

**T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**



**PROTEİN / AMİNO ASİT DİZİMLERİNİN YAPAY ZEKA İLE
DEĞERLENDİRİLMESİNDE YENİ YAKLAŞIMLARIN
GELİŞTİRİLMESİ**

Talha Burak ALAKUŞ

Doktora Tezi

YAZILIM MÜHENDİSLİĞİ ANABİLİM DALI

Yazılım Mühendisliği

HAZİRAN 2022

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yazılım Mühendisliği Anabilim Dalı

Doktora Tezi

PROTEİN / AMİNO ASİT DİZİLİMLERİNİN YAPAY ZEKA İLE
DEĞERLENDİRİLMESİNDE YENİ YAKLAŞIMLARIN
GELİŞTİRİLMESİ

Tez Yazarı
Talha Burak ALAKUŞ

Danışman
Prof. Dr. İbrahim TÜRKOĞLU

HAZİRAN 2022
ELAZIĞ

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yazılım Mühendisliği Anabilim Dalı

Doktora Tezi

Başlığı:	Protein / Amino Asit Dizilimlerinin Yapay Zeka İle Değerlendirilmesinde Yeni Yaklaşımların Geliştirilmesi
Yazarı:	Talha Burak ALAKUŞ
İlk Teslim Tarihi:	05.05.2022
Savunma Tarihi:	06.06.2022

TEZ ONAYI

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına göre hazırlanan bu tez aşağıda imzaları bulunan jüri üyeleri tarafından değerlendirilmiş ve akademik dinleyicilere açık yapılan savunma sonucunda OYBİRLİĞİ ile kabul edilmiştir.

Danışman:	Prof. Dr. İbrahim TÜRKOĞLU Fırat Üniversitesi, Teknoloji Fakültesi	<i>İmza</i> Onayladım
Başkan:	Prof. Dr. Mehmet KAYA Fırat Üniversitesi, Mühendislik Fakültesi	Onayladım
Üye:	Prof. Dr. Bilal ALATAŞ Fırat Üniversitesi, Mühendislik Fakültesi	Onayladım
Üye:	Prof. Dr. Nejat YUMUŞAK Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi	Onayladım
Üye:	Prof. Dr. Nizamettin AYDIN Yıldız Teknik Üniversitesi, Elektrik Elektronik Fakültesi	Onayladım

Bu tez, Enstitü Yönetim Kurulunun/...../20..... tarihli toplantısında tescillenmiştir.

İmza
Prof. Dr. Kürşat Esat ALYAMAÇ
Enstitü Müdürü

BEYAN

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım “Protein / Amino Asit Dizilimlerinin Yapay Zeka İle Değerlendirilmesinde Yeni Yaklaşımların Geliştirilmesi” Başlıklı Doktora Tezimin içindeki bütün bilgilerin doğru olduğunu, bilgilerin üretilmesi ve sunulmasında bilimsel etik kurallarına uygun davrandığımı, kullandığım bütün kaynakları atıf yaparak belirttiğimi, maddi ve manevi desteği olan tüm kurum/kuruluş ve kişileri belirttiğimi, burada sunduğum veri ve bilgileri unvan almak amacıyla daha önce hiçbir şekilde kullanmadığımı beyan ederim.

06.06.2022

Talha Burak ALAKUŞ



ÖNSÖZ

Gelişen bilgi ve iletişim teknolojilerinin bir parçası olan internet teknolojileri, web sayfaları, bloklar ve sosyal medya uygulamaları pek çok veri toplayan uygulamalar sayesinde bilimsel olan ve olmayan birçok veriler toplanabilir hale gelmiştir. Verilerin değerinin anlaşılması sonucu bu verileri toplama, işleme, erişime sunma, kullanıcılara hazır hale getirme, saklama ve analiz etme gibi pek çok farklı teknikler kullanılmaktadır. Bu verilerin hız, çeşitlilik ve hacim açısından büyük artış göstermesi ile büyük veri kavramı doğmuştur. Günümüzde yapay öğrenme tabanlı çok sayıda uygulama büyük verilerin işlenmesinde etkin bir şekilde kullanılmaktadır. Bu alandaki yenilikler, gelişmeler ve başarılar, biyoenformatik alanındaki büyük verinin analiz edilmesine vesile olmuştur.

Bu tez çalışmasında protein ya da amino asit dizilimlerinin yapay zeka ile değerlendirilmesine yönelik yeni yaklaşımlar geliştirilmiş ve bu yaklaşımlar ile biyoenformatik alanına katkılar sağlanmıştır. Biyoenformatik alanında sıklıkla kullanılan protein değerlendirme çalışmalarının kapsamı; protein dizilimlerine dayalı protein etkileşimlerinin incelenmesine, protein yapılarının sınıflandırılmasına, bilinmeyen proteinlerin fonksiyonlarının tespit edilmesine, akıllı ilaç uygulamalarında kullanılan ilaç-hedef etkileşimlerinin belirlenmesine yönelik mevcut çalışmaların iyileştirilmesi ya da yeni yaklaşımların geliştirilmesi olarak tespit edilmiştir.

Çalışmalarım süresince hiçbir desteğini esirgmeden yanımda olan eşime ve aileme; fikirlerinden, bilim insanı kişiliğinden ve tecrübesinden çok şey öğrendiğim değerli danışman hocam Prof. Dr. İbrahim Türkoğlu'na; tezin önerilerinin geliştirilmesi ve belirlenmesi aşamalarında katkılarını sunan Prof. Dr. Mehmet Kaya'ya ve Prof. Dr. Bilal Alataş'a; bu tezin her aşamasında desteğini hissettiğim değerli dostlarım ve meslektaşlarıma teşekkürü borç bilirim.

Talha Burak ALAKUŞ

ELAĞ, 2022

İÇİNDEKİLER

	Sayfa
ÖNSÖZ	iv
İÇİNDEKİLER	v
ÖZET	vii
ABSTRACT	viii
ŞEKİLLER LİSTESİ	ix
TABLolar LİSTESİ	x
SİMGELER VE KISALTMALAR.....	xiii
1. GİRİŞ	1
1.1. Tezin Amacı ve Kapsamı.....	1
1.2. Yönelim Gerekçeleri.....	2
1.3. Tezin Organizasyonu	4
2. NÜKLEİK ASİTLER VE PROTEİNLER.....	6
2.1. Nükleik Asitlerin Yapısı	6
2.2. Genetik Bilginin Saklanması	7
2.3. Proteinlerin Yapı Seviyeleri.....	10
2.3.1. Birincil Yapı	10
2.3.2. İkincil Yapı	11
2.3.3. Üçüncül ve Dördüncül Yapı	12
3. PROTEİN DİZİLİMLERİNİN SAYISALLAŞTIRILMASI.....	14
3.1. Karakter Tabanlı Protein Sayısallaştırma Yöntemleri	14
3.1.1. CPNR Protein Sayısallaştırma Yöntemi	14
3.1.2. İkili Kodlama Sayısallaştırma Yöntemi.....	16
3.2. Fizikokimyasal Tabanlı Protein Sayısallaştırma Yöntemleri.....	16
3.2.1. Hidrofobiklik Protein Sayısallaştırma Yöntemi	17
3.2.2. Meiler Parametreleri Protein Sayısallaştırma Yöntemi	17
3.2.3. Atchley Faktörleri Protein Sayısallaştırma Yöntemi	18
3.2.4. EIIP Protein Sayısallaştırma Yöntemi	19
3.3. Evrim Tabanlı Protein Sayısallaştırma Yöntemleri	20
3.3.1. PAM250 Protein Sayısallaştırma Yöntemi.....	20
3.3.2. BLOSUM62 Protein Sayısallaştırma Yöntemi.....	22
3.3.3. PSSM Protein Sayısallaştırma Yöntemi	24
3.4. Yapı Tabanlı Protein Sayısallaştırma Yöntemleri	27
3.4.1. Miyazawa Enerjileri Protein Sayısallaştırma Yöntemi.....	27
3.4.2. Micheletti Potansiyelleri Protein Sayısallaştırma Yöntemi.....	28
3.5. Makine Öğrenmesi Tabanlı Protein Sayısallaştırma Yöntemleri	28
3.5.1. AESNN3 Protein Sayısallaştırma Yöntemi	29
3.5.2. ANN4D Protein Sayısallaştırma Yöntemi.....	30
3.5.3. ProtVec Protein Sayısallaştırma Yöntemi	31
4. PROTEİN DİZİLİMLERİ İLE GERÇEKLEŞTİRİLEN ÇALIŞMALAR.....	33
4.1. Protein-Protein Etkileşimlerinin Tahmin Edilmesi.....	33
4.2. Protein Ailelerinin Sınıflandırılması	38
4.3. Protein Fonksiyonlarının Belirlenmesi	41

4.4. İlaç-Hedef Etkileşimlerinin Tahmin Edilmesi	45
4.5. Kodlayan ve Kodlamayan RNA Dizilimlerinin Saptanması	49
4.6. Kanser İçeren Protein Dizilimlerinin Belirlenmesi.....	52
5. AVL AĞACI TABANLI YENİ BİR SAYISALLAŞTIRMA YÖNTEMİ.....	58
5.1. Önerilen Sayısallaştırma Yöntemi	58
5.2. COVID-19 Protein Dizilimlerine Uygulanması	61
5.2.1. Derin Öğrenme Kullanılarak Protein Etkileşimlerinin Tahmin Edilmesi	66
5.3. Uygulama Sonuçlarının Değerlendirilmesi.....	75
6. ENTROPİ TABANLI YENİ BİR SAYISALLAŞTIRMA YÖNTEMİ	79
6.1. Önerilen Sayısallaştırma Yöntemi	79
6.2. Viral Proteinler Arasındaki Etkileşimlere Uygulanması	85
6.2.1. Ebola Virüsüne Ait Protein Etkileşimlerinin Birlikte Evrim Analizi	89
6.2.2. Influenza A Virüsüne Ait Protein Etkileşimlerinin Birlikte Evrim Analizi	94
6.3. Uygulama Sonuçlarının Değerlendirilmesi.....	102
7. FİBONACCİ HASH TABANLI YENİ BİR SAYISALLAŞTIRMA YÖNTEMİ	108
7.1. Önerilen Sayısallaştırma Yöntemi	108
7.2. Protein Ailelerine Uygulanması	111
7.2.1. Derin Öğrenme Kullanılarak Protein Ailelerinin Sınıflandırılması.....	113
7.3. Uygulama Sonuçlarının Değerlendirilmesi.....	119
8. ÖNERİLEN PROTEİN SAYISALLAŞTIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI... 121	
8.1. Protein Sayısallaştırma Yöntemlerinin Viral-Konak Etkileşimlerine Uygulanması	121
8.2. Derin Öğrenme Kullanılarak Viral-Konak Etkileşimlerin Tahmin Edilmesi	126
8.3. Uygulama Sonuçlarının Değerlendirilmesi.....	129
9. SONUÇ	132
9.1. Sonuçlar ve Tartışma	134
9.2. Öneriler	135
TERİMLER SÖZLÜĞÜ	137
KAYNAKLAR	139
ÖZGEÇMİŞ	

ÖZET

Protein / Amino Asit Dizilimlerinin Yapay Zeka İle Değerlendirilmesinde Yeni Yaklaşımların Geliştirilmesi

Talha Burak ALAKUŞ

Doktora Tezi

FIRAT ÜNİVERSİTESİ

Fen Bilimleri Enstitüsü

Yazılım Mühendisliği Anabilim Dalı

Haziran 2022, Sayfa: xiii + 153

Biyoenformatik alanında protein veya amino asit üzerine gerçekleştirilen çalışmalar ve araştırmalar, proteinlerin yapılarının anlaşılabilmesi ve proteinlerin hücresel aktivitelerdeki rollerinin belirlenmesi açısından büyük bir önem arz etmektedir. Yapılan araştırmalar incelendiğinde, protein fonksiyonlarının belirlenmesi, proteinler arasındaki etkileşimlerin tahmin edilmesi, protein ikincil yapılarının bulunması, protein ailelerinin sınıflandırılması, kodlayan ve kodlamayan mRNA'ların belirlenmesi, viral-konak etkileşimlerinin tahmin edilmesi ve ilaç-hedef etkileşimlerinin tespit edilmesi ile ilgili çalışmaların aktif bir şekilde yürütüldüğü gözlemlenmiştir.

Tez çalışmasında, söz konusu araştırma alanlarına yönelik yapay zeka tabanlı yeni analiz yaklaşımları geliştirilmiştir. Bu çerçevede, protein dizilimlerini sayısallaştırmak için algoritmik-tabanlı yeni bir kategorisel yaklaşım geliştirilmiştir. Algoritmik-tabanlı yaklaşım kapsamında, üç farklı protein sayısallaştırma yöntemi önerilmiştir. Bu doğrultuda, önerilen protein dizilimlerini sayısallaştırma yöntemleri:

- Protein dizilimlerini sayısallaştırmak için AVL ağacı tabanlı protein sayısallaştırma yöntemi önerilmiş ve bu yöntemin başarımı COVID-19 hastalığına neden olan SARS-CoV-2 virüsünün proteinleri arasındaki etkileşimleri üzerinde test edilmiştir. Bu yöntem, içeriğinde bulunan ikili arama ağaç yapısı nedeniyle algoritmik-tabanlı kategoride değerlendirilmiştir.
- Entropi hesaplamasına dayanan protein sayısallaştırma yöntemi önerilmiş ve viral genomlar üzerindeki protein-protein etkileşimlerini belirlemek için kullanılmıştır. Bu yöntemde ise Shannon entropisi kullanılması nedeniyle algoritmik-tabanlı kategoride değerlendirilmiştir.
- Hem Fibonacci sayılarını hem de hash tablosunu içeren (FIBHASH) protein sayısallaştırma yöntemi önerilmiş ve bu yöntemin başarımı, protein ailelerini belirlemek için değerlendirilmiştir. Bu yöntem, Fibonacci sayılarını ve hash tablosunu içeren hibrit bir yöntemdir. Hash tablosu veri yapıları ve algoritma analizinde sıklıkla kullanıldığı için, FIBHASH yöntemi algoritmik-tabanlı yöntem olarak değerlendirilmiştir.
- Önerilen protein sayısallaştırma yöntemleri SARS-CoV-2 virüsü ve konak hücreleri arasındaki etkileşimleri tahmin etmek için kullanılmış ve hem önerilen protein sayısallaştırma yöntemlerinin hem de literatürde en çok kullanılan yöntemlerin karşılaştırılması yapılmıştır.

Önerilen protein sayısallaştırma yöntemleriyle elde edilen yapay zekâ değerlendirme başarımları, protein analizi alanlarında yapılacak olan yeni çalışmalar için ümit verici olmuştur.

Anahtar Kelimeler: Biyoenformatik, Yapay zeka, Derin öğrenme, Protein analizi, Protein sayısallaştırma yöntemleri, Algoritmik-tabanlı yöntemler, AVL-ağacı, Entropi

ABSTRACT

Development of New Approaches to Evaluate Protein/Amino Acid Sequences with Artificial Intelligence

Talha Burak ALAKUŞ

Ph.D. Thesis

FIRAT UNIVERSITY
Graduate School of Natural and Applied Sciences
Department of Software Engineering

June 2022, Pages: xiii + 153

Studies and research on protein or amino acids in the field of bioinformatics are of great importance in terms of understanding the structures of proteins and determining the roles of proteins in cellular activities. When the researches are examined, it is observed that studies related to determining protein functions, predicting interactions between proteins, finding protein secondary structures, classifying protein families, identifying coding and non-coding mRNAs, predicting viral-host interactions and detecting drug-target interactions have been actively conducted.

