Southeast Europe Journal of Soft Computing ISSN 2233 – 1859
Available online: http://scjournal.ius.edu.ba

Vol 11, No. 1, March 2022, pp. 07-12

Wavelet Transform-Based Phylogenetic Analysis of Protein Sequences

*[1]Cagin Kandemir-Cavas

[1]Dokuz Eylul University, Turkey

**Corresponding Author*: cagin.kandemir@deu.edu.tr

Article Info

Article history:

Article received on 02 February 2022

Received in revised form 02 March 2022

Keywords:

Bioinformatics; Protein sequence; Phylogenetic tree; Wavelet transform

ABSTRACT: With the acceleration of gene sequencing studies, many biological data have been accumulating. By analyzing these data, it contributes greatly to the studies on understanding the metabolic disorders in the organism and increasing the efficiency of the drugs. For this purpose, it is critical to classify the data in a way that is accurate, fast and low-cost according to its characteristics and relationships. Besides experimental methods, machine learning and bioinformatics methods are commonly used, such as artificial neural networks, support vector machines, soft-computing methods. However, the effectiveness of these methods on biosequence data depends on the method of using the method with the most appropriate parameters and converting protein sequences into numerical sequences. When the sequences are transformed with amino acid frequencies, the properties of amino acids are ignored. For this purpose, handling the physicochemical (hydrophobicity, hydrophilicity ...) properties of amino acids increases the performance of classification techniques. The phylogenetic tree is the best method to visualize the classification among species. In the study, the wavelet transform used in the analysis of digital signals has been adapted to protein sequences defined by hydrophobicity values. Each protein sequence was defined to correspond to a signal, the wavelet transform was divided into approach and detail components, and the similarities between them were calculated, then, the phylogenetic tree of the species was created. As an application, phylogenetic trees of ND5 protein sequences of 22 species were created in the MatlabR2017 program via Neighbor-Joining (NJ) and Unweighed Pair Group Method of Arithmetic Averages (UPGMA).

1. INTRODUCTION

Bioinformatics is defined as the branch of science that uses computer-based applications for biological data. The bioinformatics is widely used for in life-science nowadays. Biosequence data can be DNA or amino acid sequences. The structures and functions of proteins depend on the order of amino acids in the sequences. Proteins with similar amino acid sequences have similar functions [1].

Comparison of proteins is used in protein analysis. Because the structure of a protein can be determined through the amino acid sequence [2]. The similarity between a pair of protein sequences means the similarity between their functions and structures. This similarity can be used to find similar biological functions, structures, and to reveal relationships among organisms [3,4].

In the literature, support vector machines [5–7], artificial neural networks [8, 9], fuzzy logic [10–12], distance-based algorithms [13, 14], hidden Markov models [15], knowledge-based technology [16], statistical based algorithms [17], and multi-class support vector machines [18] have been used to describe the relationship between proteins.

Wavelet theory has been widely used in the intracellular location prediction of apoptotic proteins [19], in defining the similarity model of protein sequences [20, 21], in the functional comparison of proteins [22, 23].

Fourier transform was applied in the classification of protein structures in [24]. In this study, the primary amino acid sequence of the protein sequence was expressed as a signal; the time axis represents the amino acid position values and the frequency axis is the hydrophobicity values of the amino acids. In the next step, each signal is divided into frequencies by Fourier transform. The basic parameters that determine the protein class on the signal allocated to their frequencies were examined and the structure was classified. However, the Fourier transform does not show the time periods it is found although it shows the frequency components of the signal. It is possible to access this information through various functions in wavelet transform calculations. For this reason, in this paper, the wavelet transform used in the analysis of digital signals has been adapted to protein sequences defined by hydrophobicity values. The phylogenetic tree of the species was created by defining each protein sequence to correspond to a signal, dividing it into wavelet transform, approach and detail components and calculating the similarities between them. As an application, phylogenetic trees were created using ND5 protein sequences of 22 species using Neighbor-Joining (NJ) and Unweighed Pair Group Method of Aritmetic Averages (UPGMA) methods.

