Selection of Moment Vectors in Protein Sequence Comparison Under Binary Representation

Jayanta Pal (jayantapal1978@yahoo.com)
Narula Institute of Technology https://orcid.org/0000-0003-4459-8115
Soumen Ghosh
Narula Institute of Technology
Bansibadan Maji
National Institute of Technology, Dugapur
Dilip Kumar Bhattacharya
University of Calcutta

Research

Keywords: Fourier Transform, Perseval's Identity, Power Spectrum, Moments Vector, Euclidean Distance, Phylogenetic Tree.

Posted Date: November 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1028526/v1

Abstract

Similarity/dissimilarity study of protein and genome sequences remains a challenging task and selection of techniques and descriptors to be adopted, plays an important role in computational biology. Again, genome sequence comparison is always preferred to protein sequence comparison due the presence of 20 amino acids in protein sequence compared to only 4 nucleotides in genome sequence. So it is important to consider suitable representation that is both time and space efficient and also equally applicable to protein sequences of equal and unequal lengths. In the binary form of representation, Fourier transform of a protein sequence reduces to the transformation of 20 simple binary sequences in Fourier domain, where in each such sequence, Perseval's Identity gives a very simple computable form of power spectrum. This gives rise to readily acceptable forms of moments of different degrees. Again such moments, when properly normalized, show a monotonically descending trend with the increase in the degrees of the moments. So it is better to stick to moments of smaller degrees only. In this paper, descriptors are taken as 20 component vectors, where each component corresponds to a general second order moment of one of the 20 simple binary sequences. Then distance matrices are obtained by using Euclidean distance as the distance measure between each pair of sequence. Phylogenetic trees are obtained from the distance matrices using UPGMA algorithm. In the present paper, the datasets used for similarity/dissimilarity study are 9 ND4, 16 ND5, 9 ND6, 24 TF proteins and 12 Baculovirus proteins. It is found that the phylogenetic trees produced by the present method are at par with those produced by the earlier methods adopted by other authors and also their known biological references. Further it takes less computational time and also it is equally applicable to sequences of equal and unequal lengths.

Background

Each protein sequence is represented by 20 different amino acids. This is called primary representation of proteins. Naturally numerical representation of protein sequence is expressed by numerical representation of 20 constituent amino acids. Again the method of representation is of two types- one is called alignment free method and the other one is called alignment based method. Some such alignment based methods are given in [1] and [2]. At present, alignment based methods are rarely used owing to their limitations. So nowadays, alignment free methods are the preferable choice of discussion. In this connection, a literature survey is given in [3], where alignment free methods up to 2003 are available. Therefore the present paper concentrates on alignment free methods for the later periods only. Now numerical representations of protein sequences are found with different dimensions- mostly they are two or three dimensional. Graphical representations of the protein sequences are obtained from numerical data points of the representation. Some such graphical representations are given in [4–23]. Comparison of protein sequences depends on proper choice of numerical representation and proper choice of descriptors. Now descriptors may be obtained directly from the data points of the graph, in which case, they are called graphical descriptors. There are also matrix forms of descriptors, where first of all, some matrices are obtained from the data points of the graph and then descriptors are obtained from such matrices. But another completely different approach is also followed in protein sequence comparison.

This is based on classification of amino acids in different groups with different cardinality [22–25]. It is also important to mention a special type of twenty dimensional representations of protein sequences, which is a binary representation [26]. This is an extension of four dimensional binary representation of genome sequences [27], where the four nucleotides are represented by T = (1,0,0,0), C = (0,1,0,0), A = (0,0,1,0) and G = (0,0,0,1) respectively. Now so far as comparisons of biological sequences are concerned, Fast Fourier Transform (FFT) has a significant role. In general, it has a nice application in signal and image processing. In particular, if the original signal is a binary signal as above, it has further advantages. In fact, the genome sequence U(n) is split up into four simpler binary sequences $U_T(n)$, $U_C(n)$, $U_A(n)$, $U_G(n)$ corresponding to the nucleotides T, C, A, G respectively. So the Fourier transform of U(n) is the union of the Fourier transforms of $U_T(n)$, $U_C(n)$, $U_A(n)$ and $U_G(n)$. Based on such representation, genome sequences are compared by the use of ICD (inter coefficient distance) method [28]. The advantage of this method is that it is possible to manage comparison of genome sequences of unequal lengths by putting additional zeroes to make the lengths of the sequences equal. The reason is that zeroes in the time domain have no effect on its Fourier transform. Corresponding application of ICD method in protein sequence comparison is found in twenty dimensional binary representation of Protein sequences [26]. Let us now come back to another advantage of the aforesaid binary representations of genomes. In such a representation, the number of nucleotides of a single category present in each of $U_T(n)$, $U_C(n)$, $U_A(n)$, $U_G(n)$ is exactly equal to the sum of the number of 1's present there. This gives a very simple computable form of power spectrum from Perseval's identity. This is observed in [29], where they obtain expressions for moment vectors of different degrees conveniently. Using 12 dimensional descriptors involving moments of degree 1, 2, 3, authors compare some genome sequences successfully. In this connection, it may be noted that moments of different orders are different measures. Thus three different measures are piled up together to get the 12 component descriptors for genome sequences. Naturally for protein sequences the descriptor would be a 60 component vector. So it remains open to check, if there exists moments of a single degree equally competent for formation of descriptors in protein sequence comparison. The objective of the present paper is to select such a descriptor and to use it in protein sequence comparison. In doing so, it is to be confirmed that the method remains computationally efficient; and at the same time, it can be applied equally well on sequences of equal and unequal lengths.

Methdology

Extended Representation

In the present paper, a binary form of representation is used and it is considered for all the 20 amino acids in the order shown in Table-1. Here the order stands for alphabetical order of abbreviated form of 20 amino acids. Each amino acid is represented by a 20 component vector. Each vector contains nineteen '0's and one '1' depending on the sequence number of the amino acid as shown in the last column in Table-1. This is an extension of Voss type of representation [18]. This maps each amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y) into 20 binary indicator sequences as shown in Table-1 using

‘1’ or ‘0’ to give the binary representation. Now this 20 component vector is used on a protein sequence of length ‘n’ to get a 20*n component vector of 0 and 1.

Calculation of descriptor

To calculate the descriptor, first of all, FFT is applied on the above represented sequence of length 20*n to get 20*n component complex values using the following formula.

$$U_i(k) = \sum_{n=0}^{N-1} u_i(n) e^{-i(2\pi/N)kn}, k = 0, 1, .., N-1$$

where N is the sample size and k stands for frequency.

Next power spectrum for each component signal $U_i(n)$ is calculated using the following formula.

$$PS_{U_i}(k) = |U_i(k)|^2, k = 0, 1, 2, .., N-1; i = 1, 2, .., 20$$

The Perseval’s identity assumes a simple form given by

$$\frac{1}{N} \sum_{k=0}^{N-1} PSU_i(k) = \sum_{n=0}^{N-1} |u_i(n)|^2 = (Nu_i), i = 1, 2, \ldots, 20.$$

where Nu_i is the number of 1’s present in $u_i(n)$.