In the thesis, new artificial intelligence-based analysis approaches have been developed for these research areas. In this framework, a new algorithmic-based categorical approach has been developed to map protein sequences. As part of the algorithmic-based approach, three different protein mapping methods have been proposed. Accordingly, the proposed methods of quantifying protein sequences are:

- An AVL-based protein mapping method has been proposed to convert protein sequences, and the success of this method has been tested on interactions between proteins of the SARS-CoV-2 virus that causes COVID-19 disease. This method has been evaluated in the algorithmic-based category due to the binary search tree structure it contains.
- In another study, entropy-based protein mapping method has been proposed and used to identify protein-protein interactions on viral genomes. Since Shannon entropy is used in this method, it is evaluated in the algorithmic-based category.
- Another protein mapping method (FIBHASH), which includes both Fibonacci numbers and hash tables, has been developed and the performance of this method has been evaluated in determining protein families. This method is a hybrid method that includes Fibonacci numbers and hash table. Since the hash table is frequently used in data structures and algorithm analysis, the FIBHASH method is considered an algorithmic-based method.
- The proposed protein mapping methods were used to predict the interactions between SARS-CoV-2 virus and host cells, and both the proposed protein mapping methods and the state of art methods in the literature were compared.

The successes of artificial intelligence obtained with the proposed protein mapping methods have been promising for new studies in the fields of protein analysis.

Keywords: Bioinformatics, Artificial intelligence, Deep learning, Protein analysis, Protein mapping methods, Algorithmic-based methods, AVL-tree, Entropy

ŞEKİLLER LİSTESİ

	Sayfa
Şekil 2.1. DNA'nın çift sarmal yapısı [1].....	7
Şekil 2.2. Santral dogma süreci	9
Şekil 2.3. α -heliks yapısı [4].....	11
Şekil 2.4. Paralel ve paralel olmayan β -yaprak yapılarının üstten görünümü [5]	12
Şekil 2.5. Protein üçüncül yapı örneği [6].....	13
Şekil 2.6. Protein dördüncül yapı örneği [7]	13
Şekil 3.1. PAM 250 skor matrisi [15].....	21
Şekil 3.2. BLOSUM62 matrisi	23
Şekil 4.1. GO terimlerinin graf yapısı	42
Şekil 5.1. Amino asit kodlarının AVL ağacına eklenmesi	59
Şekil 5.2. Her bir amino asidin derinlik değerleri	60
Şekil 5.3. SARS-CoV-2 virüsünün genomik yapısı	63
Şekil 5.4. Önerilen AVL yönteminin yapay zekayla değerlendirilme süreci.....	65
Şekil 5.5. Geliştirilen BiRNN modelinin yapısı.....	67
Şekil 5.6. 10 katlı çapraz-doğrulama işleminin adımları.....	67
Şekil 5.7. Orf3a proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri	69
Şekil 5.8. Orf3b proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri	70
Şekil 5.9. Orf6 proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri	71
Şekil 5.10. Orf7a proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri	73
Şekil 5.11. Orf8 proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri	74
Şekil 6.1. Birlikte evrim analizinin genel şeması	86
Şekil 6.2. Önerilen entropi yönteminin birlikte evrim analiziyle değerlendirme süreci	89
Şekil 6.3. ÇBÖ analizi. a) Önerilen protein sayısallaştırma yöntemi. b) MirrorTree	103
Şekil 6.4. ÇBÖ analizi. a) Önerilen protein sayısallaştırma yöntemi. b) MirrorTree	105
Şekil 7.1. Önerilen FIBHASH yönteminin derin öğrenme modelleriyle değerlendirilme süreci	113
Şekil 7.2. En etkili protein sayısallaştırma yöntemlerinin ROC eğrileri.....	119
Şekil 8.1. Protein sayısallaştırma yöntemlerinin başarımlarının değerlendirme süreci	125
Şekil 8.2. En etkili protein sayısallaştırma yöntemlerinin AUC grafikleri	127

TABLÖLAR LİSTESİ

	Sayfa
Tablo 2.1. RNA kodon tablosu.....	8
Tablo 2.2. Amino asitler ve kodları.....	10
Tablo 3.1. Amino asitler ve kodon sayıları	14
Tablo 3.2. Asal sayılar arasındaki fark ve karşılık gelen sayısal ifadeler	15
Tablo 3.3. Amino asitlerin CPNR değerleri	15
Tablo 3.4. Amino asitlerin ikili-kodlama değerleri	16
Tablo 3.5. Amino asitlerin hidrofobiklik değerleri.....	17
Tablo 3.6. Amino asitlerin Meiler parametreleri.....	18
Tablo 3.7. Amino asitlerin Atchley faktörleri	19
Tablo 3.8. Amino asitlerin EIIP değerleri	20
Tablo 3.9. Amino asitlerin PAM250 değerleri	22
Tablo 3.10. Amino asitlerin BLOSUM62 değerleri	23
Tablo 3.11. ABRE motifine ait bazların pozisyon değerleri	24
Tablo 3.12. Bazların her bir pozisyon için frekans değerleri	25
Tablo 3.13. Bazların logit değerleri.....	25
Tablo 3.14. Bazların tam sayı haline getirilmiş logit değerleri	26
Tablo 3.15. Bazların eşit frekans değerleri.....	26
Tablo 3.16. Bazların PSSM değerleri.....	27
Tablo 3.17. Amino asitlerin Miyazawa enerjileri.....	27
Tablo 3.18. Amino asitlerin Micheletti potansiyelleri.....	28
Tablo 3.19. Amino asitlerin AESNN3 değeri.....	29
Tablo 3.20. Amino asitlerin ANN4D değeri	30
Tablo 3.21. ProtVec yönteminin çalışma prensibi	31
Tablo 5.1. Amino asitlerin AVL değerleri	61
Tablo 5.2. Etkileşim halinde olan proteinler ve toplam sayıları.....	64
Tablo 5.3. Orf3a proteininin etkileşim sonuçları.....	68
Tablo 5.4. Orf3b proteininin etkileşim sonuçları	69
Tablo 5.5. Orf6 proteininin etkileşim sonuçları	70
Tablo 5.6. Orf7a proteininin etkileşim sonuçları.....	72
Tablo 5.7. Orf8 proteininin etkileşim sonuçları	73
Tablo 5.8. Protein sayısallaştırma yöntemlerinin ortalama etkileşim doğruluk skorları.....	75

Tablo 6.1.	Önerilen protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları.....	89
Tablo 6.2.	PAM250 protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları.....	90
Tablo 6.3.	BLOSUM62 protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları	91
Tablo 6.4.	Miyazawa enerjileri protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları	91
Tablo 6.5.	Micheletti potansiyelleri protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları	92
Tablo 6.6.	Atchley faktörleri protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları	93
Tablo 6.7.	Hidrofobiklik protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları	93
Tablo 6.8.	Önerilen Shannon entropisine dayalı yöntem ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları	94
Tablo 6.9.	PAM250 protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları	95
Tablo 6.10.	BLOSUM62 protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları	96
Tablo 6.11.	Miyazawa enerjileri protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları	97
Tablo 6.12.	Micheletti potansiyelleri protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları	98
Tablo 6.13.	Atchley faktörleri protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları	100
Tablo 6.14.	Hidrofobiklik protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları	101
Tablo 6.15.	MirrorTree yaklaşımı ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları	102
Tablo 6.16.	MirrorTree yaklaşımı ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları.....	104
Tablo 7.1.	Amino asitlerin Fibonacci değerleri.....	109
Tablo 7.2.	Hash yöntemi örneği	110
Tablo 7.3.	Amino asitlerin FIBHASH değerleri.....	110
Tablo 7.4.	Çalışmada kullanılan protein aileleri.....	112
Tablo 7.5.	RNN sınıflandırıcısının sonuçları.....	114
Tablo 7.6.	LSTM sınıflandırıcısının sonuçları	115

Tablo 7.7.	BiRNN sınıflandırıcısının sonuçları.....	117
Tablo 7.8.	BiLSTM sınıflandırıcısının sonuçları.....	118
Tablo 8.1.	Protein sayısallaştırma yöntemlerinin başarımlarının karşılaştırılması	126
Tablo 8.2.	Protein sayısallaştırma yöntemlerinin donanım performansları.....	128



SİMGELER VE KISALTMALAR

Kısaltmalar

AAB	: Amino Asit Bileşimi
AFD	: Ayrık Fourier Dönüşümü
ARDS	: Acute Respiratory Distress Syndrome - Akut Solunum Sıkıntısı Sendromu
AUC	: Area Under Curve - Eğri Altında Kalan Alan
AVL	: Adelson, Velski and Landi
BiLSTM	: Bidirectional LSTM - Çift Yönlü LSTM
BiRNN	: Bidirectional RNN - Çift Yönlü RNN
BLOSUM	: Blocks Substitution Matrix - Blok İkame Matrisi
BS	: Biyolojik Süreç
BT	: Birleşik Triad
COVID-19	: Koronavirüs Hastalığı 2019
CPNR	: Complex Prime Number Representation - Karmaşık Asal Sayı İfadesi
ÇBÖ	: Çok Boyutlu Ölçekleme
ÇDH	: Çoklu Dizilim Hizalama
DİA	: Derin İnanç Ağları
DNA	: Deoksiribonükleik Asit
DSA	: Derin Sinir Ağları
DVM	: Destek Vektör Makineleri
EIIP	: Electron-Ion Interaction Potential - Elektron-İyon Etkileşim Potansiyeli
ESA	: Evrimsel Sinir Ağı
FIBHASH	: FibonacciHash
GAAB	: Gruplandırılmış Amino Asit Bileşimi
GAKA	: Gradyan Arttırma Karar Ağacı
GEA	: Graf Evrimsel Ağ
GO	: Gen Ontolojisi
GP	: Glikoprotein
GRU	: Gated Recurrent Unit - Geçitli Tekrarlayan Birim
HA	: Hemaglutinin
HB	: Hücresel Bileşen
KA	: Karar Ağacı

KEO	: Korelasyonlu Evrimsel Oran
KMA	: Korelasyonlu Mutasyon Analizi
KNN	: K En Yakın Komşu
L	: RNA Polimeraz
LA	: Legendre Anları
LASSO	: Least Absolute Shrinkage and Selection Operator - En Küçük Mutlak Küçültme ve Seçim Operatörü
LSTM	: Long/Short Term Memory - Uzun/Kısa Vadeli Bellek
M1	: Matris protein 1
M2	: Matris protein 2
MF	: Moleküler Fonksiyon
MKK	: Matthew's Korelasyon Katsayısı
mRNA	: Haberci RNA
NA	: Nöraminidaz
NEP	: Nükleer İhraç Protein
NMR	: Nükleer Manyetik Rezonans
NP	: Nükleoprotein
NS1	: Non Structural protein - Yapısal Olmayan protein
OK	: Oto Kovaryans
OS	: Oylama Sınıflandırıcı
ÖVK	: Öğrenmeli Vektör Kuantalama
PA	: Polimeraz Asidik
PAM	: Point Accepted Mutation - Nokta Kabul Edilen Mutasyon
PB1	: Polimeraz Bazık 1
PB2	: Polimeraz Bazık 2
PPE	: Protein-Protein Etkileşimi
PSSM	: Position Specific Scoring Matrix - Konuma Özgü Puanlama Matrisi
ReLU	: Rectified Linear Unit - Düzleştirilmiş Doğrusal Birim
RNA	: Ribonükleik Asit
RNN	: Recurrent Neural Network - Yinelemeli Sinir Ağı
RO	: Rassal Orman
RotOR	: Rotasyon Ormanı
rRNA	: Ribozomal RNA
RY	: Rastgele Yürüyüş

SA	: Sinir Ağı
SAAB	: Sözde AAB
SGD	: Stochastic Gradient Descent - Stokastik Gradyan İniş
SNB	: Saf Naive Bayes
SOM	: Self Organizing Map - Kendi Kendini Düzenleyen Harita
SSA	: Siyam Sinir Ağı
t-DSKG	: t-Dağıtılmış Stokastik Komşu Gömme
TBA	: Temel Bileşen Analizi
tRNA	: Transfer RNA
UHCK	: Ufak Hücre Ciğer Kanseri
UOHCK	: Ufak Olmayan Hücre Ciğer Kanseri
Y2H	: Yeast 2 Hybrid - Maya İki Hibrit
YDS	: Yarı Dizilim Sıralaması
YKE	: Yüksek Kalite Endeksi
YSA	: Yapay Sinir Ağı

1. GİRİŞ

Tezin bu bölümünde, tez konusunun amacı ve kapsamı verilerek, yönelim gerekçelerinden bahsedilmiştir. Bunlara ek olarak, tezin organizasyonu da bu kısımda irdelenmiştir.

1.1. Tezin Amacı ve Kapsamı

Biyoenformatik, bilgisayar ve yazılım mühendisliği, istatistik, uygulamalı matematik alanlarındaki yaklaşımların ve yöntemlerin biyolojik veriler üzerinde kullanılması ve analizi olarak ifade edilmektedir. Teknolojinin ilerlemesi ve moleküler biyoloji alanındaki yeni gelişmeler, araştırmacıların bu alana yönelmesine neden olmuştur. Bu durum, sadece biyoenformatik alanında değil aynı zamanda bilgisayar bilimleri alanında da etkisini göstermiş ve çok sayıda yapay zekaya dayalı yeni analiz yöntemleri ve yazılım algoritmaları geliştirilmiştir. Biyoenformatik alanında sıklıkla kullanılan protein ya da amino asit analizi çalışmalarının kapsamı şu şekildedir;

- Protein dizilimlerine dayalı protein etkileşimlerinin incelenmesi,
- Protein yapılarının belirlenmesi,
- Bilinmeyen protein fonksiyonlarının tespit edilmesi,
- Akıllı ilaç uygulamalarında kullanılan protein motiflerinin belirlenmesi,
- Amino asitlerin neden olduğu hastalıkların tahmin edilmesi,
- İlaç-hedef etkileşimlerinin belirlenmesi,
- Viral-konak arasındaki protein etkileşimlerinin tahmin edilmesi.