2. METHODS AND MATERIALS

The method section of the study consists of the following steps as shown in Figure 1.

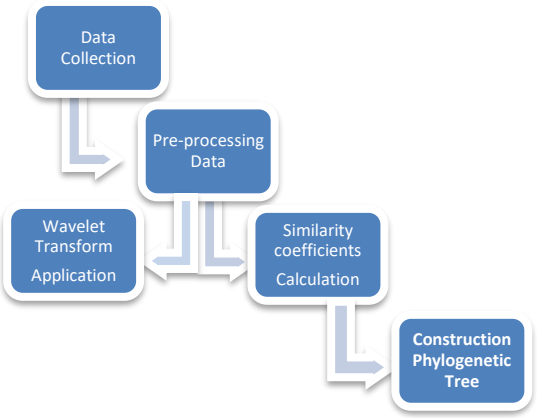


Figure 1: Analytical steps of the study

2.1. Pre-processing Data

To compare the other studies that existed in the literature, ND5 protein sequences of 22 species are chosen and given in Table 1 [25]. The sequences were obtained from the Universal Protein Source Information Base (UniProtKB) database [26] in FASTA format. The FASTA format of a protein is illustrated in Figure 2.

Table 1 Protein sequence information of 22 species

Organism	**Accession Number**
Gibbon	P03919
Horse	A5JYC4
Bornean orangutan	P03918
Fin whale	P24978
Mouse	P03921
Rhino	B7ZHW2
Gorilla	P03917
Gray seal	P38602
Human	P03915
Spain bovine	C5IX37
Tiger	F6KOP7
Cat	P48921
Opossum	P41309
Korean bovine	Q85BD6
Blue whale	P41299
Platypus	Q36459
Pigmy chimpanzee	P03916

Harbor seal	Q00542
Sumatran Orangutan	P92699
Common chimpanzee	Q35648
Rat	P11661
Wallaroo	NP_007404

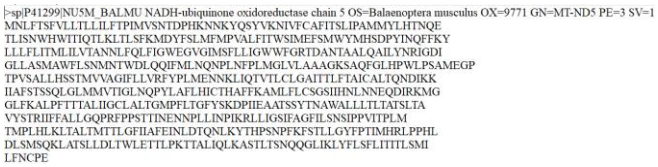


Figure 2: FASTA format of a Protein sequence (eg P41299)

The hydropathy values given in Table 2 for the constituent amino acids of proteins are as follows [27].

Table 2 Hydropathy values of amino acids

Amino Acid	Hydropathy values
I	4.5
V	4.2
L	3.8
F	2.8
C	2.5
M	1.9
A	1.8
G	-0.4
T	-0.7
S	-0.8
W	-0.9
Y	-1.3
P	-1.6
H	-3.2
D	-3.5
N	-3.5
E	-3.5
Q	-3.5
K	-3.9
R	-4.5

ND5 protein sequences of 22 species were transformed into numerical with the help of MATLAB program with these hydropathy values corresponding to each amino acid. Protein sequences that become numerical are expressed as signals using MATLAB. Protein sequences in FASTA format were converted into signals such that their hydrophobicity values were y-axis and amino acid position values were x-axis. In Figure 3, the expression of two protein sequences as signals are given.

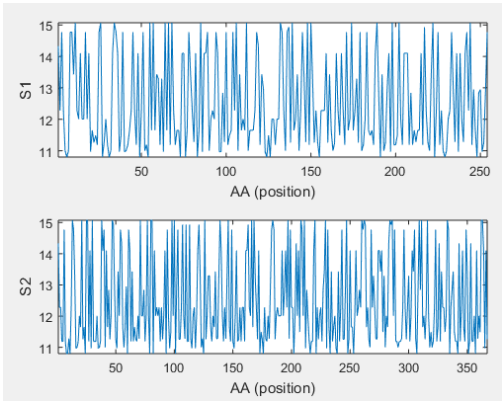


Figure 3: Representation of protein sequences as signals.