Next normalization is done on the components of the power spectrum using the following formula

$$Z_i(k) = \frac{PSU_i(k)}{N(Nu_i)}, \text{where } \sum_{k=1}^{N-1} Z_i(k) = 1$$

Now descriptors are calculated using 2nd order moments for each component of $Z_i(k)$, where the j^{th} order moment of each ($z_i(k)$), is given by

$$\frac{1}{N} \sum_{k=1}^{N-1} |Z_i(k) - A|^j, A = \text{ mean of each } (\mathrm{Z_i(k)}) = \frac{1}{N}, i = 1, 2, .., 20$$

This gives the descriptor of the protein sequence considered. This process is repeated for all the protein sequences to be compared.

Computation of distance matrix- Let S_1 and S_2 be two protein sequences of two different species to be compared. Let D_i and D_j are the two descriptors produced in the last step from S_1 an S_2. Now, distance between two sequences S_1 and S_2 is calculated using Euclidean Distance by the following formula-

$\text{Distance}(D_i, D_j) = (\Sigma(D_i - D_j)^2)^{1/2}$

Now if there are ‘m’ numbers of sequences, the distance between each pair of sequences is measured to form a matrix called distance matrix. Finally, Phylogenetic tree is generated to compare the protein sequences using the UPGMA algorithm. The whole procedure is described by the following algorithm.

Procedure MOMENT

Take m number of protein sequences one by one. Find out the maximum length of the sequences and let it be L. Make all the sequence of length L by appending required number of ‘0’ at the end.

for i=1 to m by 1 do

Take the i^{th} protein sequence of length L.

Represent all the amino acids present in the sequence individually in binary form of representation

using extended Voss typeRepresentation to give a 20 component vector of ‘0’ and ‘1’. So a sequence of length L, will be represented by 20*L size vector.

FFT is used on this binary form and subsequently power spectrum is calculated and normalized on

these represented values of length L. It will produce L/2 frequencies.

Descriptors are calculated using 2nd order moments.

End for

Construct distance matrix from the mnumber of descriptive vectors using Euclidean Distance between

each pair of protein sequences.

Finally construct Phylogenetic Tree from the distance matrix using UPGMA Algorithm.

End MOMENT

Complexity Analysis

The proposed algorithm has following steps-

a) Representation of protein sequence of length L using extended Voss type representation which involved

O(L) in the worse case.

b) Calculation of FFT of a binary sequence of length 20*L which is O(L log L) in the worse case.

c) Calculation of descriptor which is also O(L).

All these steps are executed for m number of sequences, which gives a complexity of O(m*L log L). Without loss of generality, if we assume m and L are equal, then the complexity becomes $O(m^2 \log m)$.

Results And Discussion

The results obtained by the present methods for 9 ND4, 16 ND5, 9 ND6, 12 Baculovirus and 24 TF protein sequences are shown in figure 1 to 5 in the form of Phylogenetic trees, which are similar to their known biological references. For further validation of the result, comparison is made among the results obtained by the present method with those obtained earlier by other methods on the same sequences. It is seen that the present method produces similar results in all the cases, which is discussed below, in details.

Similarity/Dissimilarity Study of ND4, ND5 and ND6 Protein Sequences

Numerous similarity/dissimilarity techniques have been proposed in the study of NADH Dehydrogenase 4 (ND4), 5 (ND5) and 6 (ND6) protein sequences [30–33]. NADH Dehydrogenase protein sequences, specially ND5, are best known for their high mutation rate. So they have been widely used in the study of protein sequences. In this paper, ND4, ND5 and ND6 proteins are used to study their similarity/dissimilarity. In the proposed method, Phylogenetic trees are produced based on extended Voss type representation by applying FFT and subsequently using 2nd order moment as the descriptor. These are shown in figure-1, figure-2 and figure-3 respectively. To check the strength of proposed techniques, a comparison is made with the results obtained by other methods as well. For example, figure- 6 and figure-7 show the phylogenetic trees of ND4 protein sequence obtained by the use of complex representation of Amino Acids under Hydrophobicity & Residue Volumes and Hydrophilicity & Residue Volume respectively as found in [34].It reflects the exact relationship of the species as obtained by the present method. Again, Figure-8 shows the phylogenetic tree of the 16 ND5 protein sequences based on the PCA-FFT analysis [35]. Comparing this with Fig. 2, it is found that the distance between F_whale and B_whale is the smallest. So it may be concluded that they are more similar compared to the other species. Also it is observed that Human – P_chimpanzee and Human- C_chimpanzee have a comparatively closer relationship than Human - Gorilla, which reflects the known biological reference. Again, it is seen that Opossum has the highest distance from all remaining 15 species, confirming that Opossum is exclusively different from all other 15 species. Lastly, figure-9 and figure-10 show the phylogenetic trees of ND6 proteins obtained by the use of complex representation of Amino Acids under Hydrophobicity & Residue Volumes and Hydrophilicity & Residue Volume of ND6 protein sequences as found in[34]. It also reflects the exact relationship of the species as obtained by the present method. The above discussions justify the soundness of the proposed technique.

Similarity/Dissimilarity Study of 12 Baculoviruses Protein Sequences

Present method is also applied in the study of 12 Baculovirus Protein Sequences and the result obtained in the form of phylogenetic tree is shown in figure-4. To claim the correctness of proposed techniques, a

comparison is made with the result obtained by other methods. First it was compared with 2 phylogenetic trees produced in [36] using geometrical centre and Sequence-Segmented Method with k=5 as shown in figure-11. It is found that results obtained by the present method are at par with the results obtained in [36]. Further the result obtained by the present method as shown in figure-4 is compared with the result obtained in [37] using new spectrum-like graph representation as shown in figure-12. It also reflects the same relationship as the present method. This further established our claim.

Similarity Study of 24 TFs

24 Transferrins (TFs) sequences are known as 24 vertebrates taken from the NCBI database. Present paper exhibits the corresponding phylogenetic tree in figure-5. It is found that using the proposed technique TF proteins and Lactoferrin (LF) proteins are separated easily. These results are also compared with that of [38], which is based on Position-Feature Energy Matrix. The same results are obtained in both the cases. These are also at par with the known biological reference. This again adds to the soundedness of our results.

