Araştırma Makalesi / Research Paper
Bayburt Üniversitesi Fen Bilimleri Dergisi
Bayburt University Journal of Science

dergipark.gov.tr/bufbd
ISSN: 2667-579X

Cilt 6, Sayı 1, Yıl 2023, Sayfa 19-28
Vol 6, Issue 1, Year 2023, Pages 19-28
doi.org/10.55117/bufbd.1192229

Geliş Tarihi / Date of Recieve: 20.10.2022
Revizyon Tarihi / Date of Revision: 25.11.2022
Kabul Tarihi / Date of Acceptance: 04.12.2022

# Protein Verilerinin Ayrık Dalgacık Dönüşümü ile Analizi

**Çağın KANDEMİR ÇAVAŞ**[1*]

[1]*Dokuz Eylül Üniversitesi, Bilgisayar Bilimleri Bölümü, İzmir, Türkiye*

***Anahtar Kelimeler:***

*Dalgacık Dönüşümü, Protein Dizileri, Filogenetik, Sınıflandırma*

**Özet**

Biyolojik veri tabanları, genomik ve proteomik çalışmalar nedeniyle büyük miktarda veri içermektedir. Verilerin analizi, organizmadaki metabolik bozuklukların anlaşılmasına ve ilaç keşif çalışmalarının artırılmasına büyük katkı sağlamaktadır. Zaman ve maliyet tasarrufu nedeniyle makine öğrenmesi ve veri analizi yöntemleri bu amaçla sıkça kullanılmaktadır. Yöntemlerin etkinliği, uygun parametre seçimine ve protein dizilerinin kodlanış tipine de bağlıdır. Bu amaçla amino asitlere ait fizikokimyasal özelliklerin dâhil edilmesi kullanılan algoritmanın performansını arttırmaktadır. Filogenetik analiz, türler arasındaki ilişkiyi görselleştirmek için kullanılan en iyi yöntemlerden biridir. Çalışmada, dijital sinyal analizinde kullanılan dalgacık dönüşümü yönteminin, protein dizilerine uyarlanması tasarlanmıştır. Dalgacık dönüşümü kullanılarak 15 türe ait SOD1 protein dizileri arasındaki genetik yakınlık Ağırlıklı Çift Grup Aritmetik Ortalamalar Yöntemi (WPGMA) yöntemiyle belirlenmiştir. Ayrıca, proteinler arası genetik uzaklıkları temel alan Jukes-Cantor (JC) uzaklığı kullanılarak elde edilen filogenetik ağaç ile elde edilen sonuçlar karşılaştırılmış, dalgacık analizi yönteminin türlere ait moleküler boyuttaki ilişkinin ortaya koyulmasında etkinliği ortaya çıkartılmıştır. Türlere ait filogenetik ağaç oluşturma süreleri Dalgacık dönüşümü ile 2.0711178 sn., Jukes-Cantor ile 2.20329 sn. olarak elde edilmiştir. Böylelikle, dalgacık dönüşümü kullanarak tanımlanan filogenetik ağaç oluşturma işlem süresinin mevcut JC yöntemine göre daha kısa olmasının büyük veri analizlerinde avantaj sağlaması beklenmektedir.

*e-posta: cagin.kandemir@deu.edu.tr

Bu makaleye atıf yapmak için:
Çağın KANDEMİR ÇAVAŞ, "Protein Verilerinin Ayrık Dalgacık Dönüşümü İle Analizi", Bayburt Üniversitesi Fen Bilimleri Dergisi, C. 6, s 1, ss. 19-28

How to cite this article:
Çağın KANDEMİR ÇAVAŞ, "Analysis of Protein Data with Discrete Wavelet Transform", Bayburt University Journal of Science, vol. 6, no 1, pp. 19-28

## Analysis of Protein Data with Discrete Wavelet Transform

***Keywords:***

*Wavelet Transform, Protein Sequences, Phylogenetic, Classification*

**Abstract**

Biological databases contain large amounts of data due to genomics and proteomics studies. The analysis of the data makes a great contribution to the understanding of metabolic disorders in the organisms and to improve drug discovery studies. Machine learning and data analysis methods are frequently used for this purpose due to the time and cost savings. The effectiveness of the methods also depends on the appropriate parameter selection and the type of coding of the protein sequences. Therefore, the inclusion of physicochemical properties of amino acids increases the performance of the algorithm used. Phylogenetic analysis is one of the best methods used to visualize the relationship between species. In the study, the wavelet transform used in digital signal analysis was designed to be adapted to protein sequences. Using wavelet analysis, genetic similarity between SOD1 protein sequences of 15 species was determined by Weighted Pair Group Arithmetic Mean Method (WPGMA). In addition, the results obtained with the phylogenetic tree obtained by using the Jukes-Cantor (JC) distance based on the genetic distances between the proteins were compared, and the effectiveness of the wavelet analysis method in revealing the molecular dimension of the species was revealed. The phylogenetic tree construction times of the species were obtained as 2.0711178 sec. with the Wavelet transform and 2.20329 sec. with the Jukes-Cantor. Thus, it is expected that the phylogenetic tree construction process defined by using wavelet transform is shorter than the current JC method, which will provide an advantage in big data analysis.

# 1 GİRİŞ

Biyoenformatik, biyolojik verilerin sınıflandırılması, yorumlanması için hesaplama ve analiz araçlarının uygulanması olarak tanımlanır. Bilgisayar bilimi, matematik, fizik ve biyolojiden yararlanan disiplinler arası bir alandır. Amaca bağlı olarak biyolojik veriler DNA veya amino asit dizileri olabilmektedir. Proteine ait yapılar ve fonksiyonlar, amino asit dizilerinin içerdiği bilgi ile ifade edilebilmektedir. Benzer amino asit dizisine sahip proteinler birbirine benzer işlevlere sahip olmaktadır [1, 2]. Bu nedenle biyoenformatik alanında, protein sınıflandırma konusu önemli bir araştırma alanıdır.