Tez çalışmasının amacı ise, bahsi geçen protein çalışmalarının kapsamı doğrultusunda, aşağıda belirtilen hedeflerin gerçekleştirilmesine yöneliktir;

- Protein dizilimlerinin sayısallaştırılmasında kullanılan mevcut sayısallaştırma tekniklerinin iyileştirilmesi ya da yeni bir protein sayısallaştırma tekniğinin geliştirilmesi,
- Protein analizi çalışmalarına yönelik genomik işaret işlemeye dayalı yöntemlerin iyileştirilmesi,
- Akıllı ilaç, protein tanımlamaları, protein yapı tahmini, hastalık, vb. alanlarda kullanılan protein-protein etkileşimlerinin belirlenmesine yönelik mevcut yöntemlerin başarımlarının artırılması için işaret işleme ve yapay zeka tekniklerini içeren yeni bir hibrit yöntemin geliştirilmesi,
- Protein hizalama, protein yapı tahmini, etkileşim belirleme, vb. çalışmalarda kullanılan mevcut modellerin iyileştirilmesi ya da yeni yöntemlerin geliştirilmesi,
- Protein dizilimlerinin hastalık tanımlama ve keşfi, tür tespiti açısından değerlendirilmesi,

1.2. Yönelim Gerekçeleri

Tezin yönelim gerekçeleri şu şekilde belirlenmiştir;

- Proteinler arasındaki etkileşimler belirlenirken ve protein analizine yönelik çalışmalar gerçekleştirilirken, genellikle çoklu dizilim hizalama işlemleri yapılmaktadır. Ancak, bu dizilim hizalama algoritmaları ile gerçekleştirilen çalışmalar sonunda, yüksek oranda yanlış pozitif etkileşimler elde edilmektedir. Bu nedenden dolayı bu tür çalışmalar için hesaplama-tabanlı yaklaşımlar kullanılmakta ve önerilmektedir.
- Protein çalışmalarında genomik işaret işlemeye dayalı yöntemler de etkili bir şekilde kullanılmaktadır. Ancak, bu yöntemlerin başarımları seçilen proteinin zincir uzunluğuna göre değişiklik göstermektedir. Çalışmalarda bu problemin önüne geçmek için moment vektörleri etkin bir şekilde kullanılmakta ancak bu yöntemlerin başarımları da seçilen protein sayısallaştırma tekniğine göre değişiklik göstermektedir.
- Çoklu hizalama hesaplaması hem işlem hem de zaman açısından maliyetli ve karmaşık olmaktadır.
- Bazı protein analizi çalışmalarında, protein sayısallaştırılma işlemi gerçekleştirilirken BLOSUM (Blocks Substitution Matrix – Blok İkame Matrisi) tablosu kullanılmaktadır ancak, bu tablo kümeleme işlemi sırasında yanlış matris hesaplamaları yapabilmektedir. Bu yanlışlıklar farklı yöntemler ile (optimal öz vektörlerin elde edilmesi, kaotik yöntemler, vb.) düzeltilebilse de performans istenilen seviyeye gelememektedir.
- Çoğu hastalıklar günümüzde protein dizilimlerinin analizi ile belirlenebilmektedir. Bu amaç için kullanılan yöntemler genellikle tıbbi yöntemler olmakla beraber laboratuvar ortamı gerektiren işlemlerden oluşmaktadır. Bundan dolayı hastalık teşhis analizi uzun zaman alabilmektedir.
- Proteinler arasındaki etkileşimleri tahmin etmek için AVL ağacına dayalı yeni bir protein sayısallaştırma yöntemi önerilmiştir. Bu yöntem ile amino asit kodları ayrı ayrı sayısal ifadelere dönüştürülmüştür. Önerilen yöntemde ilk olarak amino asit kodları alfabetik olarak sıralanmış ve ardından AVL ağacına eklenmiştir. Ekleme işleminin ardından her bir düğümün derinlik değeri hesaplanmış ve amino asit kodları sayısal hale getirilmiştir. Önerilen bu yöntemin literatüre en büyük katkısı, yeni bir kategorisel yöntem olmasıdır. Literatürde çok sayıda protein sayısallaştırma yöntemi bulunmaktadır. Bu yöntemler, karakter-tabanlı, fizikokimyasal-tabanlı, yapı-tabanlı, evrim-tabanlı ve makine öğrenmesi-tabanlı yöntemlerdir. Önerilen

yöntem ise yapısı gereği algoritmik-tabanlı bir yöntemdir. Literatür incelendiğinde bu kategoriye ait başka bir protein sayısallaştırma yöntemi bulunmamaktadır. Gerçekleştirilen çalışma ile literatüre hem yeni bir kategori hem de yeni bir protein sayısallaştırma yöntemi kazandırılmıştır.

- Viral genomlar arasındaki protein etkileşimlerini saptamak için ise entropi hesaplamasına dayalı yeni bir protein sayısallaştırma yöntemi önerilmiştir. Önerilen yöntem ile ilk olarak protein dizilimlerinde bulunan amino asit kodlarının tekrarlanma frekansları hesaplanmış ve ardından bu değerler kullanılarak Shannon entropi ile protein dizilimleri sayısal ifadelerle dönüştürülmüştür. Önerilmiş olan bu yöntemin en büyük kazanımı adaptif bir yapıya sahip olmasıdır. Bu sayede protein dizilimleri her seferinde aynı değeri almamaktadır. Bunun en büyük nedeni amino asitlerin tekrarlanma frekanslarının dizilim uzunluğuna göre değişiklik göstermesidir. Bu da yöntemin diğer protein sayısallaştırma yöntemlerinden en büyük farkını ortaya koymaktadır. Literatürdeki diğer yöntemlerde, protein dizilimlerinin uzunluğu ya da amino asit kodlarının tekrarlanma frekansı sayısallaştırma sürecini etkilememektedir. Bundan dolayı farklı amino asit kodları aynı değeri alabilmektedir. Önerilen yöntemin literatüre diğer bir katkısı ise, değerlendirildiği kategoridir. Literatürde bulunan diğer protein sayısallaştırma yöntemleri karakter-tabanlı, yapı-tabanlı, evrim-tabanlı, fizikokimyasal-tabanlı, vb. kategorilerde değerlendirilmektedir. Önerilen bu yöntem ise algoritmik-tabanlı bir yöntem olup literatürde ilk kez proteinlerin sayısal ifadelerle dönüştürülmesine uyulanmıştır.
- Protein ailelerini sınıflandırmak için ise Fibonacci sayılarını ve hash tablosunu içeren hibrit bir yapı önerilmiştir. Önerilen yöntemde ilk olarak amino asit kodları alfabetik olarak sıralanmış ve sonra her bir amino aside Fibonacci sayıları sırasıyla eklenmiştir. Ekleme işleminin ardından amino asit kodlarının sayısal ifadeleri hash tablosuna yerleştirilmiştir. Hash tablosuna yerleştirmedeki ana amaç, Fibonacci sayıları ile elde edilen değerlerin boyutlarının küçültülmesidir. Sayısal ifadelerin boyutu küçültüldükten sonra sayısallaştırma işlemi tamamlanmıştır. Önerilen yöntem algoritmik-tabanlı bir yöntemdir. Kategorik açıdan incelendiğinde literatürde bu alanda bu kategoriye ait başka bir protein sayısallaştırma yöntemine rastlanmamıştır. Bu bağlamda, FIBHASH yöntemi literatüre kazandırılmıştır. Bunun yanı sıra, literatürdeki diğer yöntemlerin çoğunda, farklı amino asit dizilimleri benzer sayısal değerler alabilmektedir. Ancak önerilen FIBHASH yönteminde böyle bir durum söz konusu değildir. Bütün amino asit kodları farklı

değerler almaktadır. Buda önerilen yöntemin diğer çoğu yöntemden farkını ortaya koymaktadır.

Tezin amaçları yukarıda maddeler halinde verilen bu problemler doğrultusunda belirlenmiştir. Bu amaçlara yönelik, tez çalışması boyunca yeni protein sayısallaştırma yöntemleri önerilmiş ve önerilen protein sayısallaştırma yöntemleri ile bahsi geçen problemlere çözüm aranmıştır. Geliştirilen protein sayısallaştırma yöntemleri ile proteinlerin birincil yapıları kullanılarak analiz işlemler gerçekleştirilmiş ve hizalama algoritması kullanılmadan protein-protein etkileşimleri tahmin edilmiştir. Buna ek olarak, protein aileleri saptanmış ve viral genomların kodlamış oldukları proteinler arasındaki etkileşimlerde belirlenmiştir.

1.3. Tezin Organizasyonu

Tezin organizasyonu şu şekildedir; **ikinci bölümde**, nükleik asitlerin yapısı, baz kavramları, proteinlerin yapıları, amino asit kodları ve dizilimleri incelenmiştir. **Üçüncü bölümde** amino asit dizilimlerinin sayısallaştırılması için kullanılan yöntemler üzerinde durulmuştur. Bu amaç doğrultusunda karakter-tabanlı, fizikokimyasal-tabanlı, evrimsel-tabanlı, yapı-tabanlı ve makine-öğrenmesi tabanlı olmak üzere, beş farklı kategoriye ait sayısallaştırma yöntemleri irdelenmiştir. **Dördüncü bölümde** ise, amino asit dizilimleri ile gerçekleştirilen çalışmalardan bahsedilmiştir. Bu kısımda proteinler arasındaki etkileşimlerin tahmin edilmesi, protein ailelerinin sınıflandırılması, protein fonksiyonlarının belirlenmesi, ilaç-hedef etkileşimlerinin tahmin edilmesi, kodlayan ve kodlamayan RNA (Ribonükleik Asit) dizilimlerinin tahmin edilmesi ve kanser içeren protein dizilimlerinin belirlenmesi üzerine yapılan çalışmaların literatür taraması gerçekleştirilmiştir. **Beşinci bölümde**, SARS-CoV-2 virüsüne ait proteinler arasındaki etkileşimi belirlemek için önerilmiş olan AVL ağacına dayalı protein sayısallaştırma yöntemi irdelenmiştir ve açıklanmıştır. Bu kısımda gerçekleştirilen çalışmanın deneysel sonuçları yorumlanmış ve literatüre katkısı üzerinde durulmuştur. **Altıncı bölümde**, önerilen başka bir protein sayısallaştırma yöntemi olan Shannon entropisine dayalı protein sayısallaştırma yöntemi üzerinde durulmuştur. Önerilen bu yöntemin performansı viral hücreler arasındaki protein etkileşimlerini belirlemek üzere test edilmiştir. Bu yöntemle ait sonuçlar bu kısımda analiz edilmiş ve literatüre katkısı sunulmuştur. **Yedinci bölümde** ise, protein ailelerini belirlemek için önerilmiş olan FIBHASH (FibonacciHash) yöntemi açıklanmıştır. Bu yöntemle ait sınıflandırma performansı bu kısımda irdelenmiştir. Yöntemin kullanım alanları ve literatüre etkisi de tartışılmıştır. Sekizinci kısımda çalışma boyunca önerilmiş olan protein yöntemlerin karşılaştırılması yapılmıştır. Bu amaç için, COVID-19'a neden olan SARS-CoV-2 virüsüne ve insana ait proteinler arasındaki viral-konak etkileşimleri tahmin edilmiştir. Hem önerilen yöntemler kendi içerisinde hem de literatürde en çok kullanılan yöntemler ile kıyaslanmıştır. Son bölüm olan **sonuç bölümünde** ise tez sonuçlandırılmış ve sonuçlar irdelenmiştir. Bunlara ek olarak, geliştirilmiş olan yöntemlerin başarımların değerleri verilmiş ve diğer

alıřmalarla kıyaslanması yapılmıřtır. Geliřtirilmiř olan yntemlerin gelecekteki kullanım alanlarına da deęinilmiřtir.

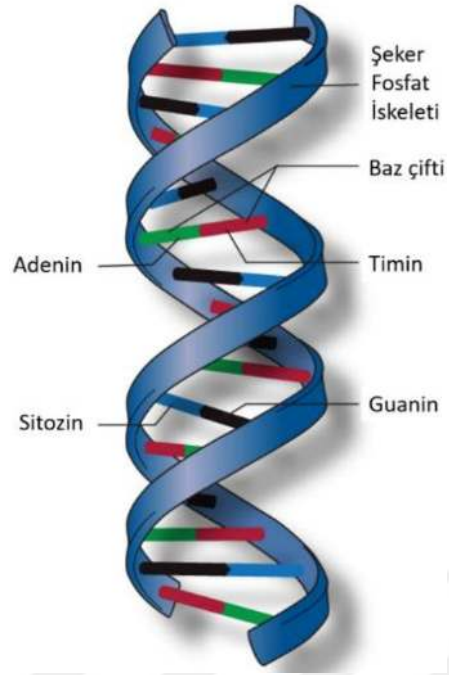


2. NÜKLEİK ASİTLER VE PROTEİNLER

Tez çalışmasının bu bölümünde; nükleik asitlerin ve proteinlerin yapısı, amino asit dizilimleri, amino asit kodları hakkında bilgiler verilmiş ve terminolojiye yönelik belirli bir altyapı oluşturulmuştur.

2.1. Nükleik Asitlerin Yapısı

Nükleik asitler ve proteinler, doğada çok önemli roller üstlenen makro moleküllerin başında gelmektedir. DNA (Deoksiribonükleik Asit) genetik bilgiyi taşımakla görevliken, RNA organizmanın hücresel aktivitelerini kontrol eden proteinlerin sentezlenmesinden sorumludur. Nükleik asitlerin temel bileşenleri nükleotitlerken, proteinlerinki ise amino asitlerdir. Nükleotitlerin yapısı DNA ve RNA'da aynıdır. Nükleotitler pentoz, fosforik asit ve heterosiklik bazdan meydana gelmektedir. Nükleotitler, bir nükleotitin pentoz şekeri ile bir sonraki nükleotitin fosforik asit kalıntısı arasındaki kimyasal bağlarla birbirlerine bağlanır. Nükleik asitler beş farklı bazdan oluşmaktadır. Bu bazlar; Adenin, Guanin, Sitozin, Timin ve Urasil bazlarıdır. Bazlar genelde ilk harfleri ile ifade edilmektedirler. Bu durumda Adenin bazı A harfi ile, Guanin bazı G harfi ile, Sitozin bazı C harfi ile, Timin bazı T harfi ile ve Urasil bazı ise U harfi ile temsil edilmektedir. Ancak bu beş adet bazın hepsi aynı anda DNA ya da RNA'da bulunmamaktadır. Urasil bazı sadece RNA dizilimlerinde görülürken, DNA dizilimlerinde Urasil bazı bulunmaz. Bunun yanı sıra, Timin bazı sadece DNA dizilimlerine özgüken, RNA dizilimlerinde bu baza rastlanmamaktadır. DNA ve RNA dizilimlerinde bulunan bazlar, bir iplikteki pürin halkası (A ve G), karşı iplikteki pirimidin (C, T ve U) halkası ile etkileşimde bulunacak biçimde eşleştirilmektedir. A ve G bazları arasında iki hidrojen bağı bulunurken, C, T ve U bazları arasında ise üç hidrojen bağı bulunmaktadır. DNA dizilimlerinde A bazının karşısına T bazı, G bazının karşısına ise C bazı gelmektedir. RNA dizilimlerinde ise A bazı ile U bazı eşleşirken, C bazı ile G bazı eşleşmektedir. Şekil 2.1'de DNA'nın çift sarmal yapısı verilmiştir.