Conclusion

In bioinformatics, protein sequence comparison remains a challenging task specially to understand the relationship between different species. In this respect, lots of techniques have been adopted by different researchers from time to time. Whenever the length of the protein sequence remains the same, it is found that alignment-based methods perform well. Even then results are not dependable in all cases. It can be said that for huge amounts of data, alignment-based techniques are not recommended due to computational complexity. Whereas alignment-free techniques are preferred, as it needs less computational time. Again all alignment free methods are not capable of dealing with sequences of unequal lengths. But the present technique is an alignment-free one, which is applicable to sequences of equal and unequal lengths. It uses a novel binary representation of protein sequences. Moreover, selection of moment vectors in Protein sequence comparison under binary representations is a pioneering attempt in the representation of protein sequences for the study of their similarity/dissimilarity. The present technique is examined on the different groups of species including 9 ND4, 16 ND5, 9 ND6, 12 Baculovirus and 24 TF proteins. On examination of the results obtained, present technique shows exceedingly fine and precise results over alignment-based and also over alignment-free methods used by other authors in the literature. It is worth mentioning that the present technique is very efficient in terms of time complexity which is O(m2 log(m)) where m is the length of the sequence. So it can be concluded that the present technique is well directed and motivated. It is time efficient and produces precise results comparable to those of other existing methods and also their known biological references. Last but not the least, the present technique is capable of dealing with huge amounts of data of the protein sequences of equal and unequal lengths.

Abbreviations

UPGMA: Unweighted Pair Group Method with Arithmetic Mean;

ND4: NADH Dehydrogenase 4

ND5: NADH Dehydrogenase 5

ND6: NADH Dehydrogenase 6

TF: Transferrins

LF: Lactoferrin

PCA: Principal Component Anaysis

FFT: Fast Fourier Transform

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication : Not applicable

Availability of data and materials : The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: I wish to confirm that there has been no financial support for this work that could have influenced its outcome.

Authors' contributions :

JP: Design and development of the work and finalization of draft.

SG: Data collection, analysis and interpretation.

BM: Initial drafting the article.

DKB: Conception of the work and critical revision of the article after final draft.

Acknowledgements: Not applicable

References

1. Edgar, R.C. Muscle: Multiple Sequence Alignment with High Accuracy and High Throughput. Nucleic Acids Res. 2004; doi: 10.1093/nar/gkh340.

2. Katoh K, Misawa K, Kuma KI, Miyata T. Mafft: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. Nucleic Acids Res. 2002. org/10.1093/nar/gkf436.
3. Vinga S, Almeida J. Alignment-Free Sequence Comparison—A Review. Bioinformatics. 2003; doi:1093/bioinformatics/btg005.
4. Nandy A. A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Application to Globin Genes. Current Science. 1994; 309-314.
5. Leong PM, Morgenthaler S. Random Walk and Gap Plots of DNA Sequences. Bioinformatics. 1995; org/10.1093/bioinformatics/11.5.503.
6. Guo X, Randic M, Basak SC. A Novel 2-D Graphical Representation of DNA Sequences of Low Degeneracy. Chem Phys Lett. 2001; doi:1016/S0009-2614(01)01246-5.
7. Yau SS, Wang J, Niknejad A, Lu C, Jin N, Ho YK. DNA Sequence Representation without Degeneracy. Nuc Acids Res. 2003; doi: 10.1093/nar/gkg432
8. Liao B. A 2D graphical representation of DNA sequence. Chem Phys Lett. 2005; 401(1-3):196-199.
9. Liao B, Tan M, Ding K. Application of 2-D Graphical Representation of DNA Sequence. Chem Phys Lett. 2005; org/10.1016/j.cplett.2005.08.079.
10. Song J, Tang H. A New 2-D Graphical Representation of DNA Sequences and Their Numerical Characterization. J biochem biophys methods. 2005; doi: 10.1016/j.jbbm.2005.04.004.
11. Randic M, Vracko M, Lers N, Plavsic D. Novel 2-D graphic representation of DNA sequences and their numerical characterization. Chem Phys Lett. 2003; doi: 1016/S0009-2614(02)01784-0.
12. Randic M, Vracko M, Lers N, Plavsic D. Analysis of Similarity/Dissimilarity of DNA Sequences Based on Novel 2-D Graphical Representation. Chem Phys Lett. 2003; doi: 1016/S0009-2614(03)00244-6.
13. Yao YH, Liao B, Wang TM. A 2D Graphical Representation of RNA Secondary Structures and the Analysis of Similarity/Dissimilarity Based on it. J Mol Structure: THEOCHEM. 2005; doi: 10.1016/j.theochem.2005.08.009.
14. Randic M, Vracko M, Nandy A, Basak SC. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. J Chem Inform Comput Sci. 2000; doi:1021/ci000034q.
15. Nandy A, Nandy P. Graphical Analysis of DNA Sequence Structure: II. Relative Abundances of Nucleotides in DNAs, Gene evolution and Duplication. Curr Sci. 1995; jstor.org/stable/24096174.
16. Yao YH, Nan XY, Wang TM. A New 2D Graphical Representation— Classification Curve and the Analysis of Similarity/Dissimilarity of DNA Sequences. J Mol Str: THEOCHEM. 2006; doi:1016/j.theochem.2006.02.007.
17. Qi ZH, Fan TR. PN-Curve: A 3D Graphical Representation of DNA Sequences and Their Numerical Characterization. Chem Phys Lett. 2007; doi:10.1016/j.cplett.2007.06.029.
18. Voss R F. Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences. Phy. Rev. Lett. 1992; doi:10.1103/PhysRevLett.68.3805.