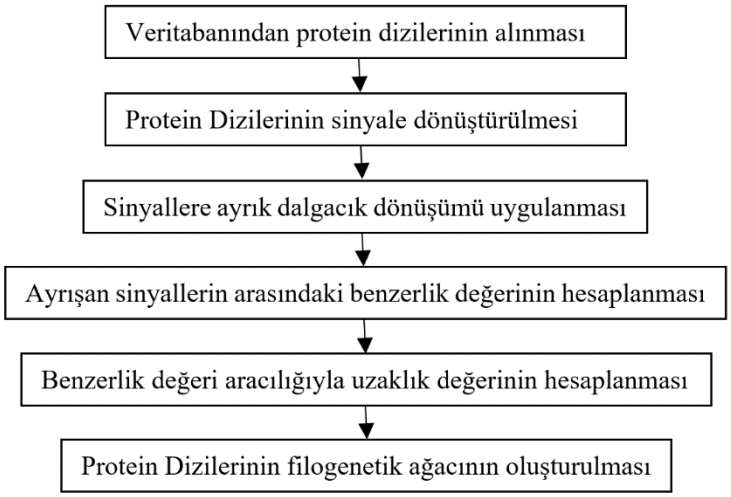
Proteinlerin karşılaştırılması, protein analizinde sıklıkla kullanılmaktadır. Çünkü bir proteinin yapısı, amino asit dizisi aracılığıyla kısmi olarak belirlenebilmektedir [2, 3]. Bir protein dizi çifti arasındaki benzerlik, işlevleri ve yapıları arasındaki benzerlik anlamına gelmektedir. Bu benzerlik, benzer biyolojik işlevleri, yapıları bulmak, organizmalar arasındaki ilişkileri ortaya koymak için kullanılabilmektedir [4, 5].

Literatürde, proteinler arasındaki ilişkiyi tanımlamak amacıyla ayırma analizi [6-9], yapay sinir ağları [10, 11], bulanık mantık [12, 14] ve genetik algoritmalar [15-17] gibi bilgi-tabanlı teknoloji kullanılmaktadır. Dalgacık analizi, biyoenformatik alanında protein kodlama bölgelerinin belirlenmesinde [18,19], protein dizilerinin benzerlik modelini tanımlamada [20], protein hücre içi bölge tayininde [21-23] kullanılmıştır.

Shu ve Yong [24] tarafından yapılan çalışmada protein yapılarının sınıflandırılmasında Fourier dönüşümü uygulanmıştır. Yapılan bu çalışmada protein dizisinin birincil amino asit dizisi, zaman ekseni amino asit konum değerlerine, frekans ekseni amino asitlerin hidropati değerlerine karşılık gelecek şekilde sinyal olarak ifade edilmiştir. Bir sonraki aşamada, her bir sinyal Fourier dönüşümü ile frekanslarına ayrılmıştır. Frekanslarına ayrılan sinyal üzerinde protein sınıfını belirleyen temel parametreler incelenmiş ve yapı sınıflandırılması yapılmıştır. Ancak bilindiği üzere Fourier dönüşümü, sinyale ait frekans bileşenlerini göstermesine rağmen bulunduğu zaman dilimlerini göstermemektedir. Dalgacık dönüşümü hesaplamalarında yer alan çeşitli fonksiyonlar vasıtasıyla bu bilgiye erişmek mümkün olmaktadır. Bu sebeple, çalışmada, dijital sinyallerin analizinde kullanılan dalgacık dönüşümü, hidropati değerleri ile tanımlanan protein dizilerine adapte edilmiştir. Her bir protein dizisinin bir sinyale karşılık gelecek şekilde tanımlanmış, dalgacık dönüşümü ile yaklaşım ve detay bileşenlerine ayrılmış şekilde aralarındaki en yüksek korelasyon olan bileşenler üzerinden benzerlikler hesaplanarak türlere ait filogenetik ağaç oluşturulmuştur. Uygulama olarak 15 türe ait SOD1 protein dizileri Ağırlıklı Çift Grup Aritmetik Ortalamalar Yöntemi (WPGMA) yöntemleri kullanılarak filogenetik ağaçları oluşturulmuştur.

# 2 MATERYAL VE METOD

Protein dizileri arasındaki genetik yakınlığın ifadesi filogenetik ağaçlar vasıtasıyla yapılmaktadır. Uzaklık-tabanlı filogenetik ağaç yöntemleri, dizilerin arasındaki uzaklık temel alınarak oluşturulmaktadır. Dalgacık dönüşümü kullanarak protein verilerinin aralarındaki ilişkinin ortaya koyulması amacıyla çalışma 6 adımdan oluşmaktadır. İşlem adımları Şekil 1'de gösterilmiştir.




**Şekil 1.** Dalgacık dönüşümü ile protein dizilerinin filogenetik analizi işlem adımları

## 2.1 Verilerin hazırlanması

Tablo 1'de verilen 15 türe ait SOD1 protein dizileri Evrensel Protein Kaynağı veri tabanı (UniProtKB) veri tabanından elde edilmiştir [25]. Evrensel Protein Kaynağı veri tabanı, proteinler hakkında gerekli olabilecek tüm bilgilerin elde edilebileceği bir veri tabanıdır. Veri tabanı, proteinin adı, yapısı, fonksiyonu, amino asit dizisi, hücre içi bölge, biyolojik ontolojileri, ait olduğu aile grubu bilgilerini içermektedir. Veri tabanından elde edilen veriler Şekil 2'de gösterildiği gibi FASTA formatta kaydedilmiştir.