Şekil 2.1. DNA'nın çift sarmal yapısı [1]

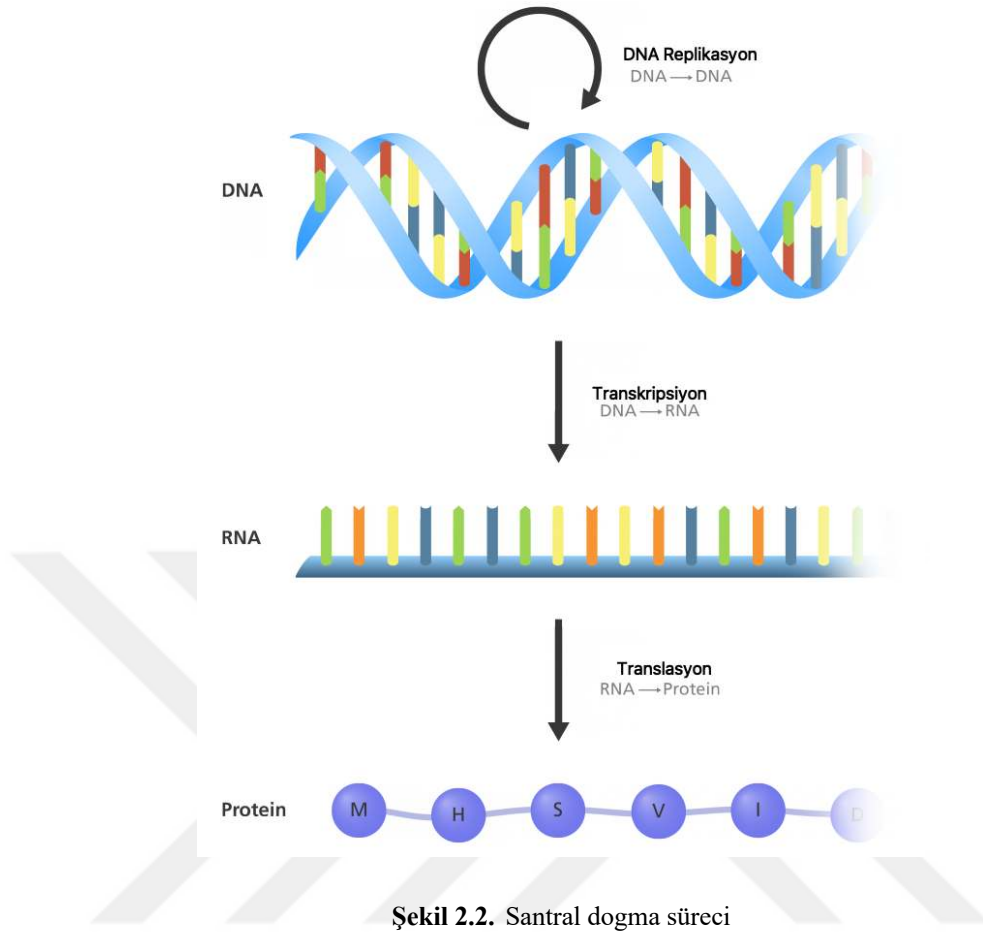
2.2. Genetik Bilginin Saklanması

Nükleotit zincirinde bulunan tek değişken eleman baz dizisidir. Bundan dolayı proteinleri oluşturmak için gerekli olan bilgileri kodlamaktadırlar. Proteinler yirmi adet amino asitten meydana gelmektedir. Her bir amino asit, kodon adı verilen bir üçlü baz tarafından kodlanmaktadır. Bir amino asit birden fazla kodon tarafından kodlanabilir. Tablo 2.1'de RNA kodon tablosu verilmiştir.

Tablo 2.1. RNA kodon tablosu

		İkinci Baz			
		U	C	A	G
Birinci Baz	U	Fenilalanin	Serin	Tirozin	Sistein
		Fenilalanin	Serin	Tirozin	Sistein
		Lösın	Serin	Dur	Dur
		Lösın	Serin	Dur	Triptofan
	C	Lösın	Prolin	Histidin	Arjinin
		Lösın	Prolin	Histidin	Arjinin
		Lösın	Prolin	Glutamin	Arjinin
		Lösın	Prolin	Glutamin	Arjinin
	A	İzolösın	Treonin	Asparajin	Serin
		İzolösın	Treonin	Asparajin	Serin
		İzolösın	Treonin	Lizin	Arjinin
		Metiyonin	Treonin	Lizin	Arjinin
	G	Valin	Alanin	Aspartik asit	Glisin
		Valin	Alanin	Aspartik asit	Glisin
		Valin	Alanin	Glutamik asit	Glisin
		Valin	Alanin	Glutamik asit	Glisin

Bir organizmadaki DNA diziliminin tamamı genom olarak ifade edilmektedir. Burada bulunan genetik bilgi ise, DNA dizilimlerinde bulunan bazlar tarafından kodlanmaktadır. Bu bilgi, transkripsiyon işlemi süresince mRNA'ya aktarılmaktadır. mRNA'dan protein oluşturma sürecine translasyon denilmektedir. Genom ve transkriptom terimlerine ek olarak, herhangi bir organizmadaki tüm protein havuzuna ise proteom denmektedir. Genel olarak, süreçten de anlaşılacağı üzere, proteinlerin amino asit bileşimleri, DNA dizilimleri tarafından belirlenmektedir. Tüm bu sürece santral dogma adı verilmektedir [2]. Şekil 2.2'de santral dogma sürecinin adımları verilmiştir.



Şekil 2.2. Santral dogma süreci

Santral dogma kısaca, proteinin oluşması için DNA'dan RNA'ya doğru gerçekleşen bilgi akışı olarak ifade edilmektedir. Santral dogmada transkripsiyon ve translasyon olmak üzere iki önemli süreç bulunmaktadır. İlk olarak, DNA'da bulunan bilgi RNA mesajlarına dönüştürülmektedir. Bu sürece transkripsiyon adı verilir. Ardından, bu mesajlar, DNA'dan ribozomlara gönderilir ve burada belirli proteinlerin oluşturulabilmesi için çevrim işlemi yani translasyon yapılır.

2.3. Proteinlerin Yapı Seviyeleri

Proteinlerin bilinen dört adet yapısı bulunmaktadır. Bu yapılar birincil (primer), ikincil (sekonder), üçüncül (tersiyer) ve dördüncül (kuaterner) yapılardır.

2.3.1. Birincil Yapı

Proteinler yirmi adet amino asitten oluşan makro moleküllerdir. Proteinlerin biyolojik özellikleri üç boyutlu yapılara dayanmaktadır. Proteinlerde birincil yapı genellikle katlanmamış yapı olarak ifade edilmektedir. Bu katlanmamış yapı sadece amino asit dizilimlerinden oluşmaktadır. Her ne kadar sadece amino asit dizilimlerinden oluşsa da birincil yapı, ikincil ve üçüncül yapının belirlenmesinde etkili olmakta ve proteinin üç boyutlu formunun oluşmasında önemli bir rol üstlenmektedir [3]. Tablo 2.2’de amino asitler, 3-harfli kodları ve 1-harfli kodları verilmiştir.

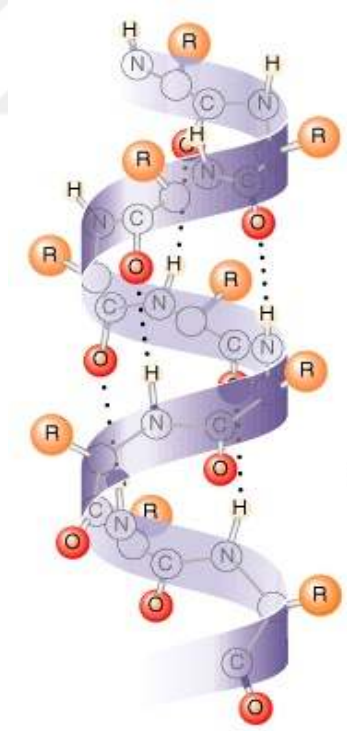
Tablo 2.2. Amino asitler ve kodları

Amino Asit	3-Harfli Kodu	1-Harfli Kodu
Alanin	Ala	A
Sistein	Cys	C
Aspartik asit	Asp	D
Glutamik asit	Glu	E
Fenilalanin	Phe	F
Glisin	Gly	G
Histidin	His	H
İzolösin	İle	I
Lizin	Lys	K
Lösin	Leu	L
Metionin	Met	M
Asparajin	Asn	N
Prolin	Pro	P
Glutamin	Gin	Q
Arjinin	Arg	R
Serin	Ser	S
Treonin	Thr	T
Valin	Val	V
Triptofan	Trp	W
Tirozin	Tyr	Y

Amino asitlerin yapısında hem amino hem karboksil grup bulunmaktadır. Her bir amino asidin bünyesinde α -karbon olarak bilinen asimetrik karbon atomu bulunmaktadır. Amino asitlerde bulunan α -karbon atomu dört farklı gruptan oluşmaktadır. Bu gruplar; amino grubu, hidrojen atomu, karboksil grubu ve genellikle R ile ifade edilen değişkendir. R grubu çoğunlukla yan zincir olarak belirtilmektedir. Amino asidin sahip olduğu özelliklere göre değişiklik gösterir. Organizmadaki en önemli görevi, amino asidin fonksiyonunu belirlemektir. Her amino asidin ilgili yan zinciri, hidrofobik, polar, asidik ya da bazik gibi kimyasal özellikleri oluşturmaktadır.

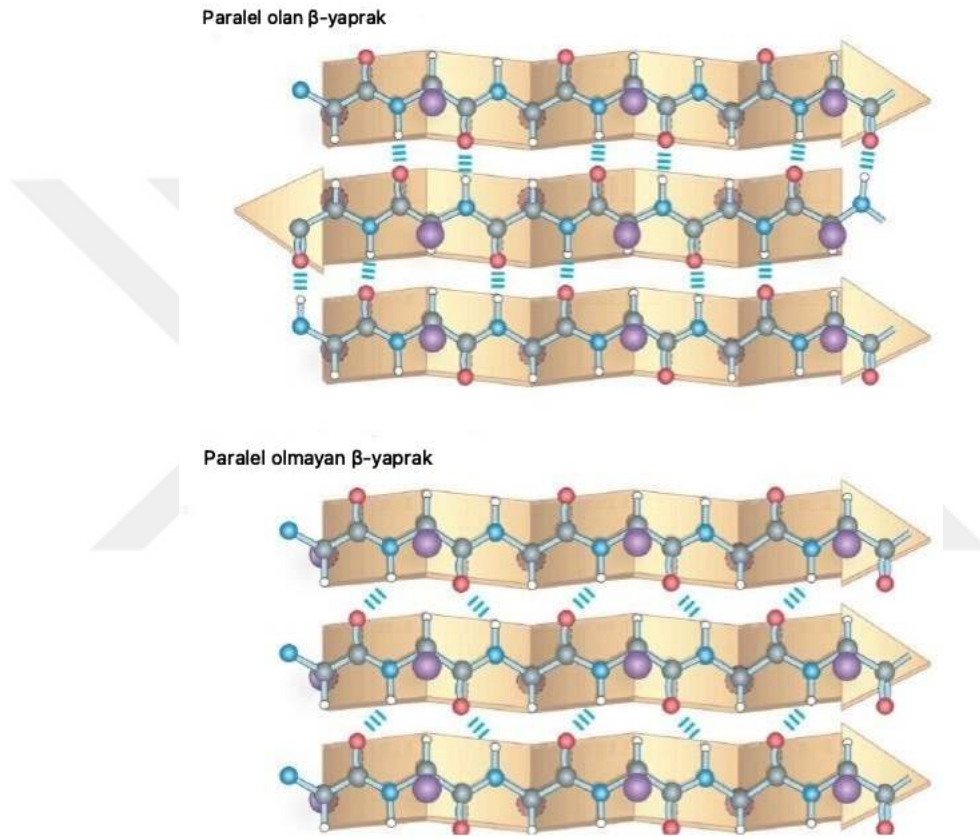
2.3.2. İkincil Yapı

Proteinin ikincil yapısı, tekrarlanan katman ya da kıvrımlardan meydana gelmektedir. Bu katmanlar hidrojen bağlarıyla oluşmaktadır. İkincil yapıda en çok α -heliks ve β -yaprak olmak üzere iki adet yapı görülmektedir. Her dört amino asitte bir kurulan ve kovalent olmayan H bağları ile oluşan yapıya α -heliks yapısı denmektedir. Bu yapının en önemli özelliklerinden birisi dayanıklı olmasıdır. Saç ve deri gibi dokularda sıklıkla görülmektedir. Şekil 2.3'te α -heliks yapısı verilmiştir.



Şekil 2.3. α -heliks yapısı [4]

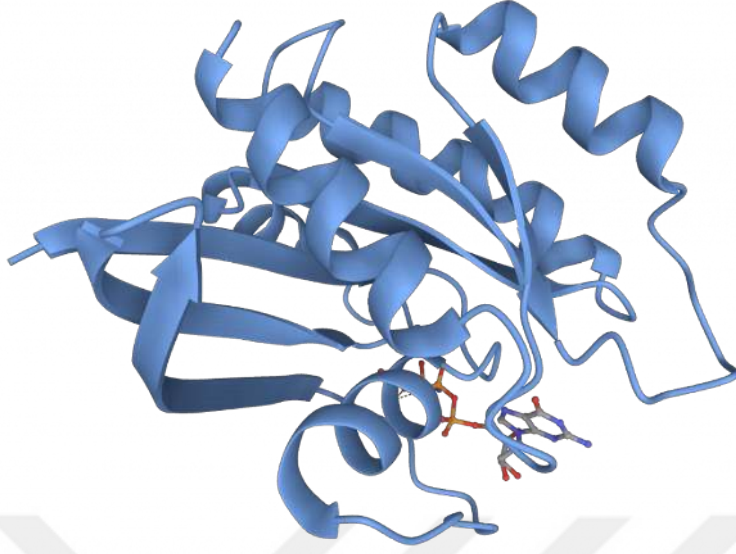
β -yaprak yapısında bulunan polipeptit zincirleri hem paralel hem de paralel olmayan yapıda bulunabilmektedir. Paralel bölgeler arasında H bağları bulunurken, bunlara ek olarak kovalent olmayan iyonik bağlar da bulunabilmektedir. Bu yapıda α -heliks yapısında olduğu gibi bir sarmal yapı bulunmamaktadır. Şekil 2.4'te paralel ve paralel olmayan β -yaprak yapılarının üstten görünümü verilmiştir.



Şekil 2.4. Paralel ve paralel olmayan β -yaprak yapılarının üstten görünümü [5]

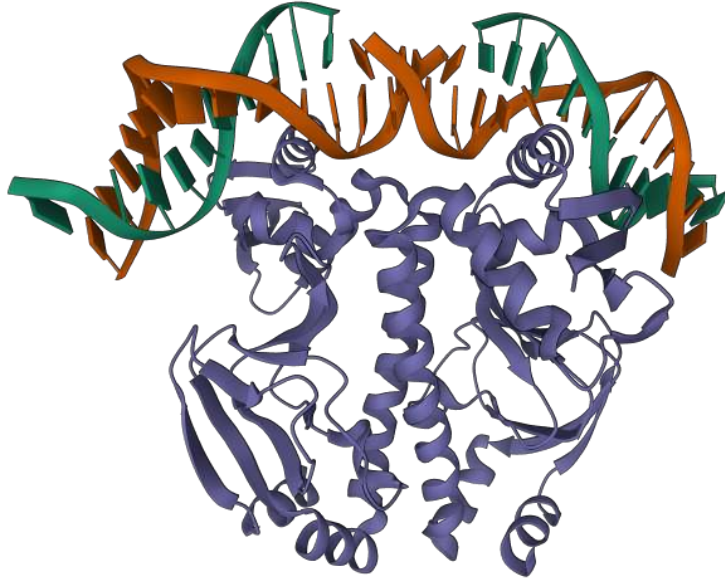
2.3.3. Üçüncül ve Dördüncül Yapı

Üçüncül yapı, ikincil yapıda bulunan elemanların üst düzey katlanmaları ile oluşmaktadır. Bu yapıda en sık görülen etkileşimler ise Van der Waals ve hidrofobik etkileşimlerdir. Bu etkileşimlerin dışında ise, tuz köprüleri, iyonik bağlar ve hidrojen bağları da bulunmaktadır. Şekil 2.5'te örnek bir üçüncül yapı verilmiştir.