19. Yu J F, Wang J H, Sun X. Analysis of Similarities/Dissimilarities of DNA Sequences Based on a Novel Graphical Representation. MATCH Commun. Math. Comput. Chem. 2010; 63: 493-512.
20. Liu X Q, Dai Q, Xiu Z, Wang T. PNN-Curve: A New 2D Graphical Representation of DNA Sequences and Its Application. J Theor Biol. 2006; 243(4):555-561.
21. Randic M, Zupan J, Balaban A T. Unique Graphical Representation of Protein Sequences Based on Nucleotide Triplet Codons. Chem Phy Lett. 2004; doi: 1016/j.cplett.2004.08.118.
22. Yu J F, Sun X, Wang J H. TN Curve: A Novel 3D Graphical Representation of DNA Sequence Based on Trinucleotides and Its Applications. J Theor Biol. 2009; doi: 1016/j.jtbi.2009.08.005.
23. Li C, Xing L, Wang X. 2-D Graphical Representation of Protein Sequences and Its Application to Corona Virus Phylogeny. BMB Reports. 2007; doi: 5483/bmbrep.2008.41.3.217.
24. Yu Z G, Anh V, Lau KS. Chaos Game Representation of Protein Sequences Based on the Detailed HP Model and Their Multifractal and Correlation Analyses. J Theor Biol. 2004; doi:10.1016/j.jtbi.2003.09.009.
25. Ghosh S, Pal J, Maji B, Bhattacharya D K. A Sequential Development towards a Unified Approach to Protein Sequence Comparison Based on Classified Groups of Amino Acids. Int J Engg Technol. 2018; doi: 14419/ijet.v7i2.9546.
26. Pal J, Ghosh S, Maji B, Bhattacharya D K. Use of FFT in Protein Sequence Comparison under Their Binary Representations. Comput Mol 2016; doi: 10.4236/cmb.2016.62003.
27. Chi R, Ding K. Novel 4D Numerical Representation of DNA Sequences. Chem Phys Lett. 2005; doi: 1016/j.cplett.2005.03.056.
28. King B R, Aburdene M, Thompson A, Warres Z. Application of Discrete Fourier Inter-Coefficient Difference for Assessing Genetic Sequence Similarity. EURASIP J Bio inform Syst Biol. 2014; doi.org/10.1186/1687-4153-2014-8.
29. Hoang T, Yin C, Zheng H, Yu C, He R L, Yau SS. A New Method to Cluster DNA Sequences Using Fourier Power Spectrum. J Theor Biol. 2015; doi: 1016/j.jtbi.2015.02.026.
30. Yao Y et al. Analysis of similarity/dissimilarity of protein sequences. Proteins. 2008; doi: 10.1002/prot.22110.
31. Lam W, Bacchus F. Learning Bayesian belief networks: An approach based on the MDL principle. Computat. Intell. 1994; 10, 269–293.
32. Xiao X et al. Using complexity measure factor to predict protein subcellular location. Amino Acids. 2005; doi: 10.1007/s00726-004-0148-7.
33. Liao B, Liao B, Sun X, Zeng Q. A novel method for similarity analysis and protein sub-cellular localization prediction. Bioinf. 2010; org/10.1093/bioinformatics/btq521.
34. Pal J, Ghosh S, Maji B, Bhattacharya D K. Protein sequence comparison under a new complex representation of amino acids based on their physio-chemical properties. International Journal of Engineering &Technology. 2018; doi: 14419/ijet.v7i1.9292.

35. Pengyao, P., Xianyou, Z. and Lei, W. Similarities/dissimilarities analysis of protein sequences based on PCA-FFT. Journal of Biological Systems 25: 29-45(2017; doi: 10.1142/S0218339017500024.
36. Yu-hua Y, Fen K, Qi D, Ping-an H. A Sequence-Segmented Method Applied to the Similarity Analysis of Long Protein Sequence. MATCH Commun. Math. Comput. Chem. 2013; 70 431-450.
37. Yuhua Y, Shoujiang Y, Huimin X, Jianning H, Xuying N, Ping-an H, Qi D. Similarity/Dissimilarity analysis of Protein sequences Based on a new spectrum-like Graphical representation. Evolutionary Bioinformatics. 2014; doi: 4137/EBO.S14713.
38. Lulu Y, Yusen Z, Ivan G, Yongtang S, Matthias D. Protein Sequence Comparison Based on Physicochemical Properties and the Position- Feature Energy Matrix. Sci Rep. 2017; 7: 46237.

Tables

Table 1: Representation of Amino Acids

Amino Acids	Abbreviation	Representation
Alanine	A	10000000000000000000
Cysteine	C	01000000000000000000
Aspartic acid	D	00100000000000000000
Glutamic acid	E	00010000000000000000
Phenylalanine	F	00001000000000000000
Glycine	G	00000100000000000000
Histidine	H	00000010000000000000
Isoleucine	I	00000001000000000000
Lysine	K	00000000100000000000
Leucine	L	00000000010000000000
Methionine	M	00000000001000000000
Asparagine	N	00000000000100000000
Proline	P	00000000000010000000
Glutamine	Q	00000000000001000000
Arginine	R	00000000000000100000
Serine	S	00000000000000010000
Tyrosine	T	00000000000000001000
Valine	V	00000000000000000100
Tryptophan	W	00000000000000000010
Threonine	Y	00000000000000000001

Figures

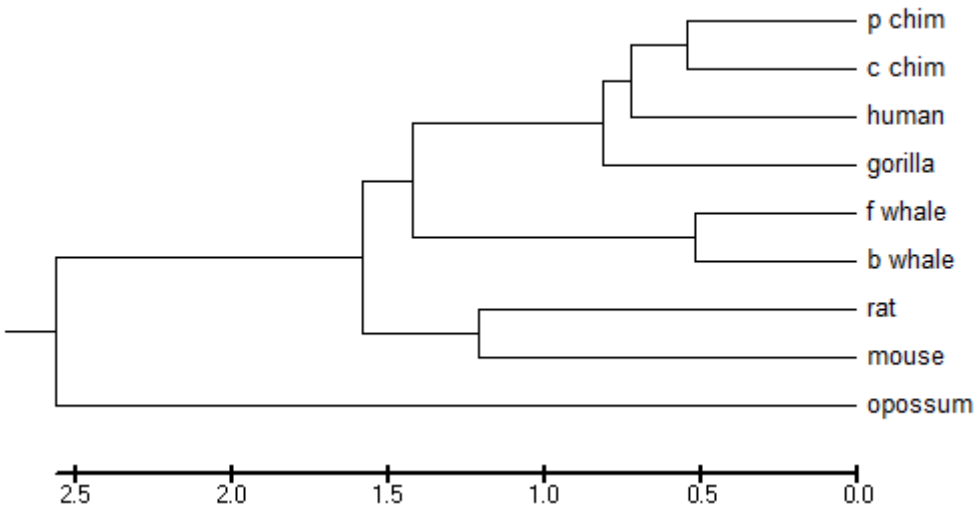

Figure 1

Phylogenetic Tree for 9 ND4 Protein Sequence using present method.