**Tablo 1.** 15 türe ait protein dizi bilgileri

| Kaynak Organizma | Protein No |
|---|---|
| İnek | P00442 |
| İnsan | P00441 |
| Şıçan | P07632 |
| Fare | P08228 |
| Kılıçbalığı | P03946 |
| At | P00443 |
| Tavuk | P80566 |
| Deve | H6BDU4 |
| Orangutan | Q8HXQ4 |
| Koyun | P09670 |
| Tavşan | P09212 |
| Şempanze | P60052 |
| Geyik | O46412 |
| Yunus | A0A340WIW6 |
| Keçi | Q5FB29 |

```
>sp|P80566|SODC_CHICK Superoxide dismutase [Cu-Zn] OS=Gallus gallus OX=9031 GN=SOD1 PE=1 SV=3
MATLKAVCVMKGDAPVEGVIHFQQQGSGPVKVTGKITGLSDGDHGFHVHEFGDNTNGCTS
AGAHFNPEGKQHGGPKDADRHVGDLGNVTAKGGVAEVEIEDSVISLTGPHCIIGRTMVVH
AKSDDLGRGGDNESKLTGNAGPRLACGVIGIAKC
```

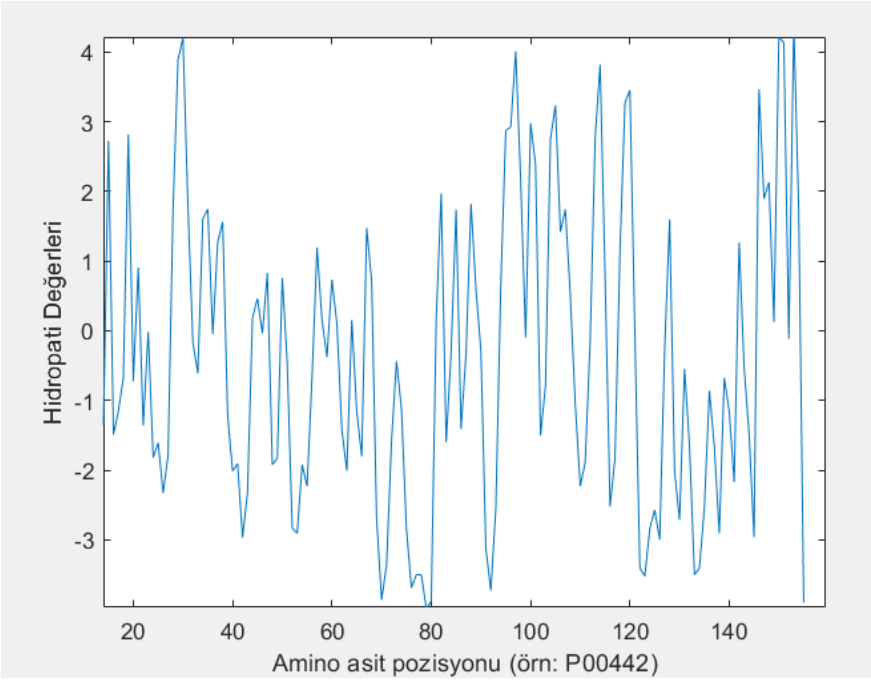
**Şekil 2.** Protein dizilerinin (örn: P80566) FASTA formatındaki gösterimi

Proteinlerin yapıtaşı amino asitler için Tablo 2'de verilen hidropati değerleri aşağıdaki gibidir [26]. Proteinlerin sayısal değere dönüştürülmesi için, protein dizisindeki her amino asidin pozisyonuna karşı gelen hidropati değerleri kullanılmaktadır.

**Tablo 2.** Amino asit hidropati değerleri

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

15 türe ait SOD1 protein dizileri MATLAB R2020b programı yardımıyla her bir amino asite karşılık gelen bu hidropati değerleri ile sayısal hale dönüştürülmüştür. Sayısal hale gelen protein dizileri sinyal olarak ifade edilmiştir. FASTA formatındaki protein dizileri, hidropati değerleri y-ekseni, amino asit konum değerleri x-ekseni olacak şekilde sinyale çevrilmiştir. Şekil 3'te protein dizisinin sinyal olarak gösterimi verilmiştir.




**Şekil 3.** Bir protein dizisinin (P00442) sinyal olarak gösterimi

## 2.2 Dalgacık dönüşümü

Dalgacık dönüşümü, görüntü işleme, sinyal işleme, zaman-frekans konularında sıkça kullanılmaktadır [27]. Fourier dönüşümü, zaman-frekans gösterimli sinyalleri frekans bileşenlerine ayırır ancak hangi zaman dilimlerinde olduklarını belirtmemektedir. Ancak dalgacık dönüşümü bu sorunu ortadan kaldırarak belirli zaman dilimlerinde frekans bileşenlerinin bilgisini de vermektedir.

Dalgacık fonksiyonları, kaydırma ve ölçekleme parametrelerinin değişimiyle kaynak bir dalgacıktan üretilmektedir. Literatürde çok çeşitli dalgacık fonksiyonları bulunmaktadır. Dalgacık dönüşümü için ölçeklendirilmiş, kaydırılabilen pencereler sinyal boyunca kullanılmakta ve yeni adımda sinyalin spektral davranış bilgisini vermektedir. Dalgacık dönüşümü yüksek frekanslarda dar zaman diliminde, düşük frekanslarda ise geniş zaman diliminde inceleme yapmaktadır [28].

Ayrık dalgacık dönüşümünü (DWT) aşağıdaki Formül (1)'deki gibi ifade edebiliriz,

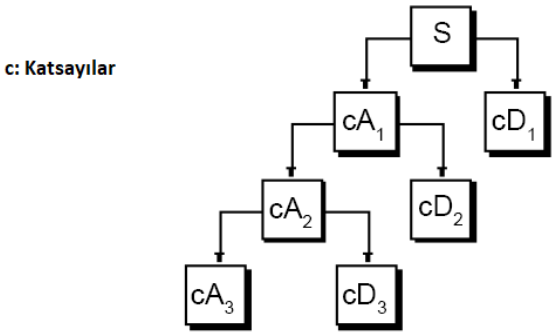
$$DWT(s,\tau) = \psi(2^s x(t) + \tau) \tag{1}$$

Buradaki $x(t)$ analizi yapılacak sinyali, $\varphi^*(2^s x(t) + \tau)$ dalgacık dönüşümü için kullanılacak fonksiyon, $s$ ve $\tau$ sırasıyla dönüşümdeki ölçekleme ve kaydırma parametrelerini temsil etmektedir [27,28].
DWT analizinde, sinyal yüksek ölçekli-düşük frekanslı yaklaşım bileşenine ve düşük ölçekli-yüksek frekanslı detay bileşenine ayrılmaktadır. Düşük ölçekler, sinyaldeki ani değişimlere sahip yüksek frekanslı bileşenlerini

gösterir. Yüksek ölçekler, sinyaldeki yavaş değişimli, düşük frekanslı bileşenleri ifade etmektedir. S sinyalinin ayrışma seviyeleri aşağıdaki Formül (2)'de verilmekte olup A, yaklaşım, D, detay frekanslarını göstermektedir.

