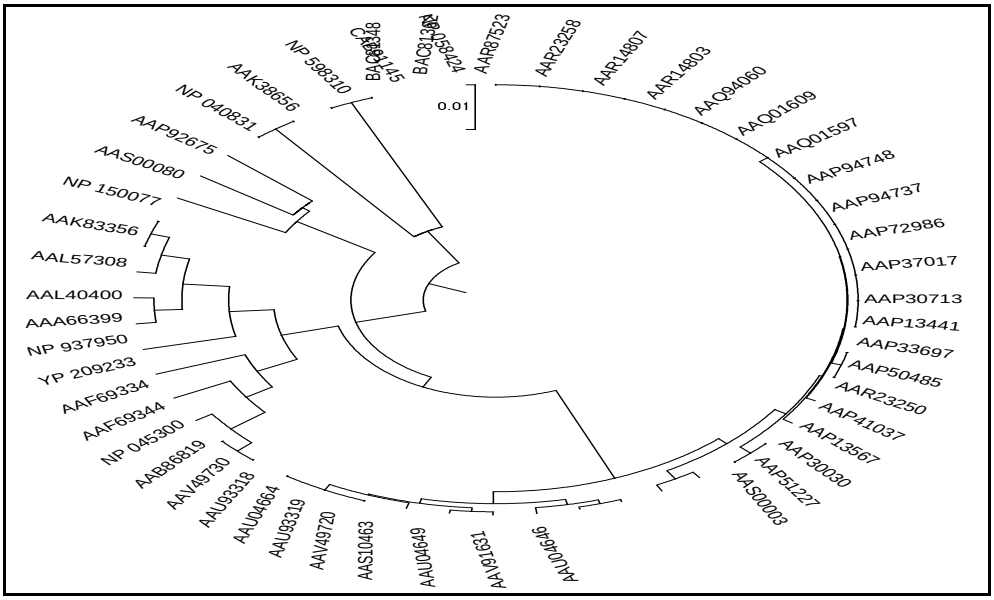


Fig. 10: Phylogenetic Tree of 50 Corona Virus Spike Proteins Based on Classification V.

5. Discussion

Figure [1] gives phylogenetic tree of {9} species of ND5 category of proteins based on [4] classified groups of amino acids obtained by an earlier method. Figure 6 represents the same for the same species of same category of proteins based on [4] classified groups of amino acids obtained by our new method. The two methods are slightly different. But the figures are almost alike. Thus, it is seen that the present method is a unified one in the sense that it can work effectively in both the cases. Further, the present method is not a complicated one.

6. Conclusion

The present method of approach to protein sequence comparison based on classified groups of amino acids is a unified one. The method is not complicated. It can be conveniently applied to compare any pair of protein sequences based on classified groups of amino acids of any cardinality.

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. https://doi.org/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. https://doi.org/10.1016/j.cplett.2004.08.118.

[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. (Five groups).

[5] Chun Li, Lili Xing, Xin Wang, "2-D graphical representation of protein sequences and its application to corona virus 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-and He, "A Sequence segmented method applied to the Simiarity analysis of Long Protein Sequnce", MATCH Commun. Math.Comput. Chem. 70 (2013) 431-450.
[8] Wang, J and Wang, W (1999) A computational approach to simplifying the protein folding problem. Nat.Struct. Biol. 1033-1038. https://doi.org/10.1038/14918.
[9] Wang, J and Wang, W (2000) Modeling study on the validity of a possibly simplified representation of proteins. Phys. Rev. E. 61, 6981-6086. https://doi.org/10.1103/PhysRevE.61.6981.
[10] W. Taylor," Identification of protein sequence boundary by consensus template alignment"- Journal of molecular Biology, vol. 188, pp. 233-258, 1985. https://doi.org/10.1016/0022-2836(86)90308-6.
[11] Yusen Zhang, Xiangtian Yu – Analysis of Protein Sequence similarity- 978-1-4244-6439-5/19/$26.00(c) 2010 IEEE.
[12] S. Ghosh, J. Pal, S. Das, D. K. Bhattacharya- Differentiation of Protein Sequence Comparison Based on Biological and Theoretical Classification of Amino Acids in Six Groups- International Journal of Advanced Research in Computer Science and Software Engineering: Volume 5, Issue 6, June 2015, pp. 695-698
[13] 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.
[14] Chun Li, Lili Xing & Xin Wang, "2-D graphical representation of protein sequences and its application to corona virus phylogeny", BMB reports, October 2007, pp. 217-222.