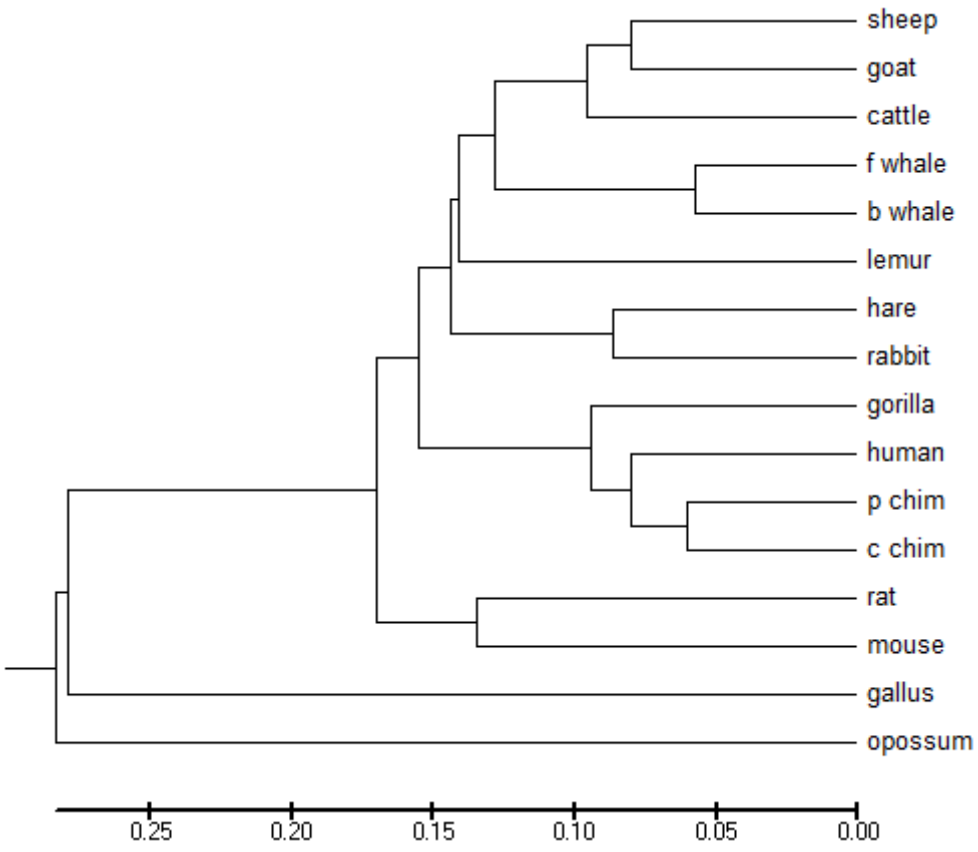

Figure 2

Phylogenetic Tree for 16 ND5 Protein Sequence using present method.

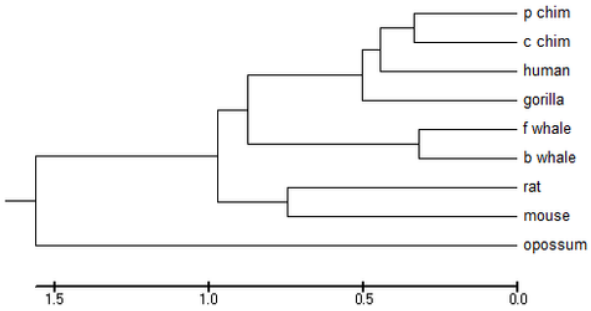

Figure 3

Phylogenetic Tree for 9 ND6 Protein Sequence using present method.

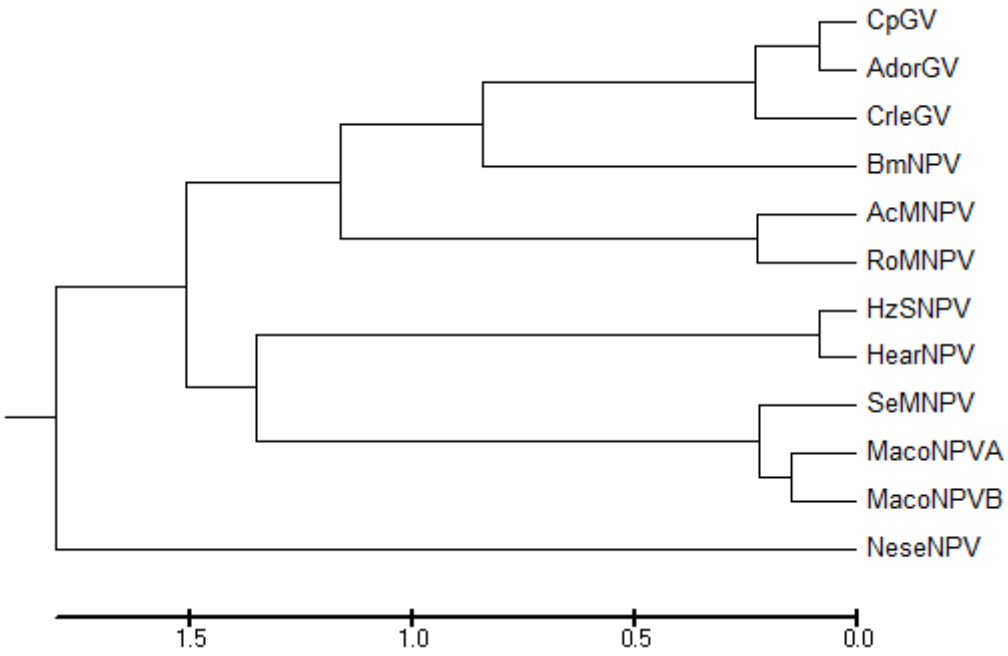

Figure 4

Phylogenetic Tree for 12 baculo virus Protein Sequence using present method.

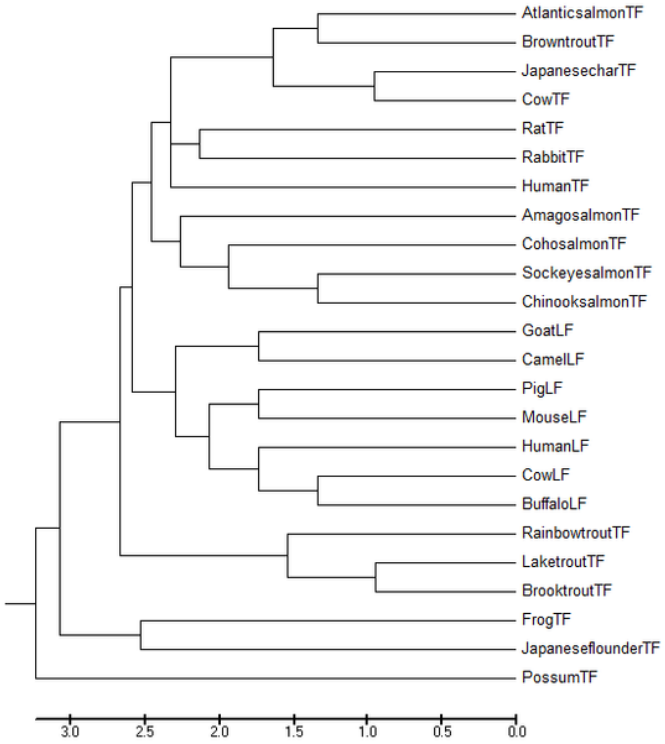

Figure 5

Phylogenetic Tree for 24 TF Protein Sequence using present method.

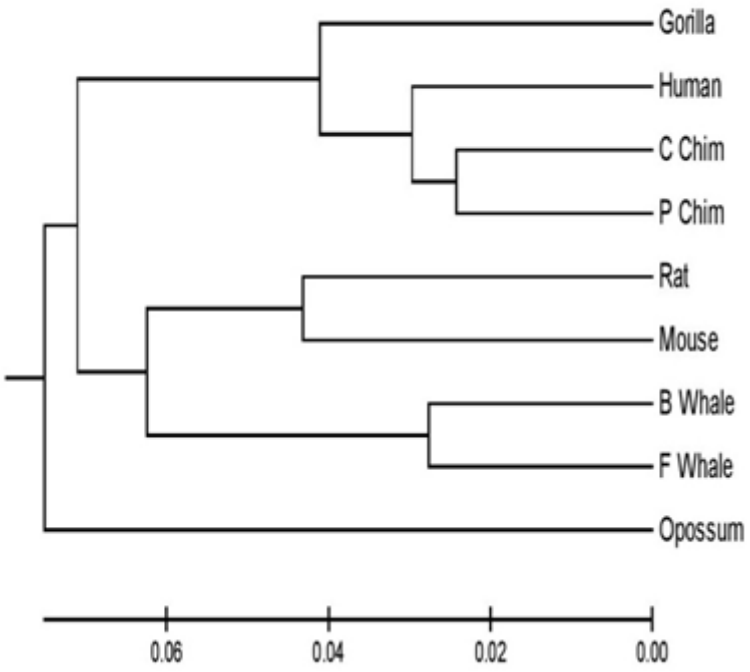

Figure 6

Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids Under Hydrophobicity and Residue Volumes of ND4[34].