Şekil 2.5. Protein üçüncül yapı örneği [6]

Proteinlerin bazılarında ise, fonksiyonel form iki ya da daha çok polipeptit zincirinin alt birimlerinin bir araya gelmesiyle meydana gelmektedir. Bu da proteinlerin dördüncül yapısını oluşturmaktadır. Şekil 2.6’da ise dördüncül yapı örneği verilmiştir.



Şekil 2.6. Protein dördüncül yapı örneği [7]

3. PROTEİN DİZİLİMLERİNİN SAYISALLAŞTIRILMASI

Bu bölümde, protein dizilimlerini sayısal ifadelere dönüştürmek için kullanılan protein sayısallaştırma yöntemlerinden bahsedilmiştir. Bu doğrultuda çeşitli kategorilere ait protein sayısallaştırma yöntemleri irdelenmiştir. Yapılan çalışmada, karakter-tabanlı, fizikokimyasal-tabanlı, evrim-tabanlı, yapı-tabanlı ve makine-öğrenmesi tabanlı olmak üzere, toplamda beş farklı kategoriye ait yöntem incelenmiştir.

3.1. Karakter Tabanlı Protein Sayısallaştırma Yöntemleri

Karakter-tabanlı protein sayısallaştırma yöntemlerinde, protein dizilimleri belirli bir bilgiye ihtiyaç duyulmadan sayısallaştırılmaktadır. Diğer kategoriye ait protein sayısallaştırma yöntemlerinde, proteinler yapı bilgilerine, evrimsel bilgilere ya da kimyasal bilgilere dayanarak sayısallaştırılırken karakter-tabanlı yöntemlerde böyle bir yaklaşım bulunmamaktadır. Çalışmada bu kategoride CPNR (Complex Prime Number Representation – Karmaşık Asal Sayı İfadesi) ve ikili-kodlama olmak üzere iki adet protein sayısallaştırma yöntemi irdelenmiştir.

3.1.1. CPNR Protein Sayısallaştırma Yöntemi

CPNR protein sayısallaştırma yöntemi ilk kez protein fonksiyonlarını kıyaslamak için önerilmiş ve kullanılmıştır [8]. Önerilen protein sayısallaştırma yönteminde, protein dizilimleri kodonlara göre sayısallaştırılmaktadır. Sayısallaştırma işleminin ilk aşamasında amino asitler kodon sayılarına ayrıştırılmıştır. Tablo 3.1’de amino asitler ve onlara karşılık gelen kodon sayıları verilmiştir.

Tablo 3.1. Amino asitler ve kodon sayıları

Kodon Sayıları	Amino Asitler
1	M, W
2	F, Y, C, H, Q, N, D, K, E
3	I
4	P, T, V, A, G
6	L, S, R

Tablo 3.1’de verilen bilgiler doğrultusunda, M amino asidinin kodon sayısı 1’dir. Benzer şekilde W amino asidinin kodon sayısı da 1’dir. L amino asidinin kodon sayısı 6’yken, P amino asidinin kodon sayısı 4’tür. Tablo 3.2’de asal sayılar arasındaki fark ve onlara karşılık gelen sayısal değerler verilmiştir.

Tablo 3.2. Asal sayılar arasındaki fark ve karşılık gelen sayısal ifadeler

Kodon Sayıları	Amino Asitler
1	1, 2
2	3, 5, 11, 17, 29, 41, 59
3	Ø
4	7, 13, 19, 37, 43
6	23, 31, 47, 53, 61

1 ve 67 arasında bulunan asal sayılar şu şekildedir; [1, 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67].

Verilen listeden de görüldüğü üzere yirmi adet sayı bulunmaktadır. 1'den sonra gelen ilk asal sayı 2'dir. Bu iki sayı arasındaki fark 1 olduğu için, 1 değeri kodon sayısı bir olan ilk amino asit koduna (M) atanmıştır. 2'den sonra gelen ilk asal sayı 3'tür. Bu iki sayı arasındaki fark 1 olduğu için, 2 değeri de kodon sayısı bir olan ikinci amino asit koduna (W) verilmiştir. 3'ten sonra gelen ilk asal sayı ise 5'tir. Bu iki sayı arasındaki fark 2 olduğu için, 3 değeri kodon sayısı iki olan ilk amino asit koduna (F) atanmıştır. Bütün bu işlem yirmi amino asit için de yapılmıştır. Tablo 3.3'te ise her bir amino asit kodunun CPNR değerleri verilmiştir.

Tablo 3.3. Amino asitlerin CPNR değerleri

Amino Asit Kodu	CPNR Değeri	Amino Asit Kodu	CPNR Değeri
M	1	Q	29
W	2	S	31
F	3	A	37
Y	5	N	41
P	7	G	43
C	11	R	47
T	13	I	53
H	17	D	59
V	19	E	61
L	23	K	67

Tablo 3.3.'te verilen değerler doğrultusunda, $P(N) = [M, W, F, Y, P, C, \dots]$ şeklindeki bir protein dizilimi, CPNR yöntemi ile $M(N) = [1, 2, 3, 5, 7, 11, \dots]$ biçiminde sayısal hale getirilmektedir.

3.1.2. İkili Kodlama Sayısallaştırma Yöntemi

İkili kodlama sayısallaştırma yönteminde, protein dizilimlerindeki amino asitler “0” ve “1” değerleri ile temsil edilmektedir. En yaygın olarak kullanılan ikili kodlama yöntemi ise ortogonal kodlama olarak da bilinen tek-sıcak (one-hot) kodlama yöntemidir [9]. Bu yöntemde, yirmi adet standart amino asidin her biri yirmi boyutlu ikili vektör ile sayısallaştırılmaktadır. Sayısallaştırma işleminin ilk aşamasında amino asitler alfabetik bir şekilde sıralanır. Ardından i . sıradaki amino asit değeri “1” değerini alırken, diğerleri “0” değerini almaktadır. Tablo 3.4’te amino asitlerin ikili-kodlama değerleri verilmiştir.

Tablo 3.4. Amino asitlerin ikili-kodlama değerleri

Amino Asit Kodu	İkili-Kodlama Değeri	Amino Asit Kodu	İkili-Kodlama Değeri
A	10000000000000000000	M	00000000001000000000
C	01000000000000000000	N	00000000000100000000
D	00100000000000000000	P	00000000000010000000
E	00010000000000000000	Q	00000000000001000000
F	00001000000000000000	R	00000000000000100000
G	00000100000000000000	S	00000000000000010000
H	00000010000000000000	T	00000000000000001000
I	00000001000000000000	V	00000000000000000100
K	00000000100000000000	W	00000000000000000010
L	00000000010000000000	Y	00000000000000000001

Tablo 3.4’te verilen değerler doğrultusunda, $P(N) = [A, C, D, \dots]$ şeklindeki bir protein diziliminin ikili kodlama yöntemi ile sayısallaştırılmış biçimi;

$$M(N) = [10000000000000000000, 01000000000000000000, 00100000000000000000]$$

3.2. Fizikokimyasal Tabanlı Protein Sayısallaştırma Yöntemleri

Fizikokimyasal-tabanlı protein sayısallaştırma yöntemlerinde ise, karakter-tabanlı yöntemlerin aksine, protein dizilimleri, proteinlerin kimyasal bilgilerine dayanarak sayısallaştırılmaktadır. Çalışmada, fizikokimyasal-tabanlı protein sayısallaştırma kategorisinde hidrofobiklik, Meiler parametreleri, Atchley faktörleri ve EIIP (Electron-Ion Interaction Potential-Elektron-İyon Etkileşim Potansiyeli) olmak üzere dört farklı protein sayısallaştırma yöntemi irdelenmiştir.

3.2.1. Hidrofobiklik Protein Sayısallaştırma Yöntemi

Bir proteinin biyolojik fonksiyonu, protein diziliminin kendisi ve protein diziliminin kimyasal özellikleri tarafından belirlenmektedir. Yirmi adet amino asidin yüzlerce fizikokimyasal özellikleri bulunmaktadır. Bu fizikokimyasal özelliklerin en önemlileri polarize edilebilirlik, Van der Waals hacmi, iyonizasyon sabiti, erişilebilir solvent yüzey alanı ve hidrofobiklik. Hidrofobiklik protein sayısallaştırma yönteminde, yirmi adet amino asit yan zincirinin her birinin hidrofilik ve hidrofobik özellikleri ele alınmaktadır [10]. Amino asitlerin hidrofobiklik özellikleri, amino asit dizilimlerini karşılaştırmak ve analiz etmek için etkili bir yol olarak sıklıkla kullanılmaktadır. Hidrofobiklik protein sayısallaştırma yöntemi bir boyutlu sayısal verilerden oluşmaktadır. Tablo 3.5'te amino asitlerin hidrofobiklik değerleri verilmiştir.

Tablo 3.5. Amino asitlerin hidrofobiklik değerleri

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

Tablo 3.5'te verilen değerler doğrultusunda, $P(N) = [M, W, F, Y, P, C, \dots]$ biçimindeki bir protein dizilimi, hidrofobiklik yöntemi ile $M(N) = [1.9, -0.9, 2.8, -1.3, -1.6, 2.5, \dots]$ şeklinde sayısallaştırılır.

3.2.2. Meiler Parametreleri Protein Sayısallaştırma Yöntemi

Başka bir fizikokimyasal-tabanlı sayısallaştırma yöntemi olan Meiler parametreleri ise amino asitlerin indirgenmiş temsillerini ifade etmek için önerilmiştir [11]. Bu yöntemin en büyük avantajı, sayısallaştırma işleminin hızlı ve verimli olmasıdır. Meiler parametreleri toplamda; polarize edilebilirlik (pe), hacim (h), hidrofobiklik (hi), sterik (s), yaprak olasılığı (yo), helix olasılığı (ho) ve izoelektronik nokta (in) olmak üzere 7 farklı parametreden meydana gelmektedir. Bu nedenle 7-boyutludur. Tablo 3.6'da amino asitlerin Meiler parametreleri verilmiştir.

Tablo 3.6. Amino asitlerin Meiler parametreleri

Amino Asit Kodu	s	pe	h	hi	in	ho	yo
A	1.28	0.05	1.00	0.31	6.11	0.42	0.23
R	2.34	0.29	6.13	-1.01	10.74	0.36	0.25
N	1.60	0.13	2.95	-0.60	6.52	0.21	0.22
D	1.60	0.11	2.78	-0.77	2.95	0.25	0.20
C	1.77	0.13	2.43	1.54	6.35	0.17	0.41
E	1.56	0.15	3.78	-0.64	3.09	0.42	0.21
Q	1.56	0.18	3.95	-0.22	5.65	0.36	0.25
G	0.00	0.00	0.00	0.00	6.07	0.13	0.15
H	2.99	0.23	4.66	0.13	7.69	0.27	0.30
I	4.19	0.19	4.00	1.80	6.04	0.30	0.45
L	2.59	0.19	4.00	1.70	6.04	0.39	0.31
K	1.89	0.22	4.77	-0.99	9.99	0.32	0.27
M	2.35	0.22	4.43	1.23	5.71	0.38	0.32
F	2.94	0.29	5.89	1.79	5.67	0.30	0.38
P	2.67	0.00	2.72	0.72	6.80	0.13	0.34
S	1.31	0.06	1.60	-0.04	-5.70	0.20	0.28
T	3.03	0.11	2.60	0.26	5.60	0.21	0.36
W	3.21	0.41	8.08	2.25	5.94	0.32	0.42
Y	2.94	0.30	6.47	0.96	5.66	0.25	0.41
V	3.67	0.14	3.00	1.22	6.02	0.27	0.49

Tablo 3.6’da verilen parametreler doğrultusunda, $P(N) = [A, R, \dots]$ biçimindeki bir protein diziliminin sayısallaştırılmış hali;

$$M(N) = [[1.28 \ 0.05 \ 1.00 \ 0.31 \ 6.11 \ 0.42 \ 0.23], [2.34 \ 0.29 \ 6.13 \ -1.01 \ 10.74 \ 0.36 \ 0.25], \dots]$$

3.2.3. Atchley Faktörleri Protein Sayısallaştırma Yöntemi

Atchley faktörleri ise, protein dizilimlerinde sıklıkla görülen metrik problemlerini gidermek için önerilmiş bir yöntemdir [12]. Protein dizilimlerinde bulunan varyans ve kovaryans değişkenleri, son yıllarda gelişen teknolojiler ile amino asit kodlarının alfabetik sıralanmasına ve bilgi teorisine dayanarak sayısallaştırılmaktadır. Araştırmacıların çoğu, protein dizilimlerinde bulunan bu varyans ve kovaryans değişkenlerini belirleyebilmek için çok değişkenli analiz yöntemini kullanmaktadır. Ancak, bu yöntem ile protein dizilim siteleri arasında bulunan negatif bağlantıları belirlemek zor olmaktadır [12]. Bu problemin önüne geçebilmek için Atchley faktörleri sunulmuştur. Atchley faktörlerinde toplamda beş adet faktör bulunmaktadır. Bu faktörler; ikincil yapı (iy), moleküler hacim (mh), polarite (p), protein elektrostatik yükü (ey) ve kodon yoğunluğu

(ky) faktörleridir. Beş adet faktör bulunduğu için 5-boyutlu bir yapıya sahiptir. Tablo 3.7’de amino asitlerin Atchley faktörleri verilmiştir.

Tablo 3.7. Amino asitlerin Atchley faktörleri

Amino Asit Kodu	p	iy	mh	ky	ey
A	-0.591	-1.302	-0.733	1.570	-0.146
R	1.538	-0.055	1.502	0.440	2.897
N	0.945	0.828	1.299	-0.169	0.933
D	1.050	0.302	-3.656	-0.259	-3.242
C	-1.343	0.465	-0.862	-1.020	0.255
E	1.357	-1.453	1.477	0.113	-0.837
Q	0.931	-0.179	-3.005	-0.503	-1.853
G	-0.384	1.652	1.330	1.045	2.064
H	0.336	-0.417	-1.673	-1.474	-0.078
I	-1.239	-0.547	2.131	0.393	0.816
L	-1.019	-0.987	-1.505	1.266	-0.912
K	1.831	-0.561	0.533	-0.277	1.648
M	-0.663	-1.524	2.219	-1.005	1.212
F	-1.006	-0.590	1.891	-0.397	0.412
P	0.189	2.081	-1.628	0.421	-1.392
S	-0.228	1.399	-4.760	0.670	-2.647
T	-0.032	0.326	2.213	0.908	1.313
W	-0.595	0.009	0.672	-2.128	-0.184
Y	0.260	0.830	3.097	-0.838	1.512
V	-1.337	-0.279	-0.544	1.242	-1.262

Tablo 3.7’de verilen faktörlere göre, $P(N) = [A, R, \dots]$ şeklindeki bir protein diziliminin, Atchley faktörleri ile sayısallaştırılmış biçimi;

$$M(N) = [-0.591 - 1.302 - 0.733 \ 1.570 - 0.146], [1.538 - 0.055 \ 1.502 \ 0.440 \ 2.897], \dots]$$

3.2.4. EIIP Protein Sayısallaştırma Yöntemi

EIIP protein sayısallaştırma yöntemi ilk olarak DNA ve proteinler arasındaki etkileşimleri belirlemek için kullanılmıştır [13]. EIIP yönteminin ilk aşamasında protein dizilimleri Fourier dönüşümü ile sinyallere ayrıştırılmıştır. Ayrıştırılma işleminin ardından sinyallerden güç spektrum yoğunluğu değerleri hesaplanmıştır. Son olarak her bir güç spektrum değeri, ilgili amino asitlere tanımlanmış ve sayısallaştırma işlemi gerçekleştirilmiştir. Tablo 3.8’de amino asitlerin EIIP değerleri verilmiştir.