Şekil 4'te gösterildiği gibi, sinyalden elde edilen yaklaşım bileşenleri her adımda yaklaşım ve detay bileşenlerine ayrılarak, algoritma ardışık bir şekilde dijital değer tek bir değere ulaşana dek devam edilir.

$$\begin{aligned} S &= A + D \\ S &= A_1 + D_1 \\ S &= A_2 + D_2 + D_1 \\ S &= A_3 + D_3 + D_2 + D_1 \end{aligned} \quad (2)$$



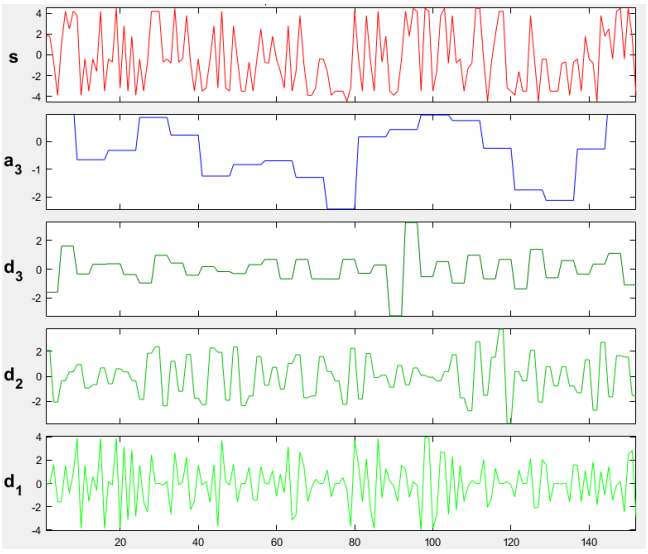

**Şekil 4.** Dalgacık Ayrışımı

Sinyaller arasındaki benzerlik için çapraz-korelasyon fonksiyonu kullanılacaktır. Çapraz-korelasyon, $S_1$ ve $S_2$ sinyalleri için Formül (3)'teki gibi tanımlanmaktadır:

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

Formül (3)'teki $N, j$ gecikmeli iki sinyalin kesişimlerinin uzunluğunu; $N_1$, $S_1$ sinyalinin uzunluğunu, $N_2$, $S_2$ sinyalinin uzunluğunu göstermektedir. Tüm korelasyon katsayılarının mutlak değeri en büyük olanı iki proteinin benzerliği olarak ifade edilir. Hesaplanan benzerlik değerleri uzaklık değerlerine dönüştürülerek türler arası uzaklık matrisi Formül (4)'teki gibi hesaplanmıştır.

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




**Şekil 5.** Üç seviyeli Haar ayrık dalgacık dönüşümü aracılığıyla bir protein dizisi (P00442) için yaklaşım ($a_3$) ve detay bileşenleri ($d_1$, $d_2$, $d_3$)

Çalışmada, sinyal olarak tanımlanan protein dizilerine 3 seviyeli Haar ayrık dalgacık dönüşümü uygulanarak, her bir sinyal 3. seviye yaklaşım ve 1'den 3. Seviyeye kadar detay bileşenlerine ayrıştırılmıştır. P00442 protein dizisi

için Şekil 5'te yaklaşım ($a_3$) ve detay ($d_1$, $d_2$, $d_3$) bileşenleri gösterilmektedir. Korelasyon, bir durumun diğerine karşı bağımlılık derecesini nicelleştirdiği için, çapraz korelasyon katsayıları elde edilen her seviyede hesaplanarak maksimum korelasyon değerleri elde edilmiş ve benzerlik matrisi elde edilmiştir. Buradan sonraki aşamada ise, benzerlik değerleri kullanılarak iki protein dizisi arasındaki uzaklık değerleri hesaplanmış ve tüm proteinlerin birbirleri ile arasındaki uzaklık değerlerini belirten uzaklık matrisi elde edilmiştir.

### 2.3 Jukes-Cantor uzaklığı

Genetik yakınlık oranı zaman ekseni boyunca sabit ise iki protein dizisi arasındaki uzaklık artmaktadır. Biyoenformatikte, uzaklığı ölçmenin en basit yolu, iki sekans arasındaki farklı sitelerin oranıdır. Bu p değeri Formül (5)'teki gibi hesaplanmaktadır.

$$p = \frac{D}{L} \tag{5}$$

$D$: iki dizinin farklı olduğu konumların sayısı
$L$: iki dizinin her birinin uzunluğu

Bazı genetik yakınlık tabanlı modeller özel olarak oluşturulmuştur. Bu modellerden en yaygın kullanılanı Jukes-Cantor (JC) modelidir [1]. JC uzaklık modeli, tüm sitelerin bağımsız olduğunu ve aynı mutasyon oranlarına sahip olduğunu ve olası tüm nükleotid yer değiştirmelerinin aynı anda gözlendiğini varsaymaktadır. JC uzaklık değeri Formül (6)'daki gibi hesaplanmaktadır.

$$d_{JC} = -\frac{3}{4} ln\left(1 - \frac{4}{3}p\right) \tag{6}$$

## 3 BULGULAR

Filogenetik analiz, türlerin birbirleriyle olan genetik yakınlık ilişkisini ortaya koymak amacıyla kullanılmaktadır. Çalışmada, proteinlerin arasındaki benzerlikleri hesaplamak için dalgacık dönüşümü kullanılmıştır. Dalgacık dönüşümü ile proteinlerin aralarındaki benzerlik değerleri Formül (3) kullanılarak, Tablo (3)'deki gibi matris olarak elde edilmiştir. Benzerlik matrisinin, simetrik bir matris olması nedeniyle, Tablo (3) alt üçgen matris şeklinde ifade edilmiştir.