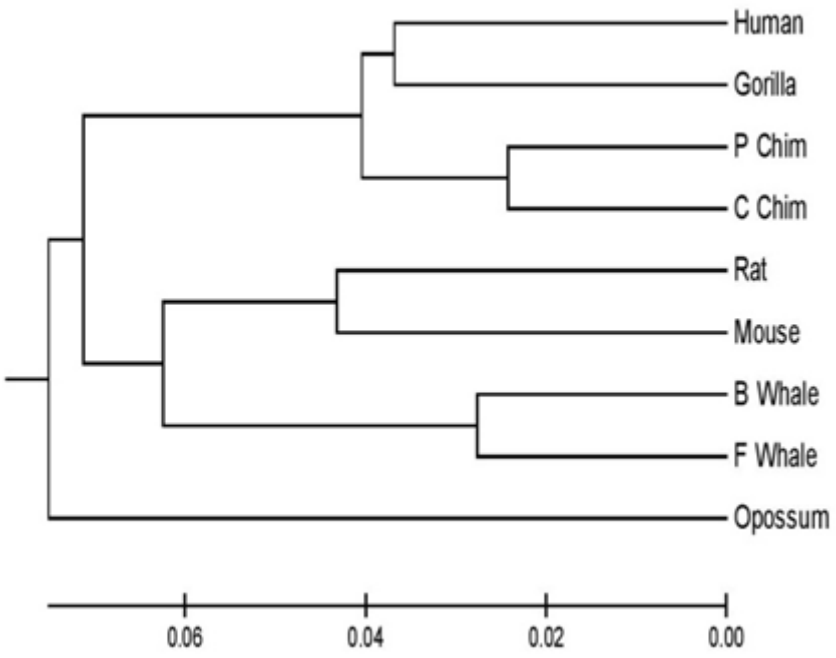

Figure 7

Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids under Hydrophilicity and Residue Volumes of ND4 [34]

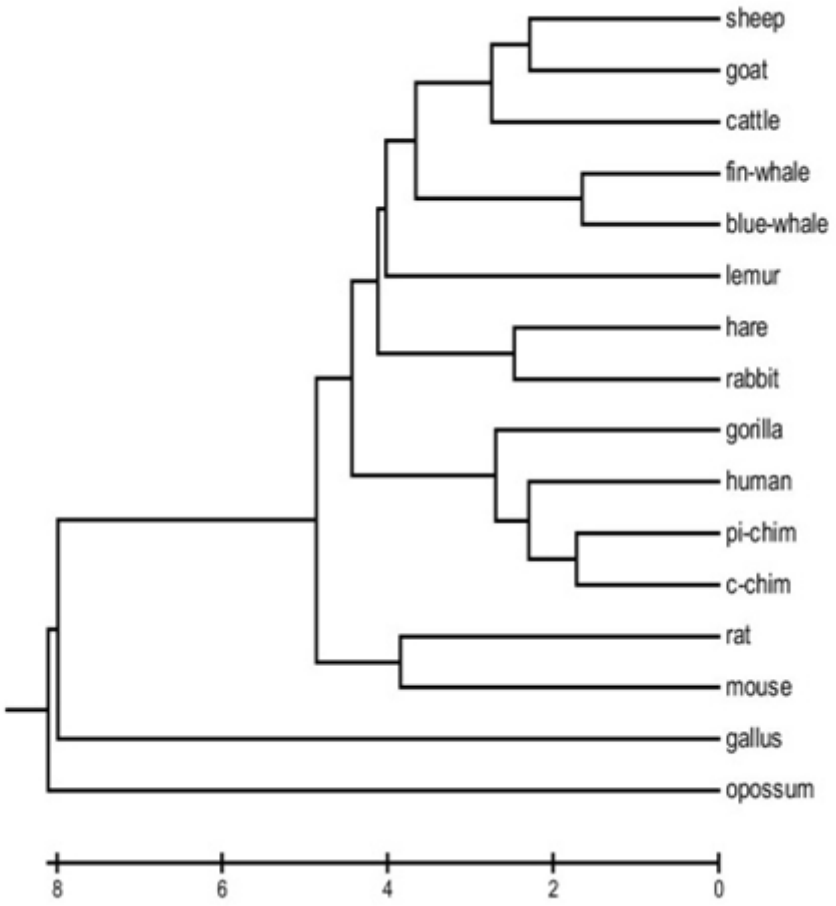

Figure 8

The phylogenetic tree of the 16 species based on the PCA-FFT analysis of their ND5 protein sequences [35].

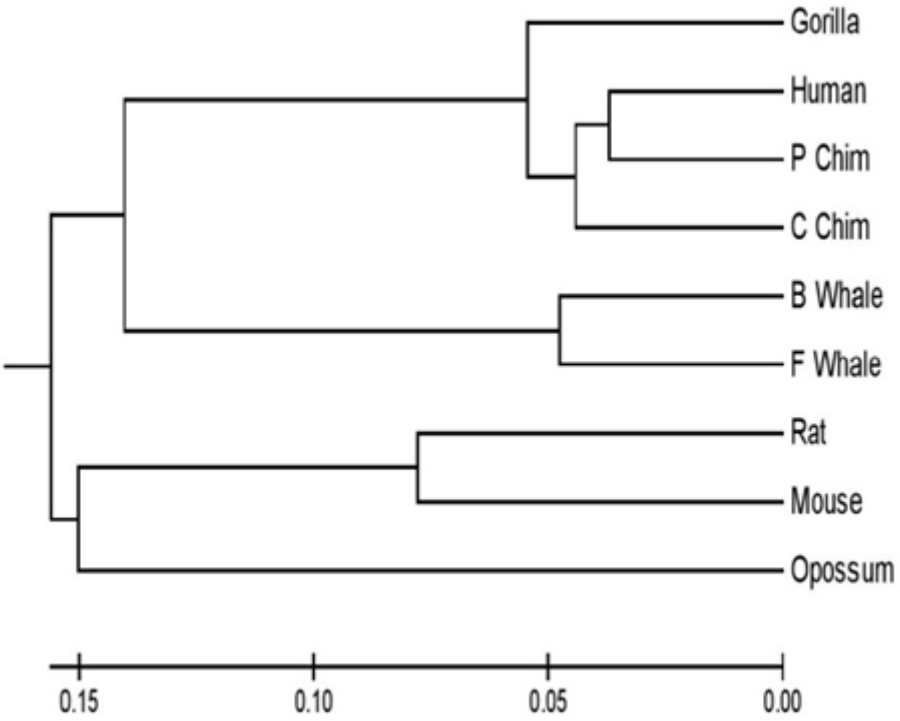

Figure 9

Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids Under Hydrophobicity and Residue Volumes of ND6[34].