Tablo 3.8. Amino asitlerin EIIP deęerleri

Amino Asit Kodu	EIIP Deęeri	Amino Asit Kodu	EIIP Deęeri
M	0.0823	Q	0.0761
W	0.0548	S	0.0829
F	0.0946	A	0.0373
Y	0.0516	N	0.0036
P	0.0198	G	0.0050
C	0.0829	R	0.0959
T	0.0941	I	0
H	0.0242	D	0.1263
V	0.0057	E	0.0058
L	0	K	0.0371

Tablo 3.8’de verilen deęerler doęrultusunda, $P(N) = [M, W, F, Y, P, C, \dots]$ şeklindeki bir protein dizilimi, EIIP yöntemi ile $M(N) = [0.0823, 0.0548, 0.0946, 0.0516, 0.0198, 0.0829, \dots]$ biçiminde sayısallaştırılır.

3.3. Evrim Tabanlı Protein Sayısallaştırma Yöntemleri

Evrimsel tabanlı protein sayısallaştırma yöntemlerinde ise, protein dizilimleri, proteinlerin evrimsel bilgilerine dayanarak sayısallaştırılmaktadır. Proteinlerin evrimsel bilgileri, filogenetik ağaçlardan ya da dizilim hizalama işlemlerinden elde edilmektedir. Konuma baęlı ve konumdan baęımsız olmak üzere iki ayrı türü bulunmaktadır. Evrim-tabanlı protein sayısallaştırma kategorisinde, PAM250 (Point Accepted Mutation – Nokta Kabul Edilen Mutasyon), BLOSUM62 ve PSSM (Position Specific Scoring Matrix – Konuma Özgü Puanlama Matrisi) olmak üzere üç farklı sayısallaştırma yöntemi incelenmiştir. PAM250 v BLOSUM62 matrisleri konumdan baęımsız yöntemlerken, PSSM matrisi ise konuma baęlı yöntemdir. Konuma baęlı yöntemlerde, amino asit dizilimleri, amino asitlerin yeri ve bileşimine göre sayısallaştırılmaktadır. Konumdan baęımsız yöntemlerde ise böyle bir durum söz konusu değildir.

3.3.1. PAM250 Protein Sayısallaştırma Yöntemi

PAM250 matrisi, hizalanan peptit dizilimlerini puanlayarak protein dizilimlerindeki benzerlikleri belirlemek için kullanılan bir yöntemdir [14]. PAM250 matrisindeki deęerler, bilinen homolojiye sahip hizalanmış protein dizilimlerinin kıyaslanmasından ve gözlemlenen nokta mutasyonlarının belirlenmesinden elde edilmiştir. Mutasyonların frekansları Denklem 3.1’de verilen formüle göre hesaplanmaktadır;

$$M_{ij} = 10(\log_{10} R_{ij}) \quad (3.1)$$

Denklem 3.1’de M_{ij} matris elemanını ifade etmektedir. Bunun yanı sıra, R_{ij} ise, i. matrisin mutasyon olasılığını belirtmektedir. Hesaplama sonucunda elde edilen değerler, en yakın tam sayı değerine yuvarlanmaktadır. Şekil 3.1’de PAM250 skor matrisi verilmiştir.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Şekil 3.1. PAM 250 skor matrisi [15]

Şekil 3.1’den de anlaşılacağı üzere, skor matrisinin hem dikey kısmında hem de yatay kısmında amino asitler ve onların skorları bulunmaktadır. Proteinleri sayısallaştırmak için ise, ilgili amino asidin karşısına denk gelen aynı amino asidin skoru göz önünde bulundurulmaktadır. Örnek verilecek olursa, PAM250 skor matrisinde verilen değerlere göre, Y amino asidinin sayısal değeri 10 olmaktadır. Y amino asidinin karşısına gelen Y amino asidinin değeri 10’dur. Tablo 3.9’da amino asitlerin PAM250 skorları verilmiştir.

Tablo 3.9. Amino asitlerin PAM250 deęerleri

Amino Asit Kodu	PAM250 Deęeri	Amino Asit Kodu	PAM250 Deęeri
A	2	L	6
R	6	K	5
N	2	M	6
D	4	F	9
C	4	P	6
Q	4	S	3
E	4	T	3
G	5	W	17
H	6	Y	10
I	5	V	4

Tablo 3.9’da verilen deęerlere gre, $P(N) = [A, R, N, D, C, Q, \dots]$ eklindeki bir protein dizilimi, PAM250 skor matrisi ile $M(N) = [2, 6, 2, 4, 4, 4, \dots]$ biiminde ifade edilir.

3.3.2. BLOSUM62 Protein Sayısallařtırma Yntemi

BLOSUM matrisi, biyoenformatik alanında sıklıkla kullanılan ve protein dizilimlerini hizalamak iin bařvurulan yntemlerin bařında gelmektedir. BLOSUM matrisleri, evrimsel olarak birbirinden farklı protein dizilimleri arasındaki hizalamaları puanlamak iin kullanılmaktadır. Konumdan bağımsız bir yntemdir ve ilk kez 1992 yılında nerilmiřtir [16]. BLOSUM80, BLOSUM45 gibi farklı trde matrisleri bulunmaktadır. Matrislerin yanındaki sayılar benzerlik deęerini ifade etmektedir. Evrimsel aıdan birbirine en yakın iki organizma arasındaki benzerlikler yksek skor deęerleri almaktadır. Bundan dolayı, BLOSUM matrislerinde skor ne kadar yksek ise, benzerlik de o kadar fazla olmaktadır. rneęin, birbirine tr olarak yakın olan iki organizmanın protein dizilimleri hizalanırken BLOSUM80 matrisi kullanılmaktadır. Buna zıt olarak ise, birbirine tr olarak uzak olan iki organizmanın proteinleri ise BLOSUM45 matrisi ile hizalanabilmektedir. Ancak, organizmaların uzaklık ya da yakınlık durumları tam olarak bilinmedięinde ise BLOSUM62 matrisine bařvurulmaktadır. BLOSUM62 matrisinde proteinler arasındaki skorlar hesaplanırken iki temel durum gz nnde bulundurulmaktadır. Bu durumlar;

- Bir amino asidin karřısına, bařka bir amino asidin ne kadar sıklıkla geldięi,
- Amino asidin ortamda grlme olasılıęıdır.

řekil 3.2’de BLOSUM62 matrisi verilmiřtir.

Şekil 3.2. BLOSUM62 matrisi

Tablo 3.10. Amino asitlerin BLOSUM62 deęerleri

Tablo 3.10’da verilen değerlere göre, $P(N) = [A, R, N, D, C, Q, \dots]$ biçimindeki bir protein dizilimi, BLOSUM62 skorları ile $M(N) = [4, 5, 6, 6, 9, 5, \dots]$ şeklinde sayısallaştırılır.

3.3.3. PSSM Protein Sayısallaştırma Yöntemi

Homolog protein dizilimlerini hizalamak ve protein fonksiyonlarını bulmak için çeşitli hizalama uygulamaları kullanılmaktadır. Bunlara ek olarak, benzer fonksiyonlara sahip proteinleri belirleyebilmek için bazen protein motifleri de kullanılmaktadır. Protein dizilimleri homolog dizilimler olabilirken, aynı zamanda evrimsel olarak birbirlerine yakın olabilmektedirler. Bundan dolayı, protein dizilimlerinde kısa motifler aranırken, genellikle ortak kökenlerinden ziyade ortak fonksiyonlara bakılmaktadır. Protein dizilimlerindeki kısa motifleri bulabilmek için en çok kullanılan yöntemlerden birisi PSSM yöntemidir. PSSM yönteminde, belirli pozisyonlarda bulunan belirli amino asitler ya da nükleotitler puanlanmaktadır. PSSM yönteminde toplamda dört adet satır ve dokuz adet sütun bulunmaktadır. Satırlar, A, C, G ve T olmak üzere dört adet nükleotit bazını ifade etmektedir. Sütunlar ise protein motiflerindeki dokuz adet pozisyonu belirtmektedir. En yaygın kullanılan protein motiflerinden birisi ABRE motifidir. ABRE motifinin konsensüs dizisi ACGTG G/TC şeklindedir [17]. PSSM yöntemi ile işlem yapılırken önce motifinin pozisyon matrisi belirlenir. ABRE motifine ait kırk yedi adet dizilimin pozisyon matrisi ise Tablo 3.11’de verilmiştir [17];

Tablo 3.11. ABRE motifine ait bazların pozisyon değerleri

Baz	1	2	3	4	5	6	7	8	9
A	15	4	47	0	0	0	3	0	14
C	15	20	0	47	0	0	1	1	20
G	8	3	0	0	47	0	43	18	5
T	9	20	0	0	0	47	0	28	8

Pozisyon matrisinden anlaşılaçağı üzere, kırk yedi adet ABRE motifine ait dizilimlerin birinci pozisyonunda on beş adet A bazı bulunmaktadır. İkinci pozisyonunda ise A bazının sayısı dörtttür. Üçüncü pozisyonunda ise tamamıyla A bazı bulunduğu için bu pozisyonun değeri kırk yedidir. Dört, beş ve altıncı pozisyonlarda ise A bazı hiç bulunmamaktadır. Pozisyon matrisi oluşturulduktan sonra ise frekans matrisi hesaplanarak, her bir bazın, her bir pozisyondaki frekansı belirlenmektedir. Tablo 3.12’de bazların her bir pozisyondaki frekans değerleri verilmiştir.

Tablo 3.12. Bazların her bir pozisyon için frekans değerleri

Baz	1	2	3	4	5	6	7	8	9
A	0.32	0.09	1.00	0.00	0.00	0.00	0.06	0.00	0.30
C	0.32	0.43	0.00	1.00	0.00	0.00	0.02	0.02	0.43
G	0.17	0.06	0.00	0.00	1.00	0.00	0.91	0.38	0.11
T	0.19	0.43	0.00	0.00	0.00	1.00	0.00	0.60	0.17

Frekans değerleri bulunurken, pozisyonadaki baz sayısı, dizilimin uzunluğuna bölünmektedir. Örneğin, A bazının üçüncü pozisyonadaki değeri 47'dir. Dizilimin uzunluk değeri 47 olduğu için, 47/47'den, A bazının üçüncü pozisyonadaki frekans değeri 1 olarak hesaplanmaktadır. Başka bir örnekte ise, G bazının ikinci pozisyonadaki değeri 3'tür. Bu değer 47'ye bölünmesiyle 0.06 değeri elde edilmektedir. Bu değer de G bazının üçüncü pozisyonadaki frekans değerini ifade etmektedir. Hesaplamaların hepsi her bir bazın, her bir pozisyonadaki değeri için tek tek yapılmaktadır. Frekans değerleri hesaplandıktan sonra ise bazların logit değerleri belirlenmektedir. Denklem 3.2'de logit fonksiyonunu hesaplama adımları verilmiştir.

$$\text{logit} = 2\log_2 \left(\frac{P_i}{0.25} \right) \quad (3.2)$$

Denklem 3.2'de verilen P_i değeri, bazın o pozisyonadaki olasılığını yani frekansını ifade etmektedir. 0.25 değeri ise toplamda 4 adet baz bulunduğu için, her bir bazın o pozisyonda bulunma olasılığını belirtmektedir. Bu denklemde verilen formül doğrultusunda hesaplanan logit değerleri Tablo 3.13'te verilmiştir.

Tablo 3.13. Bazların logit değerleri

Baz	1	2	3	4	5	6	7	8	9
A	0.70	-3.11	4.00	$-\infty$	$-\infty$	$-\infty$	-3.94	$-\infty$	0.51
C	0.70	1.53	$-\infty$	4.00	$-\infty$	$-\infty$	-7.11	-7.11	1.53
G	-1.11	-3.94	$-\infty$	$-\infty$	4.00	$-\infty$	3.74	1.23	-2.47
T	-0.77	1.53	$-\infty$	$-\infty$	$-\infty$	4.00	$-\infty$	2.51	-1.11

Logit değerleri hesaplandıktan sonra, değerler en yakın tam sayı değerine yuvarlanır ve bazların tam sayı haline getirilmiş logit değerleri elde edilmektedir. Tablo 3.14'te bazların tam sayı haline getirilmiş logit değerleri verilmiştir.

Tablo 3.14. Bazların tam sayı haline getirilmiş logit değerleri

Baz	1	2	3	4	5	6	7	8	9
A	1	-3	4	$-\infty$	$-\infty$	$-\infty$	-4	$-\infty$	1
C	1	2	$-\infty$	4	$-\infty$	$-\infty$	-7	-7	2
G	-1	-4	$-\infty$	$-\infty$	4	$-\infty$	4	1	-2
T	-1	2	$-\infty$	$-\infty$	$-\infty$	4	$-\infty$	3	-1

Bu işlemden sonra, ABRE motifiyle eşleşen herhangi bir nükleotit dizisinin logit skoru hesaplanabilmektedir. Örneğin, $S(N) = [GCACGTGGC]$ şeklinde verilen bir dizilimin logit skoru $[-1 + 2 + 4 + 4 + 4 + 4 + 4 + 4 + 1 + 2] = 24$ olur. Ancak, $S(N) = [CCACGTGAC]$ şeklinde verilen dizilimin logit skoru hesaplanamamaktadır. Çünkü bu dizilimde A bazının sekizinci pozisyonundaki değeri $-\infty$ 'dur. Böylece dizilimin logit skoru $-\infty$ olmakta ve dizilim sayısallaştırılamamaktadır. Bu problemin önüne geçebilmek için en çok kullanılan yöntemlerden birisi ortalama yöntemidir. Denklem 3.3'te ortalama işleminin formülü verilmiştir.

$$\text{Ortalama} = 2\log_2 \left(\sum_{i=1}^4 P_i \frac{P_{ij}}{q_i q_j} \right) \quad (3.3)$$

Denklem 3.3'te verilen eşitlikte, P_i ifadesi bazın o konumdaki olasılığını ifade etmektedir. P_{ij} değeri ise bazın i . satır ve j . sütundaki bulunma olasılığını göstermektedir. q_i değerleri ise bazların eşit frekans sonuçlarını göstermektedir. Ortalama hesabı her bir baz için yapılmaktadır. Ortalama hesabını gerçekleştirebilmek için öncelikle bazların eşit frekans değerleri hesaplanmaktadır. Bazların eşit frekans değerleri Tablo 3.15'te verilmiştir.