2.2. Wavelet Transform

Wavelet transform is frequently used in image processing, signal processing, time-frequency. Although the Fourier transform divides signals with time-frequency notation into frequency components, it does not show in which time slots the signals exist. However, the wavelet transformation eliminates this problem and gives information about frequency components in certain time periods.

Wavelet functions are produced from a source wavelet by changing the shifting and scaling parameters. There are various wavelet functions in the literature. Scaled, scrollable windows for wavelet transform are used throughout the signal and provide spectral behavior information of the signal in the new step. Wavelet transformation examines in narrow frequencies at high frequencies and wide periods in low frequencies as in Figure 4 [28].

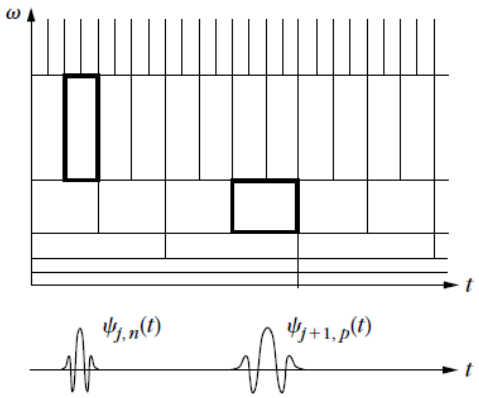


Figure 4: Time-frequency graph in the wavelet transform

The discrete wavelet transform (DWT) can be expressed as detailed in Eq. 1.

$$DWT(s,\tau) = \psi(2^s x(t) + \tau) \quad (1)$$

Here, the signal to be analyzed is $x(t)$ defined as the function to be used for wavelet rotation.$\varphi^*(2^s x(t) + \tau)$. Here s and τ represent the scaling and shifting parameters in the transformation, respectively [28, 29].

In the DWT analysis, the signal is divided into a high-scale-low frequency approach component and a low-scale-high frequency detail component. Low scales indicate high frequency components with sudden changes in the signal. High scales denote slow change, low frequency components in the signal. The decomposition levels of the S signal are given as shown in Eq. 2 and Figure 5.

$$S = A + D$$
$$S = A_1 + D_1$$
$$S = A_2 + D_2 + D_1 \quad (2)$$
$$S = A_3 + D_3 + D_2 + D_1$$

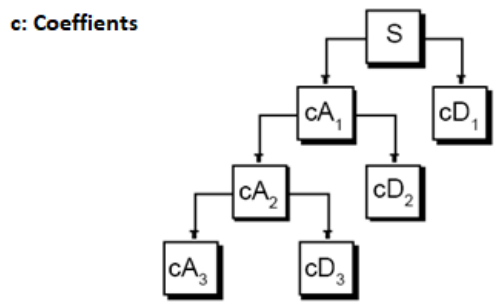


Figure 5: Wavelet Decomposition

Approach components obtained from the signal are separated into approach and detail components at each step, and the algorithm is continued sequentially until the digital value reaches a single value.

Cross-correlation function will be used for similarity between signals. Cross-correlation is defined for S_1 and S_2 signals as seen in Eq. 3:

$$r_j^{12} = \frac{\sum_{n=0}^{N-1} S_1(n) S_2(n-j)}{\sum_{n=0}^{N_1-1} S_1^2(n) \sum_{n=0}^{N_2-1} S_2^2(n)} \ , \ j = 0, \pm 1, \ldots, \quad (3)$$

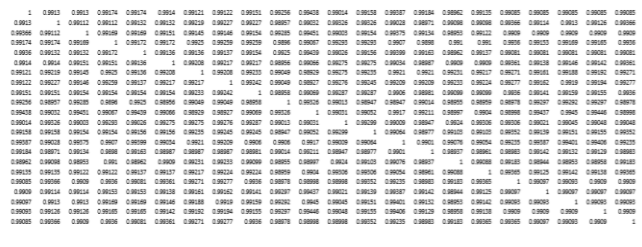
In Eq. 3, N is the length of the intersection of two signals; N_1 indicates the length of the S_1 signal, N_2 indicates the length of the S_2 signal. The greatest absolute value of all correlation coefficients is expressed as the similarity of the two proteins. By converting the calculated similarity values to distance values, the distance matrix between species was calculated. Distance values are calculated as shown in Eq. 4.