Tablo (3) ve Tablo (4)'te satırlar ve sütunlar 15 adet türe ait proteinlerdir.

**Tablo 3.** 15 türe ait benzerlik matrisi

| | P00442 | P00441 | P07632 | P08228 | P03946 | P00443 | P80566 | H6BDU4 | Q8HXQ4 | P09670 | P09212 | P60052 | O46412 | A0A340WIW6 | Q5FB29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **P00442** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P00441** | 0.9507 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P07632** | 0.7948 | 0.7658 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P08228** | 0.7491 | 0.7329 | 0.5253 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P03946** | 0.7797 | 0.7891 | 0.963 | 0.526 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P00443** | 0.7716 | 0.774 | 0.5036 | 0.876 | 0.5036 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P80566** | 0.3027 | 0.354 | 0.1656 | 0.3456 | 0.1979 | 0.4205 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **H6BDU4** | 0.7447 | 0.7465 | 0.5101 | 0.8584 | 0.5116 | 0.8805 | 0.3991 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Q8HXQ4** | 0.7569 | 0.7548 | 0.5142 | 0.8496 | 0.5156 | 0.8732 | 0.399 | 0.9888 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P09670** | 0.9228 | 0.8908 | 0.7156 | 0.7228 | 0.7064 | 0.7336 | 0.2988 | 0.7386 | 0.7466 | 1 | 0 | 0 | 0 | 0 | 0 |
| **P09212** | 0.7775 | 0.7821 | 0.9543 | 0.5172 | 0.9898 | 0.4927 | 0.1952 | 0.4987 | 0.5037 | 0.7032 | 1 | 0 | 0 | 0 | 0 |
| **P60052** | 0.7491 | 0.7329 | 0.5253 | 1 | 0.526 | 0.876 | 0.3456 | 0.8584 | 0.8496 | 0.7228 | 0.5172 | 1 | 0 | 0 | 0 |
| **O46412** | 0.3144 | 0.3129 | 0.2763 | 0.3772 | 0.2519 | 0.3661 | 0.1574 | 0.3481 | 0.3526 | 0.3276 | 0.244 | 0.3772 | 1 | 0 | 0 |
| **A0A340WIW6** | 0.7549 | 0.7594 | 0.9362 | 0.4931 | 0.9721 | 0.4687 | 0.1916 | 0.4745 | 0.4792 | 0.6801 | 0.9813 | 0.4931 | 0.2659 | 1 | 0 |
| **Q5FB29** | 0.4962 | 0.5127 | 0.4236 | 0.6818 | 0.442 | 0.6211 | 0.4931 | 0.6491 | 0.656 | 0.4676 | 0.4403 | 0.6818 | 0.2975 | 0.4204 | 1 |

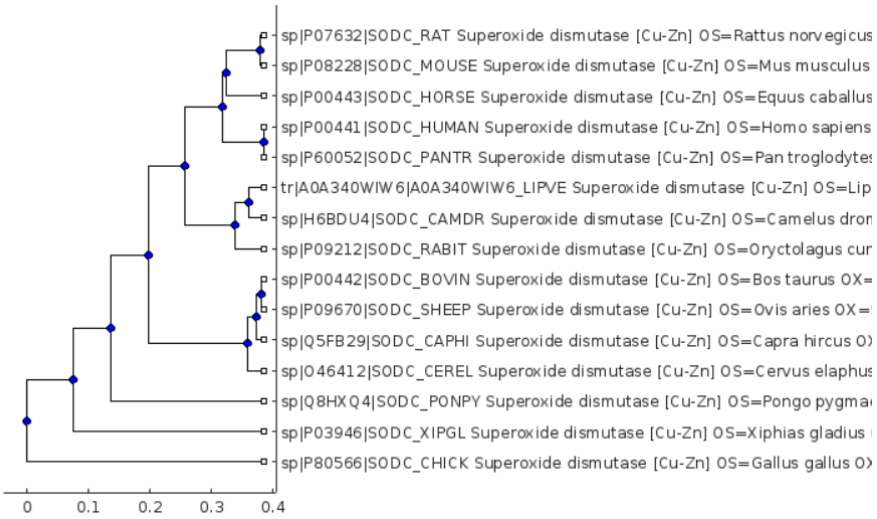
Protein dizileri arasındaki uzaklık değerleri, Formül (4) kullanılarak elde edilmiş, Tablo (4)'teki gibi uzaklık matrisine dönüştürülmüştür. Uzaklık matrisinin simetrik bir matris olması nedeniyle, Tablo (4) alt üçgen matris şeklinde ifade edilmiştir.

Tablo (4)'teki uzaklık matrisi kullanılarak, Ağırlıklı Çift Grup Aritmetik Ortalamalar Yöntemi (WPGMA) ile filogenetik ağaç Şekil 6'daki gibi oluşturulmuştur.

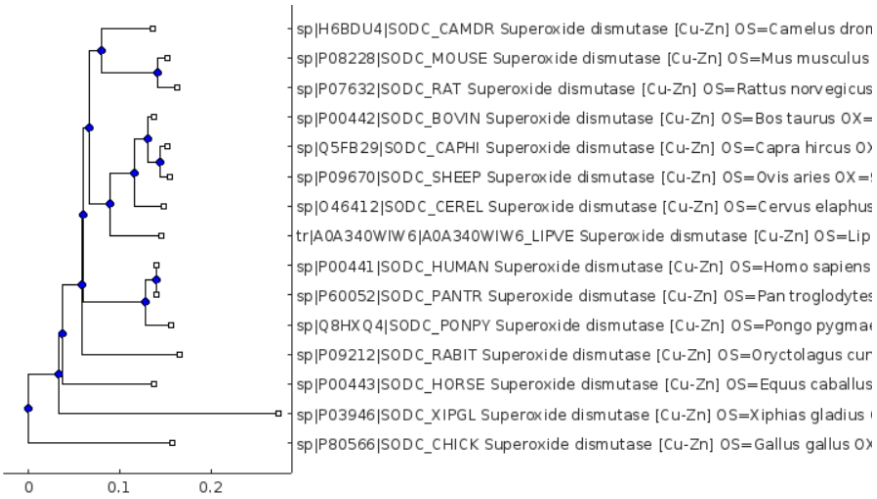
Protein dizileri arasındaki genetik uzaklığı temel alarak elde edilen JC uzaklığı ile Ağırlıklı Çift Grup Aritmetik Ortalamalar Yöntemi uygulanarak Şekil 7'deki gibi filogenetik ağaç elde edilmiştir.