Tablo 3.15. Bazların eşit frekans değerleri

Baz	A	C	G	T
A	0.13	0.03	0.03	0.06
C	0.03	0.13	0.06	0.03
G	0.03	0.06	0.13	0.03
T	0.06	0.03	0.03	0.13

Denklem 3.3'te verilen formüle göre hesaplama yapıldığında sekizince pozisyonda bulunan A bazının yeni değeri şu şekilde hesaplanır; $2\log_2((0.00 * 0.13 + 0.02 * 0.03 + 0.38 * 0.03 + 0.60 * 0.06) * 16) = -0.77$. Bu hesaplama işlemi sayesinde $-\infty$ olan değer -0.77 olarak bulunur. Ortalama hesabı tüm bazlar için yapılır ve $-\infty$ değerleri göz ardı edilir. Son durumda ise bazların PSSM değerleri Tablo 3.16'da verilmiştir.

Tablo 3.16. Bazların PSSM deęerleri

Baz	1	2	3	4	5	6	7	8	9
A	0	-1	2	-2	-2	0	-2	-1	0
C	0	1	-2	2	0	-2	0	-1	1
G	0	-1	-2	0	2	-2	2	0	0
T	0	1	0	-2	-2	2	-2	1	0

3.4. Yapı Tabanlı Protein Sayısallaştırma Yöntemleri

Yapı-tabanlı protein sayısallaştırma yöntemlerinde, proteinler, proteinlerin yapı bilgisine dayanarak sayısallaştırılır. Bu bilgiler, proteinlerin birincil, ikincil, üçüncül ve dördüncül yapılarından gelmektedir. Bu kategoride çeşitli sayıda yöntem bulunsa da diğer kategorilere nazaran daha az kullanılmaktadır. Bunun nedenlerinden birisi protein yapı bilgilerinin elde edilmesindeki zorluktur. Bu kategoride Miyazawa enerjileri ve Micheletti potansiyelleri en çok kullanılan yöntemlerdendir. Bu nedenle çalışmada sadece bu iki yöntem üzerinde durulmuştur.

3.4.1. Miyazawa Enerjileri Protein Sayısallaştırma Yöntemi

Miyazawa enerjileri, protein dizilimlerindeki kalıntılar arasındaki enerjileri belirlemek için önerilen bir yöntemdir [18]. Bu yöntemde protein dizilimlerinin temas enerjileri regresyon katsayıları ile elde edilir. Miyazawa enerjilerinde enerji deęerleri, ikincil yapı enerjileri ve üçüncül yapı enerjileri olmak üzere iki terime ayrılmıştır. Üçüncül yapı enerjileri, proteinlerin kalıntı-kalıntı arasındaki temas enerjilerinin toplanmasıyla elde edilir. İkincil yapı enerjileri ise atom zincirleri arasındaki etkileşimlere ve atom zincir-yan zincir arasındaki etkileşimlerine göre hesaplanır. Tablo 3.17’de proteinlerin Miyazawa enerji deęerleri verilmiştir.

Tablo 3.17. Amino asitlerin Miyazawa enerjileri

Amino Asit Kodu	Miyazawa Enerji Deęeri	Amino Asit Kodu	Miyazawa Enerji Deęeri
A	-0.02	L	-0.32
R	0.08	K	0.30
N	0.10	M	-0.25
D	0.19	F	-0.33
C	-0.32	P	0.11
Q	0.21	S	0.11
E	0.15	T	0.05
G	-0.02	W	-0.27
H	-0.02	Y	-0.23
I	-0.28	V	-0.23

Tablo 3.17’de verilen değerlere göre, $P(N) = [A, R, N, D, C, Q, \dots]$ biçimindeki bir protein dizilimi, Miyazawa enerjileri ile $M(N) = [-0.02, 0.08, 0.10, 0.19, -0.32, 0.21, \dots]$ şeklinde sayısal hale getirilir.

3.4.2. Micheletti Potansiyelleri Protein Sayısallaştırma Yöntemi

Diğer bir yapı-tabanlı protein sayısallaştırma yöntemi olan Micheletti potansiyelleri, proteinler arasında ortaya çıkan etkileşimlerdeki potansiyel enerjiye dayanmaktadır [19]. Micheletti potansiyellerinin ana fikri, en uygun protein etkileşimlerini belirlemektir. Tablo 3.18’de proteinlerin Micheletti potansiyel değerleri verilmiştir.

Tablo 3.18. Amino asitlerin Micheletti potansiyelleri

Amino Asit Kodu	Micheletti Potansiyel Değeri	Amino Asit Kodu	Micheletti Potansiyel Değeri
A	-0.001461	L	-0.000782
R	0.009875	K	0.005109
N	-0.001962	M	0.031655
D	-0.000531	F	-0.013128
C	-0.002544	P	-0.003621
Q	0.006456	S	-0.000802
E	0.008438	T	0.003269
G	0.000990	W	0.131813
H	0.001314	Y	-0.007699
I	0.006801	V	0.001445

Tablo 3.18’de verilen değerler doğrultusunda, $P(N) = [A, R, N, D, \dots]$ şeklindeki bir protein dizilimi, Micheletti potansiyelleri ile $M(N) = [-0.001461, 0.009875, -0.001962, -0.000531]$ olacak şekilde sayısal ifadelerle dönüştürülür.

3.5. Makine Öğrenmesi Tabanlı Protein Sayısallaştırma Yöntemleri

Bu kategoride kullanılan protein sayısallaştırma yöntemlerinde sayısallaştırma işlemi makine öğrenmesine dayanmaktadır. Evrimsel, yapısal ya da fizikokimyasal yöntemlerde olduğu gibi belirli bir bilgiye ihtiyaç duyulmamaktadır. Bu kategoride fazla yöntem olmasa da çeşitli çalışmalarda bu yöntemlerin kullanıldığı gözlemlenmiştir. Tez çalışması kapsamında, bu kategoriye ait yöntemlerden sadece AESNN3, ANN4D ve ProtVec olmak üzere üç farklı protein sayısallaştırma yöntemi incelenmiştir.

3.5.1. AESNN3 Protein Sayısallaştırma Yöntemi

AESNN3 protein sayısallaştırma yöntemi, denetimli geri-beslemeli sinir ağı modeli kullanılarak geliştirilmiştir [20]. Geliştirilen bu yöntemde, ilk olarak, benzer biyofiziksel özelliklere sahip amino asitler bir arada kümelenmiştir. Kümeleme işleminin ardından araştırmacılar, eğitim ve test verisini oluşturmuşlar ve bu doğrultuda 800 adet eğitim verisi kullanırlarken, 220 adet test verisi kullanmışlardır. Veri seti oluşturulduktan sonra hizalama işlemi yapılmış ve hizalama işlemi için 0 ve 1 aralığında değerler belirlenmiştir. Eğitim işlemi için 6 katlı çapraz-doğrulama yöntemi kullanılmış ve her bir doğrulama 10 iterasyon ile gerçekleştirilmiştir. En düşük hata veren model ile analiz işlemi devam etmiştir. Sinir ağı eğitildikten sonra, her bir amino asit tipi girdi katmanına verilmiş ve ağı beslenerek, 3 gizli katmandan elde edilen ağırlık değerleri ile amino asitler sayısallaştırılmıştır. Tablo 3.19’da amino asitlerin AESNN3 değerleri verilmiştir. Her bir amino asit 3 adet değerden oluştuğu için 3-boyutlu bir vektör yapısı vardır.

Tablo 3.19. Amino asitlerin AESNN3 değeri

Amino Asit Kodu	Gizli Katman 1	Gizli Katman 2	Gizli Katman 3
A	-0.99	-0.61	0.00
R	0.28	-0.99	-0.22
N	0.77	-0.24	0.59
D	0.74	-0.72	-0.35
C	0.34	0.88	0.35
Q	0.12	-0.99	-0.99
E	0.59	-0.55	-0.99
G	-0.79	-0.99	0.10
H	0.08	-0.71	0.68
I	-0.77	0.67	-0.37
L	-0.92	0.31	-0.99
K	-0.63	0.25	0.50
M	-0.80	0.44	-0.71
F	0.87	0.65	-0.53
P	-0.99	-0.99	-0.99
S	0.99	0.40	0.37
T	0.42	0.21	0.97
W	-0.13	0.77	-0.90
Y	0.59	0.33	-0.99
V	-0.99	0.27	-0.52

Tablo 3.19’da verilen ifadeler doğrultusunda, $P(N) = [A, R, \dots]$ biçimindeki bir protein dizilimi, AESNN3 değerleri ile $M(N) = [[-0.99 - 0.61 \ 0.00], [0.28 - 0.99 - 0.22]]$ şeklinde sayısal hale getirilir.

3.5.2. ANN4D Protein Sayısallaştırma Yöntemi

ANN4D protein sayısallaştırma yöntemi, AESNN3 protein sayısallaştırma yöntemi gibi makine öğrenmesi-tabanlı bir yöntemdir. Ancak, AESNN3 yönteminin aksine, ANN4D yönteminde, protein bilgileri fizikokimyasal ve istatistiksel bilgilere dayanarak elde edilmiş ve sinir ağında geliştirilmiştir [21]. Eğitim aşamasında YSA (Yapay Sinir Ağı) kullanılmış ve eğitim 5 katmanlı ağ yapısı ile gerçekleştirilmiştir. AESNN3 yönteminde olduğu gibi, ağırlık değerleri amino asit kodlarına atanmış ve amino asitler bu değerlere göre sayısallaştırılmıştır. Her bir amino asit için dört adet değer olduğundan, 4-boyutlu bir sayısallaştırma yöntemidir. Tablo 3.20’de amino asitlerin ANN4D değerleri verilmiştir.

Tablo 3.20. Amino asitlerin ANN4D değeri

Amino Asit Kodu	Gizli Katman 1	Gizli Katman 2	Gizli Katman 3	Gizli Katman 4
A	0.56	1.00	0.86	0.19
G	0.05	0.70	0.59	0.17
V	0.48	0.77	0.30	0.99
L	0.70	0.78	0.89	0.49
I	0.94	0.83	0.35	1.00
F	0.69	0.50	0.81	0.58
Y	0.58	0.23	0.70	0.52
W	1.00	0.37	0.89	0.66
T	0.23	0.55	0.38	0.67
S	0.19	0.55	0.57	0.41
R	0.34	0.00	1.00	0.01
K	0.27	0.11	0.86	0.08
H	0.39	0.25	0.76	0.39
D	0.51	0.41	0.11	0.00
E	0.90	0.87	0.32	0.01
N	0.19	0.18	0.68	0.18
Q	0.46	0.49	0.84	0.21
M	0.66	0.67	0.89	0.42
P	0.00	0.82	0.00	0.99
C	9.09	0.54	0.65	0.74

Tablo 3.20’de verilen değerlere göre, $P(N) = [A, G, \dots]$ biçimindeki bir protein dizilimi, ANN4D sayısallaştırma yöntemi ile $M(N) = [[0.56 \ 1.00 \ 0.86 \ 0.19], [0.05 \ 0.70 \ 0.59 \ 0.17], \dots]$ şeklinde sayısal ifadelere dönüştürülür.

3.5.3. ProtVec Protein Sayısallaştırma Yöntemi

ProtVec, proteinlerden özellik elde etmek için önerilmiş makine öğrenmesi tabanlı bir protein sayısallaştırma yöntemidir [22]. Daha çok biyolojik dizilimler için kullanılmaktadır. Bu yöntemde her bir dizilim n-boyutlu bir vektör olarak ifade edilmektedir. Daha çok proteinlerin biyofiziksel ve biyokimyasal özelliklerini temsil etmektedir. Doğal dil işleme uygulamalarında sıklıkla kullanılan Word2Vec yönteminden esinlenmiştir. Prot2Vec yönteminde ilk olarak dizilimdeki amino asit kodları 3’erli gruplar halinde ayrıştırılmıştır. Bu ayrıştırma işlemi Tablo 3.21’de verilmiştir.

Tablo 3.21. ProtVec yönteminin çalışma prensibi

Orijinal Protein Dizilimi	MAFSAEDVLKEYDRRRRMEAL...
Birinci Ayrıştırma	MAF, SAE, DVL, KEY, DRR, RRM, EAL, ...
İkinci Ayrıştırma	AFS, AED, VLK, EYD, RRR, RME, AL...
Üçüncü Ayrıştırma	FSA, EDV, LKE, YDR, RRR, MEA, L...

Her bir ayrıştırma işlemi toplamda 100-boyutlu vektör elde edilene kadar devam etmiştir. 100-boyutlu vektör elde edildikten sonra ise, yapay sinir ağı modeli ile vektörler eğitilmiş ve her bir 3’lü amino asit kodunun sayısal değerleri elde edilmiştir.

Tezin bu kısmında literatürde en çok kullanılan protein sayısallaştırma yöntemleri incelenmiş ve her birinin kategorisi hakkında bilgiler verilerek, bu yöntemlerin kullanım alanları ve hesaplamaları açıklanmıştır. Çalışmada değerlendirilen kategoriler, karakter-tabanlı, makine öğrenmesi-tabanlı, fizikokimyasal-tabanlı, yapı-tabanlı ve evrim-tabanlı kategorilerdir. Karakter-tabanlı sayısallaştırma işleminde protein dizilimlerinden belirli bilgiler elde edilmemektedir. Protein dizilimleri sadece amino asitlerin kodlarına göre sayısal hale getirilmektedir. Fizikokimyasal-tabanlı yöntemlerde ise, protein dizilimlerinden kimyasal bilgiler elde edilir ve bu bilgiler doğrultusunda dizilimler sayısal ifadelere dönüştürülür. Yapı-tabanlı yöntemlerde de tıpkı fizikokimyasal-tabanlı yöntemlerde olduğu gibi protein dizilimlerinden belirli bilgiler elde edilir ve sayısallaştırma işlemi gerçekleştirilir. Bu kategoride, protein dizilimlerinin yapısal bilgilerine ihtiyaç duyulmaktadır. Evrim-tabanlı kategoriye ait protein sayısallaştırma yöntemleri ise proteinlerin evrim bilgilerine ihtiyaç duymaktadır. Son olarak makine öğrenmesi-tabanlı yöntemlerde ise protein dizilimleri makine öğrenmesi algoritmalarına dayanarak sayısal hale

dönüştürülmektedir. Yöntemlerin türüne protein dizilimlerinden çeşitli kimyasal ya da yapısal bilgiler elde edilebilmektedir.

Tez kapsamında üç farklı protein sayısallaştırma yöntemi önerilmiş ve her biri ayrı bir kategoride değerlendirilmiştir. Çalışmada gerçekleştirilen ilk protein sayısallaştırma yöntemi AVL-ağacı yöntemidir. Geliştirilmiş olan bu yöntem algoritmik-tabanlı bir yöntemdir. Bu yöntemin literatürde bulunan diğer yöntemlerden farkı, sayısallaştırma işleminin algoritmik bir yapıya dayandırılmasıdır. Bu yöntemle ilgili detaylı bilgiler tezin 5. bölümünde verilmiştir. İkinci önerilen yöntem ise Shannon entropisine dayalı protein sayısallaştırma yöntemidir. Bu yöntem ile viral genomlar arasındaki protein etkileşimleri saptanmıştır. Karakter-tabanlı yöntemlerde olduğu gibi belirli bir bilgiye dayandırılmadan sayısallaştırma işlemi gerçekleştirmektedir. Bunun yanı sıra bu yöntemin en büyük avantajı, diğer yöntemlerin aksine dinamik yani adaptif bir yapıya sahip olmasıdır. Bu yöntemle ilgili detaylı bilgiler tezin 6. bölümünde verilmiştir. Çalışma kapsamında önerilen son yöntem ise FIBHASH yöntemidir. Bu yöntemde Fibonacci sayıları ve hash tablosu kullanılmıştır. Tıpkı önerilen diğer yöntemler gibi bu yöntemde de belirli bir protein bilgisi gerekmemektedir. Bu yöntemle ilgili detaylı bilgiler tezin 7. bölümünde verilmiştir.