$$d_j^{12} = 1 - r_j^{12} \ , j = 0, \pm 1, .., \quad (4)$$

3. RESULTS

Phylogenetic analysis is used to reveal the evolutionary relationship of species to each other. In the study, wavelet analysis was used to calculate the similarities between proteins. The similarity values between species is calculated as a matrix as shown in Table 3.

Table 3 Similarity matrix of 22 species



Phylogenetic trees were created by converting the similarity matrix to the distance matrix with Eq. 4. With the Unweighed Pair Group Method of Aritmetic Averages (UPGMA) and Neighbor-Joining (NJ) methods, phylogenetic trees are created in MatlabR2017 program, as shown in Figure 6 and Figure 7, respectively.

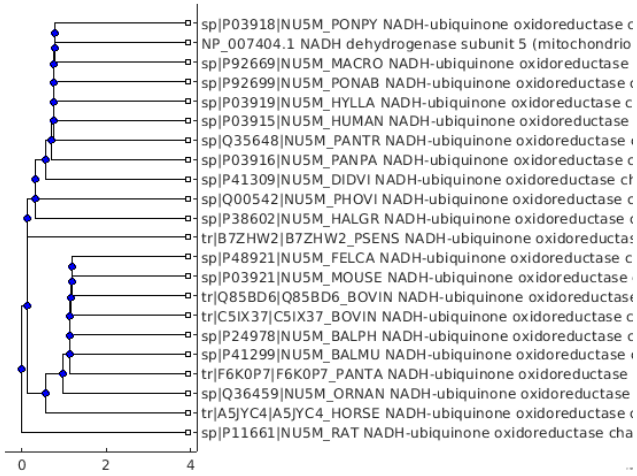


Figure 6: Phylogenetic tree obtained using Unweighed Pair Group Method of Aritmetic Averages (UPGMA) method

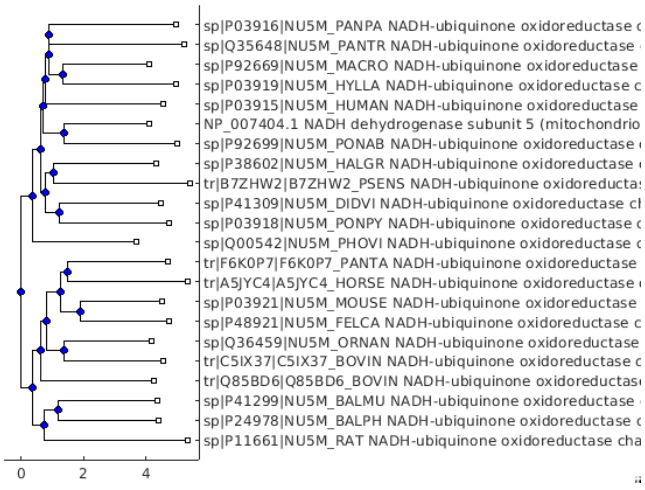


Figure 7: Phylogenetic tree obtained by Neighbor-Joining (NJ) method

In Figure 6, the shortest distances are obtained between the common chimpanzee and the pygmy chimpanzee, the Korean bovine and the Spanish bovine, the blue whale and the fin whale. This result proves that these species are quite evolutionarily close.

In Figure 7, it is seen that human, pygmy chimpanzee, common chimpanzee, gorilla, Sumatran orangutan, gibbon are in the same cluster, blue whale and fin whale are in a separate cluster.