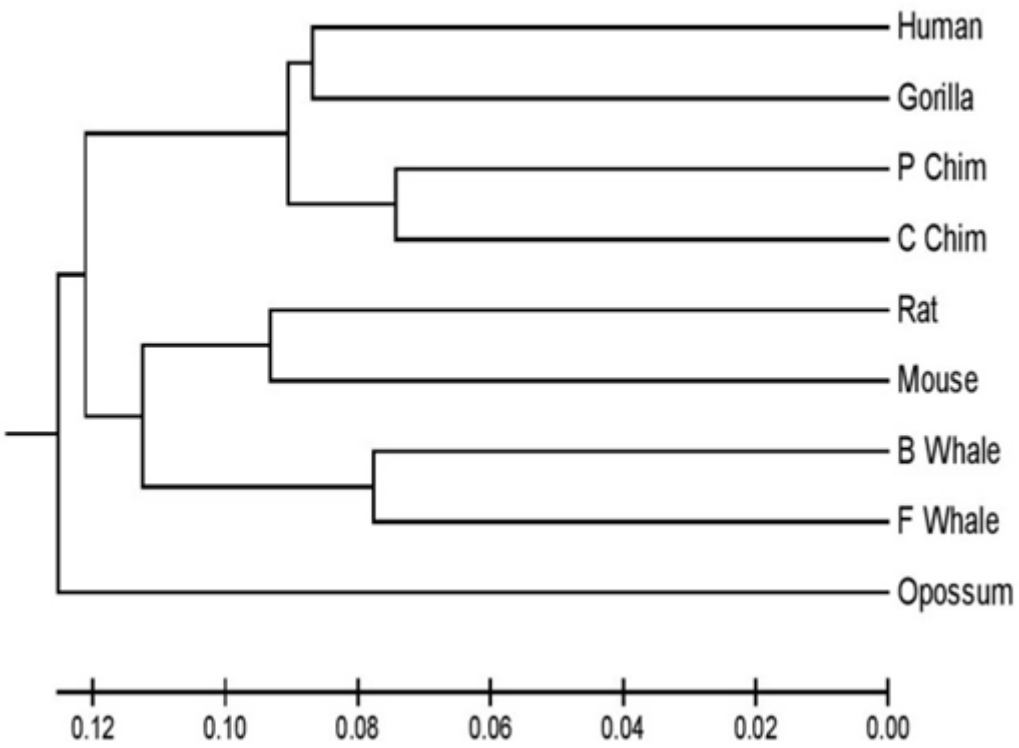

Figure 10

Phylogenetic Tree Obtained by the Use of Complex Representation of Amino Acids Under Hydrophilicity and Residue Volumes of ND6[34].

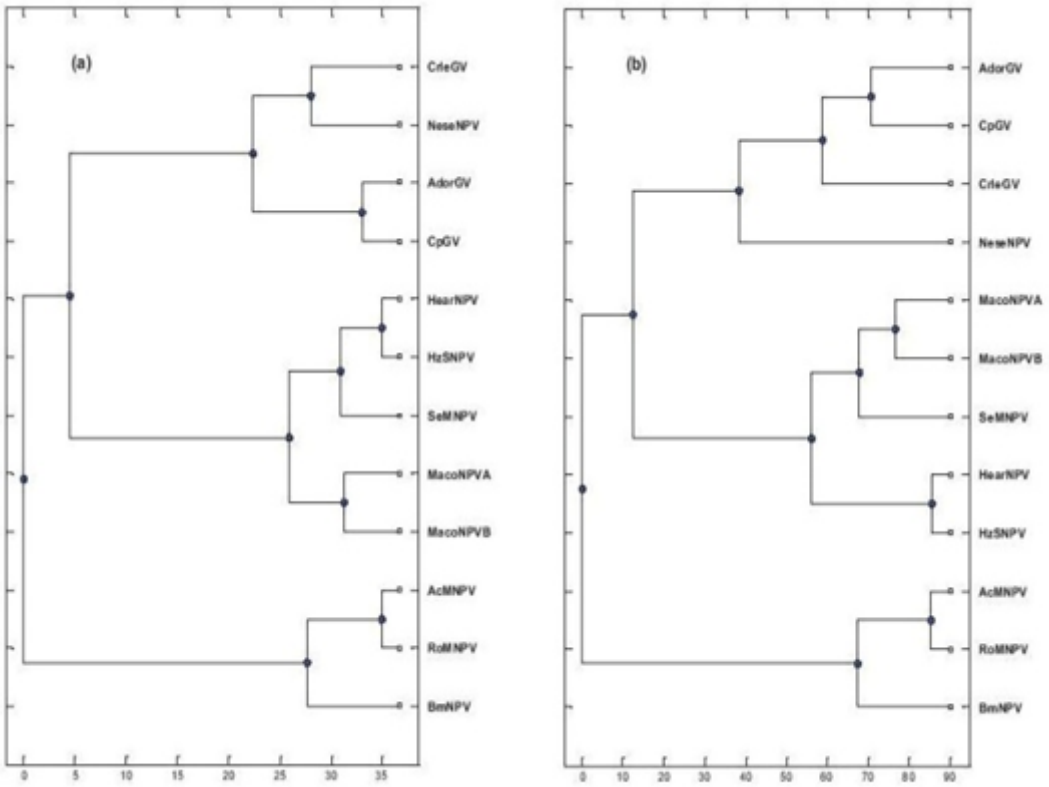

Figure 11

Two phylogenetic tree of 12 baculovirus, (a) geometrical center,(b) Sequence-Segmented Method with k=5 [36].