**Tablo 4.** 15 türe ait uzaklık matrisi

| | P00442 | P00441 | P07632 | P08228 | P03946 | P00443 | P80566 | H6BDU4 | Q8HXQ4 | P09670 | P09212 | P60052 | O46412 | A0A340WIW6 | Q5FB29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **P00442** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P00441** | 0.0493 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P07632** | 0.2052 | 0.2342 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P08228** | 0.2509 | 0.2671 | 0.4747 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P03946** | 0.2203 | 0.2109 | 0.037 | 0.474 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P00443** | 0.2284 | 0.226 | 0.4964 | 0.124 | 0.4964 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P80566** | 0.6973 | 0.646 | 0.8344 | 0.6544 | 0.8021 | 0.5795 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **H6BDU4** | 0.2553 | 0.2535 | 0.4899 | 0.1416 | 0.4884 | 0.1195 | 0.6009 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Q8HXQ4** | 0.2431 | 0.2452 | 0.4858 | 0.1504 | 0.4844 | 0.1268 | 0.601 | 0.0112 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P09670** | 0.0772 | 0.1092 | 0.2844 | 0.2772 | 0.2936 | 0.2664 | 0.7012 | 0.2614 | 0.2534 | 0 | 0 | 0 | 0 | 0 | 0 |
| **P09212** | 0.2225 | 0.2179 | 0.0457 | 0.4828 | 0.0102 | 0.5073 | 0.8048 | 0.5013 | 0.4963 | 0.2968 | 0 | 0 | 0 | 0 | 0 |
| **P60052** | 0.2509 | 0.2671 | 0.4747 | 0 | 0.474 | 0.124 | 0.6544 | 0.1416 | 0.1504 | 0.2772 | 0.4828 | 0 | 0 | 0 | 0 |
| **O46412** | 0.6856 | 0.6871 | 0.7237 | 0.6228 | 0.7481 | 0.6339 | 0.8426 | 0.6519 | 0.6474 | 0.6724 | 0.756 | 0.6228 | 0 | 0 | 0 |
| **A0A340WIW6** | 0.2451 | 0.2406 | 0.0638 | 0.5069 | 0.0279 | 0.5313 | 0.8084 | 0.5255 | 0.5208 | 0.3199 | 0.0187 | 0.5069 | 0.7341 | 0 | 0 |
| **Q5FB29** | 0.5038 | 0.4873 | 0.5764 | 0.3182 | 0.558 | 0.3789 | 0.5069 | 0.3509 | 0.344 | 0.5324 | 0.5597 | 0.3182 | 0.7025 | 0.5796 | 0 |




**Şekil 6.** Dalgacık dönüşümü ile elde edilen filogenetik ağaç




**Şekil 7.** Jukes-Cantor uzaklık ile elde edilen filogenetik ağaç

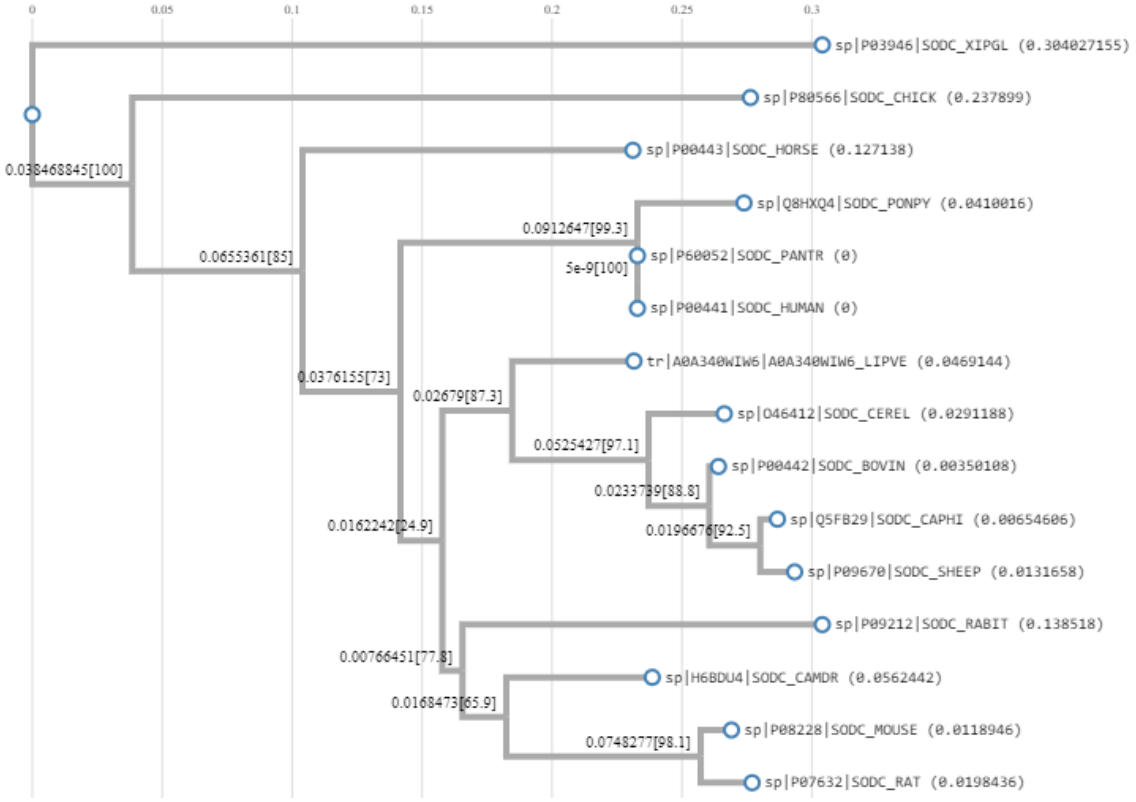
Şekil 6'te, insan ile şempanze, sıçan ile fare, koyun ile keçi arasında en kısa uzaklıklar elde edilmiştir. Bu sonuç, bu türlerin genetik açıdan oldukça yakın olduğunu ispatlamaktadır.