According to the results of both phylogenetic trees, similar species (pygmy and common chimpanzee and Korean and Spanish bovine) were found in very close branches. All these obtained classifications of the species support evolutionary theory.

4. CONCLUSIONS

In this study, it is aimed that the protein sequences that are converted into the signal is separated into their components by wavelet transformation and then classify species based on the similarity between them.

Similarity values between protein sequences belonging to different species were obtained by dividing into detail and approximate components with wavelet transformation applied in signal processing problems of protein sequences defined as signals based on the hydrophobicity values of amino acids. Later, phylogenetic trees were created by converting similarity values into interspecies. The phylogenetic results of the analysis were compared with other studies and similar results were obtained. Thus, the integration of wavelet analysis and protein expression brings a new perspective to phylogenetic studies based on protein similarity.

Acknowledgement

This study was supported by Dokuz Eylul University, Scientific Research Projects Coordination Unit, 2019.KB.FEN.001.

REFERENCES

[1] A. Lesk, "Introduction to bioinformatics," Oxford university press, 2nd edition, New York, USA, 2006.

[2] D. Baker, and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294 no. 5540, pp. 93–96, 2001, doi: 10.1126/science.1065659

[3] M. S. Rosenberg, "Evolutionary distance estimation and fidelity of pair wise sequence alignment," *BMC Bioinformatics*, vol. 6, no. 102, 2005, doi: 10.1186/1471-2105-6-102

[4] D. J., Rigden, and D. J. Rigden, "From protein structure to function with bioinformatics," 2nd ed., Springer, Heidelberg, 2017.

[5] S. Xie, Z .Li, and Hu, H., "Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization," *Gene*, vol. 642, pp. 74–83, 2018, doi: 10.1016/j.gene.2017.11.005.

[6] R. Kumar, A. Srivastava, B. Kumari, and M. Kumar, "Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine," *J. Theor. Biol.*, vol. 365, pp. 96–103, 2015, doi: 10.1016/j.jtbi.2014.10.008.

[7] P. D. Dobson and A. J. Doig, "Distinguishing Enzyme Structures from Non-enzymes Without Alignments," *J. Mol. Biol.*, vol. 330, pp. 771–783, 2003, doi: 10.1016/s0022-2836(03)00628-4.

[8] M. S. Patel, and H. S. Mazumdar, "Knowledge base and neural network approach for protein secondary structure prediction," *J. Theor. Biol.*, vol. 361, pp. 182–189, 2014, doi: 10.1016/j.jtbi.2014.08.005.

[9] M. Can and O. Gürsoy, "Artificial Neural Networks in Bacteria Taxonomic Classification," *Southeast Eur. J. Soft Comput.*, vol. 7, no. 2, pp. 1–7, 2018, doi: 10.21533/scjournal.v7i2.144

[10] W. L. Huang, H. M. Chena, S. F. Hwang, and S. Y. Ho, "Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method," *Biosystems*, vol. 90, pp. 405–413, 2007, doi: 10.1016/j.biosystems.2006.10.004

[11] E. Nasibov, and C. Kandemir-Cavas, "Protein subcellular location prediction using optimally weighted fuzzy k-NN algorithm," *Comput. Biol. Chem.*, vol. 32, no. 6, pp. 448–451, 2008, doi: 10.1016/j.compbiolchem.2008.07.011.

[12] R. Tripathy, D. Mishra, and V. B. Konkimalla, "A novel fuzzy C-means approach for uncovering cholesterol consensus motif from human G-protein coupled receptors (GPCR)," *Karbala Int. J. Mod. Sci.*, vol. 1, no. 4, pp. 212–224, 2015, doi: 10.1016/j.kijoms.2015.11.006.

[13] W. J. Bruno, N. D. Socci, and A. L. Halpern, "Weighted neighbor joining a likelihood-based approach to distance-based phylogeny reconstruction," *Mol. Biol. Evol.*, vol. 17, no.1, pp. 189–197, 2000, doi: 10.1093/oxfordjournals.molbev.a026231

[14] E. Nasibov, and C. Kandemir-Cavas, "Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction," *Comput. Biol. Chem.*, vol. 33, no. 6, pp. 461–464, 2009, doi: 10.1016/j.compbiolchem.2009.09.002.