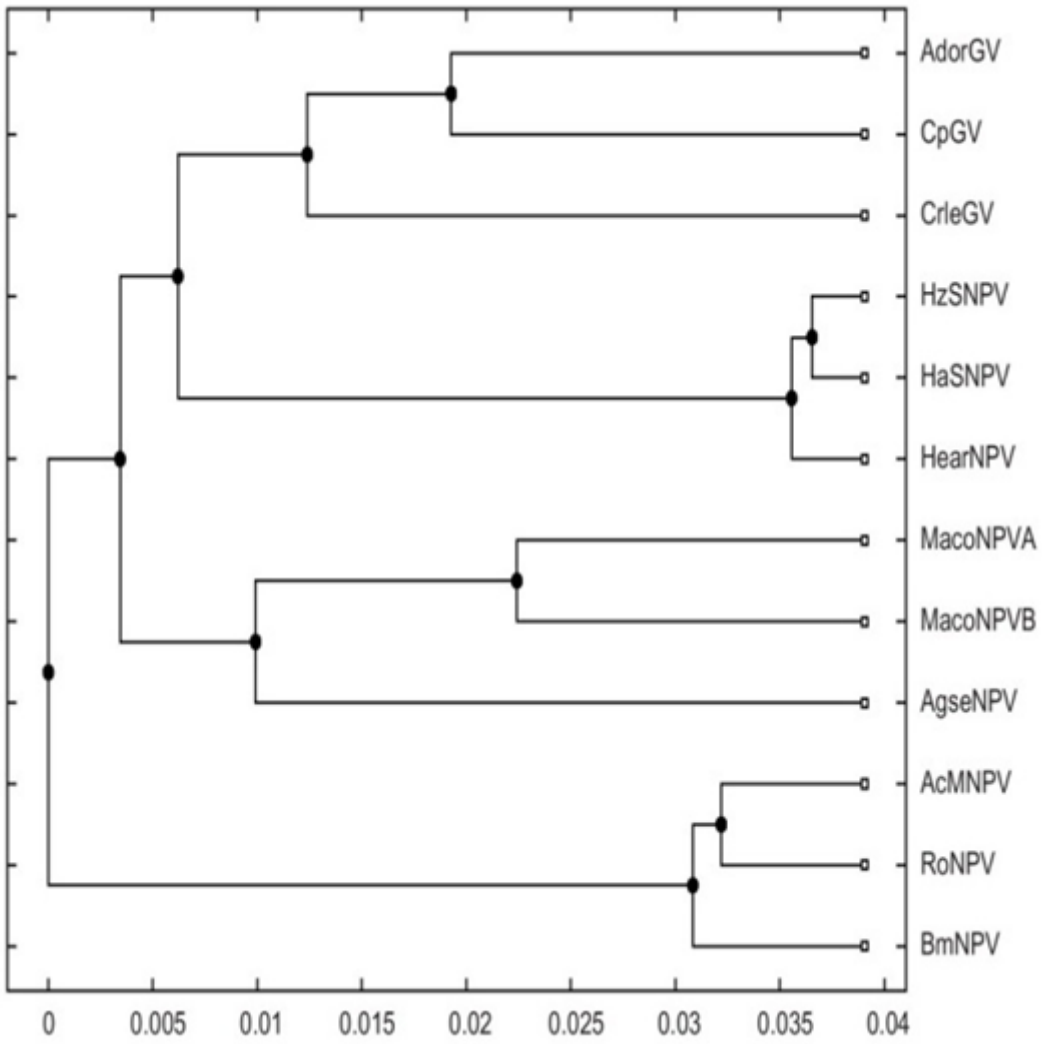

Figure 12

The phylogenetic tree based on protein sequences of 12 baculoviruses based on new spectrum-like graph representation [37]

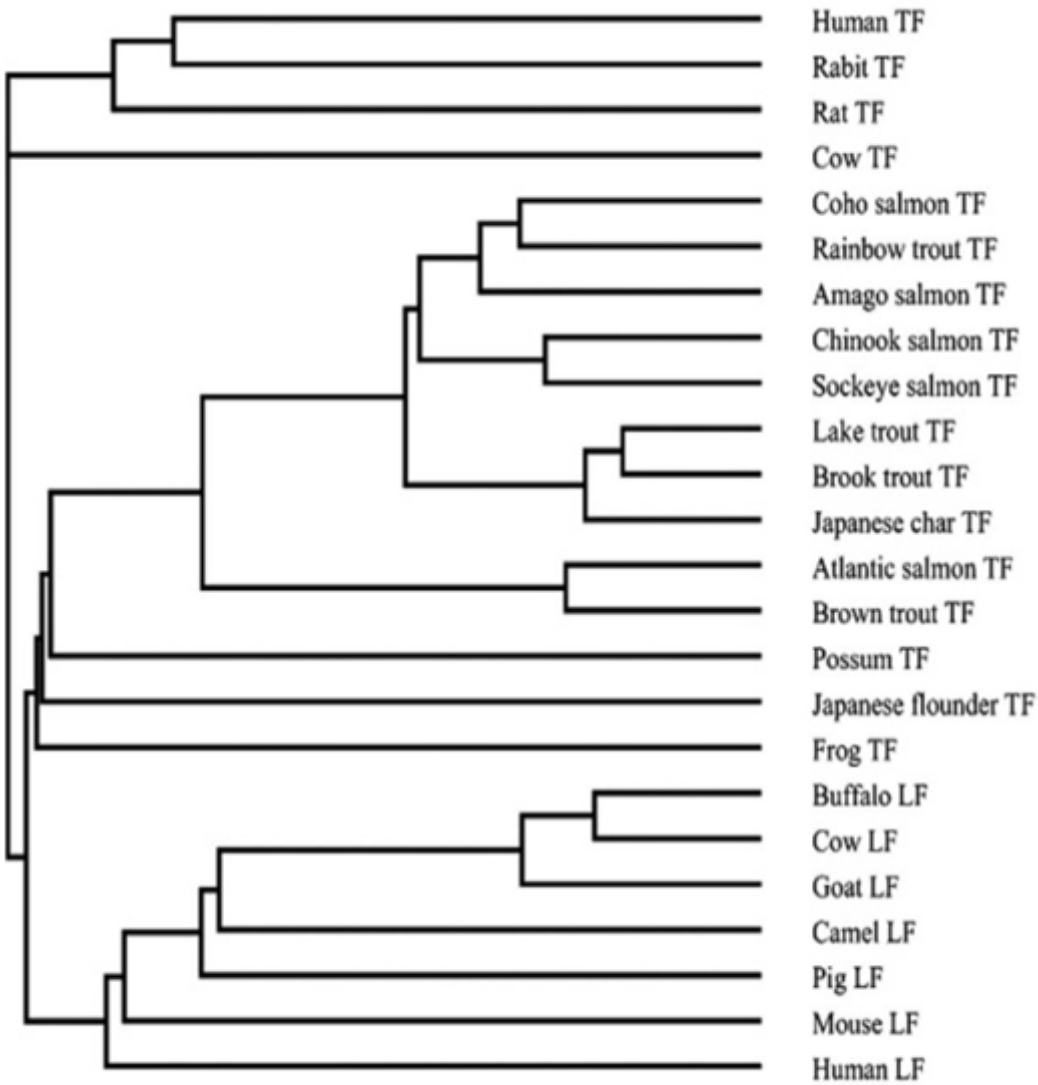

Figure 13

Phylogenetic tree of 24 TFs constructed by Position-Feature Energy Matrix [38].