Şekil 7'te, insan ile şempanze, keçi ile koyun, fare ile sıçan aynı kümede yer almaktadır.

Şekil 6 ve Şekil 7'de elde edilen filogenetik ağaç sonuçlarına göre de benzer türler (sıçan ve fare, keçi ve koyun, insan ve şempanze) birbirine çok yakın dallarda bulunmuştur. Ayrıca iki ağaçta da tavuk grup dışı tür olarak bulunmaktadır. Bu sonuçlar, genetik yakınlığın, önerilen yöntem aracılığıyla açıklanabildiğini desteklemektedir.

Yeni geliştirilen bir filogenetik ağacın doğruluğunu değerlendirmek için, geçerliliği kesin olan bir filogenetik ağaç oluşturma yöntemiyle kıyaslanmaktadır [29]. Bu nedenle çalışmada, önerilen filogenetik ağaç yöntemi çoklu dizi hizalama tabanlı ClustalW yöntemine [30] göre oluşturulmuş filogenetik ağaç ile karşılaştırılmaktadır. ClustalW ile oluşturulmuş filogenetik ağaç Şekil 8'de gösterilmektedir. Buna göre, çalışmada önerilen yöntemle benzer sonuçlar alınmıştır: insan ile şempanze, koyun ile keçi, fare ile sıçan en yakın dallarda bulunmaktadır.

Ayrıca, 15 türe ait filogenetik ağaç oluşturma süreleri Dalgacık dönüşümü ile 2.0711178 sn., Jukes-Cantor ile 2.200329 sn. olarak elde edilmiştir.




**Şekil 8.** ClustalW ile oluşturulan filogenetik ağaç

## 4 SONUÇLAR

Literatürde, protein benzerlik analizleri üzerine uygulanan yöntemler arasında, spektral yarıçap-tabanlı [31], Markov-tabanlı [32], dengeli ROC [33], Fourier dönüşümleri [34] yer almaktadır. Uygulanan modellerin seçimi olduğu kadar, proteinlerin algoritmada kullanılması için uygun şekilde tanımlanması da önem taşımaktadır.

Bu çalışmada, sinyale dönüştürülen protein dizilerinin dalgacık dönüşümü ile bileşenlerine ayrılması ve daha sonra aralarındaki benzerliklere göre türlere ait protein dizilerinin benzerliklerinin filogenetik ağaçla ifade edilmesi amaçlanmıştır.

Amino asitlerin hidropati değerlerine dayalı olarak sinyal olarak tanımlanan protein dizilerinin sinyal işleme problemlerinde uygulanan dalgacık dönüşümü ile bileşenlerine yaklaşım ve detay bileşenlerine ayrıştırılarak farklı türlere ait protein dizileri arasındaki benzerlik değerleri elde edilmiştir. Daha sonra, benzerlik değerleri uzaklık değerlerine dönüştürülerek filogenetik ağaçlar oluşturulmuştur. Önerilen yöntemden elde edilen ağaç, literatürde mevcut olan uzaklık değeri, JC, kullanılarak elde edilmiş ağaç ile karşılaştırılmış ve benzer bulgular elde edilmiştir. Aynı zamanda, çalışmada önerilen yaklaşımın doğruluğunu değerlendirmek amacıyla geçerliliği olan, çoklu dizi hizalama tabanlı bir yöntem ClustalW ile filogenetik ağaç oluşturulmuş, benzer sonuçlar elde edilmiştir. Ayrıca dalgacık dönüşümü kullanarak tanımlanan filogenetik ağaç oluşturma işlem süresinin mevcut JC yöntemine göre daha kısa olmasının büyük veri analizlerinde avantaj sağlaması beklenmektedir. Çalışmadaki sonuçlar, dalgacık analizinin, protein benzerliğine dayalı filogenetik çalışmalarda doğru ve hızlı sonuçlar elde etmek amacıyla kullanılabilirliğini göstermektedir.

## Teşekkür
Bu çalışma 2019.KB.FEN.001 numaralı proje kapsamında yapılmış olup Dokuz Eylül Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi tarafından desteklenmiştir.

## Yazar Katkıları
**Çağın KANDEMİR ÇAVAŞ:** Kavramlaştırma, Metodoloji, Yazılım, Doğrulama, Veri analizi, Araştırma, Materyaller / Kaynaklar, Veri İyileştirme, Yazım - Özgün Taslak, Yazım - Değerlendirme & Düzenleme, Görselleştirme, Süpervizyon, Proje yönetimi

Yazar makalenin son halini okuyup onaylamıştır.

## Çıkar Çatışması Beyanı
Yazar herhangi bir çıkar çatışması olmadığını beyan eder.

## Kaynakça

[1] A. Lesk, "*Introduction to bioinformatics*", New York, USA: Oxford University Press, 2004.

[2] S. A. Krawetz, and D. D. Womble, "*Introduction to bioinformatics: a theoretical and practical approach*", New Jersey, USA: Humana Press, 2003.

[3] D. Baker, and A. Sali, "Protein structure prediction and structural genomics", *Science*, vol. 294, no. 5540, pp. 93-96, 2001.

[4] M. S. Rosenberg, "Evolutionary distance estimation and fidelity of pair wise sequence alignment", *BMC Bioinformatics*, vol. 6, no. 102, 2005.

[5] D. J Rigden, and D. J. Rigden, "*From protein structure to function with bioinformatics*", Heidelberg-Almanya: Springer, 2017.

[6] H. Lin, "The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition", *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 350-356, 2008.