[15] M. Lasfar, and H. Bouden, "A method of data mining using Hidden Markov Models (HMMs) for protein secondary structure prediction," *Procedia Comput. Sci.*, 127, pp. 42–51, 2018, doi: 10.1016/j.procs.2018.01.096.

[16] C. R. Munteanu, H. Gonzalez-Dıaz, and A. L. Magalhaes, "Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices," *J. Theor. Biol.*, vol. 254, pp. 476–482, 2008, doi: 10.1016/j.jtbi.2008.06.003.

[17] M. Can, " Conformational Parameters for Amino Acids in Helical, β-Sheet, and Random Coil Regions Calculated from Proteins: After 40 Years," *Southeast Eur. J. Soft Comput.*, vol. 4, no. 1, pp. 1–6, 2015, doi: 10.21533/scjournal.v4i1.83

[18] D. Pradhan, S. Padhy, and B. Sahoo, "Enzyme classification using multiclass support vector machine and feature subset selection," *Comput. Biol. Chem.*, vol. 70, pp. 211–219, 2017, doi: 10.1016/j.compbiolchem.2017.08.009.

[19] S. Chaohong, and S. Feng, "Wavelet transform for predicting apoptosis proteins subcellular location," *J. Nat. Sci.*, vol. 15, no. 2, pp. 103–108, 2010, doi: 10.1007/s11859-010-0203-z.

[20] J. Su, and J. Bao, "A wavelet transform based protein sequence similarity model," *Appl. Math. Inf. Sci.*, vol. 7, no. 3, pp. 1103–1110, 2013, doi: 10.12785/amis/070330.

[21] L. Yang, Y. Y. Tang, Y. Lu, and H. Luo, "A Fractal dimension and wavelet transform based method for protein sequence similarity analysis," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 12, no. 2, pp. 348–359, 2015, doi: 10.1109/tcbb.2014.2363480.

[22] C. H. De Trad, Q. Fang, and I. Cosic, "Protein sequence comparison based on the wavelet transform," *Protein Eng.*, vol. 15, no. 3, pp. 193–203, 2002, doi: 10.1093/protein/15.3.193.

[23] S. Zhu, and S. Zhu, "Functional comparisons of proteins using the wavelet packet transform," *10th Int. Conf. Fuzzy Syst. Knowl. Discov.*, pp. 724–729, 2013, doi: 10.1109/fskd.2013.6816290.

[24] J. J. Shu, and K. Y. Yong, "Fourier-based classification of protein secondary structures," *Biochem. Biophys. Res. Commun.*, vol. 485, pp. 731–735, 2017, doi: 10.1016/j.bbrc.2017.02.117.

[25] W. Hou, Q. Pan, Q. Peng, and M. He, "A new method to analyze protein sequence similarity using Dynamic Time Warping," Genomics, vol. 109, pp. 123–130, 2017, doi: 10.1016/j.ygeno.2016.12.002.

[26] A. Bairoch, (2000), "The ENZYME database in 2000. Nucleic Acids Research," vol. 28, pp. 304–305, 2000, doi: 10.1093/nar/28.1.304.

[27] P. K. Ponnuswamy, "Hydrophobic characteristics of folded proteins," *Prog. Bio-phys. Mol. Biol.*, vol. 59, no. 1, pp. 57–103, 1993, doi: 10.1016/0079-6107(93)90007-7.

[28] D. C. Hong, "MATLAB Wavelet Analysis Theory and Application of MATLAB application toolbox series," Defense Industry Pub., 2000, ISBN-13: 978-7118033656

[29] Daubechies I, "Orthonormal Bases of Compactly Supported Wavelets," *Commun. Pure Appl. Math.*, vol. 41, 909–996, 1988, doi: 10.1002/cpa.3160410705.