[7] J. Jin, and J. An, "Robust discriminant analysis and its application to identify protein coding regions of rice genes", *Mathematical Biosciences*, vol. 232, no. 2, pp. 96-100, 2011.

[8] A. Pavesi, "New insights into the evolutionary features of viral overlapping genes by discriminant analysis", *Virology*, vol. 546, pp. 51-66, 2020.

[9] C. Rhodes, C. Lewis, J. Szekely, A. Campbell, M. R. A. Creighton, E. Boone, and S. Seashols-Williams, "Developmental validation of a microRNA panel using quadratic discriminant analysis for the classification of seven forensically relevant body fluids", *Forensic Science International: Genetics*, vol. 59, no. 102692, 2022.

[10] S. T. Sara, M. M. Hasan, A. Ahmad, and S. Shatabda, "Convolutional neural networks with image representation of amino acid sequences for protein function prediction", *Computational Biology and Chemistry*, vol. 92, no. 107494, 2021.

[11] G. Orlando, D. Raimondi, F. Codice, F. Tabaro, and W. Vranken, "Prediction of disordered regions in proteins with recurrent neural networks and protein Dynamics", *Journal of Molecular Biology*, vol. 434(12), no. 167579, 2022.

[12] E. Nasibov, and C. Kandemir-Cavas, "Protein subcellular location prediction using optimally weighted fuzzy k-NN algorithm", *Computational Biology and Chemistry*, vol. 32, no. 6, pp. 448-451, 2008.

[13] Y. Ding, J. Tang, and F. Guo, "Human protein subcellular localization identification via fuzzy model on kernelized neighborhood representation", *Applied Soft Computing*, vol. 96, no. 106596, 2020.

[14] Z. B. Ozger, and P. Cihan, "A novel ensemble fuzzy classification model in SARS-CoV-2 B-cell epitope identification for development of protein-based vaccine", *Applied Soft Computing*, vol. 116, no. 108280, 2022.

[15] M. L. Islam, S. Shatabda, M. A. Rashid, M. G. Khan, and M. S. Rahman, "Protein structure prediction from inaccurate and sparse NMR data using an enhanced genetic algorithm", *Computational Biology and Chemistry*, vol. 79, pp. 6-15, 2019.

[16] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier", *Artificial Intelligence in Medicine*, vol. 98, pp. 35-47, 2019.

[17] B. Bošković, and J. Brest, "Genetic algorithm with advanced mechanisms applied to the protein structure prediction in a hydrophobic-polar model and cubic lattice", *Applied Soft Computing*, vol. 45, pp. 61-70, 2016.

[18] M. R. Kumar, and N. K. Vaegae, "A new numerical approach for DNA representation using modified Gabor wavelet transform for the identification of protein coding regions", *Biocybernetics and Biomedical Engineering*, vol. 40, no. 2, pp. 836-848, 2020.

[19] Q. Zheng, T. Chen, W. Zhou, L. Xie, and H. Su, "Gene prediction by the noise-assisted MEMD and wavelet transform for identifying the protein coding regions", *Biocybernetics and Biomedical Engineering*, vol. 41, no.1, pp. 196-210, 2021.

[20] B. Yu, L. Lou, S. Li, Y. Zhang, W. Qiu, X. Wu, M. Wang, and B. Tian, "Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising", *Journal of Molecular Graphics and Modelling*, vol. 76, pp. 260-273, 2017.

[21] G. A. Arango-Argoty, J. A. Jaramillo-Garzón, and G. Castellanos-Domínguez, "Feature extraction by statistical contact potentials and wavelet transform for predicting subcellular localizations in gram negative bacterial proteins", *Journal of Theoretical Biology*, vol. 364, pp. 121-130, 2015.

[22] B. Yu, S. Li, C. Chen, J. Xu, W. Qiu, X. Wu, and R. Chen, "Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition", *Chemometrics and Intelligent Laboratory Systems*, vol. 167, pp. 102-112, 2017.

[23] S. Chaohong, and S. Feng, "Wavelet transform for predicting apoptosis proteins subcellular location", *Journal of Natural Sciences*, vol. 15, no. 2, pp. 103-108, 2010.

[24] J. J. Shu, and K. Y. Yong, "Fourier-based classification of protein secondary structures", *Biochemical and Biophysical Research Communications*, vol. 485, pp. 731-735, 2017.

[25] A. Bairoch, "The ENZYME database in 2000", *Nucleic Acids Research*, vol. 28, pp. 304–305, 2000.

[26] J. Kyte, and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein", *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–32, 1982.

[27] D. F. Walnut, "*An introduction to wavelet analysis*", Boston, USA: Springer, 2002.

[28] N. Arı, Ş. Özen, and Ö. H. Çolak, "*Dalgacık Teorisi (Wavelet), Matlab uygulamaları ile*", Ankara, Türkiye: Palme Yayıncılık, 2008.

[29] F. Pardi, and O. Gascuel, "Distance-based methods in phylogenetics". Richard M. Kliman. Encyclopedia of Evolutionary Biology, Elsevier, pp.458-465, 2016.

[30] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", *Nucleic Acids Research*, vol 11, no. 22, pp. 4673-4680, 1994.

[31] C. Wu, R. Gao, Y. De Marinis, and Y. Zhang, "A novel model for protein sequence similarity analysis based on spectral Radius", *Journal of Theoretical Biology*, vol. 446, pp. 61-70, 2018.

[32] J. Wu, T. Zhou, J. Tao, Y. Hai, F. Ye, X. Liu, and Q. Dai, "Similarity/dissimilarity analysis of protein structures based on Markov random fields", *Computational Biology and Chemistry*, vol. 75, pp. 45-53, 2018.

[33] R. Busa-Fekete, A. Kertész-Farkas, A. Kocsor, and S. Pongor, "Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities", *Journal of Biochemical and Biophysical Methods*, vol. 70, no. 6, pp. 1210-1214, 2008.

[34] J. Zhao, J. Wang, W. Hua, and P. Ouyang, "Algorithm, applications and evaluation for protein comparison by Ramanujan Fourier transform", *Molecular and Cellular Probes*, vol. 29, no. 6, pp. 396-407, 2015.