4. PROTEİN DİZİLİMLERİ İLE GERÇEKLEŞTİRİLEN ÇALIŞMALAR

Bu bölümde, protein dizilimleri ile gerçekleştirilen çalışmalar incelenmiş ve literatür hakkında bilgiler verilmiştir. İncelenen literatür doğrultusunda aşağıda verilen çalışmalar üzerinde durulmuştur;

- Protein-protein etkileşimlerinin tahmin edilmesi,
- Protein ailelerinin sınıflandırılması,
- Protein fonksiyonlarının belirlenmesi,
- İlaç-hedef etkileşimlerinin tahmin edilmesi,
- Kodlayan ve kodlamayan RNA dizilimlerinin saptanması,
- Kanser içeren protein dizilimlerin belirlenmesi.

Bahsi geçen çalışmalar incelenmiş ve farklı başlıklar altında verilmiştir. Bu çalışmaların incelenmesiyle literatür hakkında bir alt yapı oluşturulmuş ve çalışmalarda karşılaşılan problemler ve onlara yönelik önerilen çözümler üzerinde durulmuştur.

4.1. Protein-Protein Etkileşimlerinin Tahmin Edilmesi

Organizmadaki hücresel süreçlerin birçoğu, proteinler tarafından yerine getirilmekte ve kontrol edilmektedir. Tüm bu süreçler, proteinlerin diğer proteinler, nükleik asitler ve hücre bileşenleri ile gerçekleştirdikleri etkileşimlere dayanmaktadır. Bu nedenle proteinlerin önemi büyüktür. Proteinler arasındaki etkileşimleri belirleyebilmek için genellikle in-vitro (laboratuvar ortamı ya da yapay koşullar), in-vivo (ölü bir organizma ya da organizmadan parça) ve in-silico (bilgisayarda gerçekleşen, hesaplama-tabanlı) olmak üzere üç farklı yöntem kullanılmaktadır. In-vitro yöntemlerinde, çalışmalar laboratuvar ortamında ya da yapay koşullar altında gerçekleştirilir [23]. In-vitro yöntemlerinde, test tüpleri, mikro titre plakalar ve petri kapları kullanılmaktadır [24]. In-vitro ile yapılan çalışmalar daha detaylı ve daha uygun olmaktadır. Bu durumun en büyük nedeni, bu yöntemde yapay koşullar ya da laboratuvar ortamı bulunduğu için, biyolojik bir ortamdan izole edilmiş bir organizma kullanılmasıdır. Ancak, in-vitro ile yapılan çalışmaların sonucunda, organizmanın tamamındaki etkileşimler bazı durumlarda net ve tam olarak belirlenemez [25-28]. Bu yöntemde görülen başka bir dezavantaj ise, organizmanın sahip olduğu hassas hücresel durumlar oluşturulamamaktadır [29]. In-vitro tabanlı yöntemler kullanılırken çeşitli yaklaşımlar bulunmaktadır. Bu yaklaşımlar aşağıda verilmiştir;

- **Mikro dizi:** Protein mikro dizileri ile, proteinler ve fonksiyonları tespit edilmekte, proteinlerin ekspresyon seviyeleri takip edilmekte ve proteinler arasındaki etkileşimler belirlenebilmektedir. Mikro dizi yönteminde ilk olarak, hedef protein dizilimlerine floresan

işaretleyiciler eklenir. Ardından, bu dizilimler çiplerde bulunan problemler ile reaksiyona girer. Reaksiyon kuran dizilimler belirlenir ve çipin problemlerine aktarılır. Son olarak etkileşim kuran dizilimler lazer ışını teknolojisiyle belirlenir. Bu sayede çok sayıda dizilim hızlı ve eş zamanlı bir şekilde tanımlanır [30, 31].

- **X-ışını:** X-ışını teknolojisiyle, proteinler arasındaki etkileşimler ve enzimlerin durumlarında görülen değişimler gözlemlenebilmektedir. Bu yöntem ile, protein yapıları görselleştirilir ve etkileşimler kristal yapı aracılığıyla belirlenir. Görsele dayalı bir yöntem olduğu için proteinler arasındaki etkileşimleri belirlemek zor olmaktadır. Bundan dolayı hata oranı yüksek olan bir yöntemdir [32-34].
- **NMR (Nükleer Manyetik Rezonans):** Bu yöntemin temeli manyetik alana dayanmaktadır. Proteinlerin 3-boyutlu ve 4-boyutlu yapıları analiz edilebilmektedir. Protein yapıları belirlenerek, proteinler arasındaki etkileşimler saptanabilmektedir. Bu yöntem ile yüksek doğruluk oranıyla protein etkileşimleri tahmin edilebilmektedir [35-37].
- **Faj ekranı:** Faj ekranı yönteminde, faj partikülleri üzerinde bulunan polipeptitler değerlendirilerek fenotip ve genotip arasında bir bağlantı kurulur. Faj ekranı yönteminin en önemli özelliği kütüphane üretebilmesidir. Bu kütüphaneler sayesinde, bir amino asit dizisi basit bir şekilde belirlenebilir [38].
- **Parça tamamlama:** Proteinler arasındaki etkileşimleri belirlemek için sıklıkla kullanılan yöntemlerin başında gelmektedir. Bu yöntem içerisinde kütle spektroskopisi kullanılarak proteinler belirlenmektedir. Ancak bazı durumlarda yöntem kararsız olmakta ve yanlış pozitif değerler üretebilmektedir [39-41].
- **Tandem afinite arıtma kütle spektroskopisi:** Proteinlerin etiketlenmesine dayanan bir yöntemdir. Proteinler etiketlendikten sonra ise arıtma ve kütle analizi gerçekleştirilir [42].
- **Afinite kromatografi:** Afinite kromatografi yöntemi tersinir bir yöntemdir. Safılaştırma ve yakalama olmak üzere iki adımdan meydana gelmektedir. İşlemin sonunda yüksek oranda yanlış pozitif değerler elde edilebilmektedir [29]. Bu yöntemin en önemli özelliği zayıf protein etkileşimlerini bile belirleyebilmesidir.
- **Eş-immüno çökeltme:** Bu yöntemde yem ve av olmak üzere iki türlü protein bulunmaktadır. Hedef proteinler yem olarak ifade edilirken, etkileştiği protein ise av olarak tanımlanmaktadır. Yöntem kapsamında proteinler parçalanır ve birbiriyle etkileşim halinde olan proteinler belirlenir [43].

Genel olarak incelendiğinde, in-vitro yöntemlerinde yanlış pozitif değerler üretilmektedir. Bundan dolayı, in-vitro ile elde edilen sonuçların, in-vivo yöntemler ile doğrulanması gerekmektedir.

In-vivo yöntemlerinde ise, proteinler arasındaki etkileşimler, canlı organizmalar ya da hücreler kullanılarak belirlenir [23]. Genellikle hayvanlar üzerinde gerçekleştirilir. In-vivo tabanlı yöntemler çoğu zaman in-vitro tabanlı yöntemlerden sonra uygulanmaktadır. Bunun en önemli

nedeni ise, in-vitro yöntemlerinde elde edilen sonuçların, canlı hücreler ya da organizmalar üzerinde test edilmesidir. In-vitro yöntemlerinde bazı durumlarda yanlış sonuçlar ya da yüksek yanlış pozitif değerler elde edildiği için, bu yöntemlerin in-vivo yöntemler ile doğrulanması gerekmektedir. Ancak, bazı durumlarda in-vivo yöntemler de yanıltıcı sonuçlar üretebilmektedir. In-vivo da ise en çok kullanılan yöntemler şunlardır;

- **Afinite etiketi:** Daha çok proteinlerin yapılarını ve fonksiyonlarını belirlemek için kullanılan bir yöntemdir. Afinite etiketi yönteminde, ilk olarak bir yem proteini, plazmid ile transfeksiyona uğrar. Ardından, hücreler parçalanır ve proteinler ile yem protein izole edilir. Son adımda ise, proteinler elektroforez yardımıyla ayrıştırılır ve etkileşimler belirlenir [44, 45].
- **Kimyasal çapraz bağlama:** Yapısal tabanlı bir yöntemdir. Bu yöntemde, proteinin fonksiyonları kovalent olarak bağlanır ve yapısal bir etkileşimin olması sağlanır [46, 47].
- **Y2H (Yeast Two Hybrid – Maya İki Hibrit):** Proteinler arasındaki etkileşimleri belirlemek için kullanılan en popüler yöntemdir. Bu yöntemin en önemli avantajlarından birisi, proteinler arasındaki etkileşimler doğrudan belirlenebilmektedir. Bunun yanı sıra, çok sayıda yanlış pozitif değerler de elde edilebilmektedir [48].
- **Sentetik latelit:** Bu yöntemde proteinler arasındaki etkileşimler fiziksel koşullardan ziyade, fonksiyonel koşullara dayanmaktadır. Proteinler arasındaki etkileşimler gende mutasyonlar oluşturularak belirlenir [49].

In-vivo ve in-vitro yöntemler ile gerçekleştirilen deneysel uygulamalarda çok sayıda veri oluşmakta ve bu verileri işlemek için hem zaman hem de laboratuvar ihtiyacı gerekmektedir [50]. Ayrıca, elde edilen protein etkileşimleri, kullanılan deneysel yöntemlere (in-vivo ve in-vitro) göre değişiklik göstermektedir. Bunlara ek olarak, deneysel yaklaşımlar hem çevresel hem de operasyonel süreçlere duyarlı oldukları için yanlış-pozitif ve yanlış-negatif değerler üretilmektedir [51]. Bu nedenlerden dolayı, in-silico tabanlı yaklaşımların önemi artmıştır [52,53].

In-silico yöntemlerinde ise, protein etkileşimleri bilgisayarlar ya da simülasyonlar aracılığıyla belirlenmektedir [54]. In-silico yöntemler ile sadece proteinler arasındaki etkileşimler belirlenmemektedir. Bunun yanında, proteinlerin tasarımı, biyoproses optimizasyonu ve organizmaların davranışlarının modellenmesi gibi amaçlar için de kullanılmaktadır [55, 56]. En çok kullanılan in-silico yaklaşımlar şunlardır;

- **Dizilim-tabanlı yaklaşımlar:** Teknolojinin gelişmesiyle, protein dizilimleri daha kolay elde edilebilmekte ve proteinler arasındaki etkileşimleri belirleyebilmek için gerekli olan veriler toplanabilmektedir. Protein dizilimlerinin sayısallaştırılmasına dayanan bir yaklaşımdır. Bu yaklaşımlarda protein yapı bilgilerini, proteinlerin fizikokimyasal bilgilerini, gen ontoloji bilgilerini kullanan algoritmalar bulunmaktadır [57, 58].

- **Yapı-tabanlı yaklaşımlar:** Yapı-tabanlı yaklaşımlarda, proteinler arasındaki etkileşimler belirlenirken, proteinlerin yapılarına bakılmaktadır. Eğer proteinler benzer yapıya sahipse, etkileşim halinde oldukları proteinler de benzer yapıya sahiptir düşüncesine dayanmaktadır. İlk aşamada, proteinlerin yapıları belirlenir. Ardından, etkileşim halinde oldukları bilinen proteinler ile bir küme oluşturulur. Daha sonra, hedef proteinin yapı bakımında benzer olabileceği proteinler belirlenir. Belirlenen proteinlerin etkileşim halinde oldukları proteinler bulunur ve hedef proteinin bu proteinlerle etkileşim halinde olduğu kabul edilir [59, 60].
- **Gen ifadesi:** Bu yöntemde genler ve kodladıkları proteinler arasındaki etkileşimlere bakılmaktadır. Gen ifadesi, belirli bir genin bir hücrede, dokuda veya organizmada çeşitli deneysel koşullarda ve zaman aralıklarında ifade edilme derecesidir. Kapsamlı gen ekspresyonu ve proteinler arasındaki etkileşim verilerinin istatistiksel yöntemler ile değerlendirilmesiyle, genler tarafından kodlanan protein çiftlerinin rastgele proteinlerden daha sık etkileşime girdiği gözlemlenmiştir [61, 62].
- **Gen birleşimi:** Füzyon geni, çeşitli nedenlerle iki ayrı genden veya bunların bölümlerinden oluşan hibrit bir gendir. Organizmada bulunan benzersiz proteinler, başka organizmalarda bulunan proteinlerle etkileşim kurarak, protein kompleksleri oluşturur. Bu metot ile, proteinler arasındaki etkileşimler, farklı genomlardan gelen düzenleyici bilgiler kullanılarak tahmin edilir [63, 64].
- **Gen komşuluğu:** Çoklu genomlar arasında yeterli yakınlık sağlandığında, genler tarafından kodlanan proteinler arasında fonksiyonel anlamda yüksek bir bağlanma görülür. Bu yöntemde, bilinmeyen protein etkileşimlerini belirleyebilmek için, gen yakınlığı bilinen proteinler arasındaki etkileşimlere bakılır [65, 66].
- **Filogenetik ağaçlar:** Filogenetik ağaçlar, proteinlerin evrimi hakkında bilgiler sağlar. Etkileşim halinde olan proteinler, birlikte evrimleştiği için benzer filogenetik ağaçlar oluşturur. Bu yöntemde, etkileşimleri tahmin edilecek olan proteinlerin benzerlik matrisleri ve filogenetik ağaçları üretilir. Ardından benzer filogenetik ağaç yapılarına bakılır ve etkileşim halinde olan proteinler belirlenir [35, 67, 68].

In-vivo, in-vitro ve in-silico tabanlı yöntemler genel olarak karşılaştırıldığında, in-silico tabanlı yöntemlerin hem ekipman hem bütçe hem de laboratuvar ihtiyacı bakımından daha az maliyetli olduğu gözlemlenmiştir. Bu nedenle, araştırmacılar proteinler arasındaki etkileşimleri belirlemek için daha çok in-silico tabanlı yöntemlere yönelmişlerdir. Proteinler arasındaki etkileşimleri belirleyebilmek için makine öğrenmesi algoritmaları sıklıkla kullanılmaktadır. [69] numaralı çalışmada araştırmacılar, radyal-tabanlı fonksiyonel yapay sinir ağı kullanarak proteinler arasındaki etkileşim bölgelerini tahmin etmişlerdir. Çalışmanın ilk aşamasında protein dizilimleri sayısallaştırılmış ve her bir amino asidin frekans değerleri elde edilmiştir. Ardından, bağıl entropi

değerleri hesaplanmış ve toplamda 1.000 adet öznitelik elde edilmiştir. Son aşamada ise sınıflandırma yapılmış ve sınıflandırıcının performansı F1-skor ve doğruluk değerlendirme ölçütleri ile belirlenmiştir. Çalışmanın sonunda %99 oranında F1-skor elde edilirken, %80 oranında doğruluk değeri hesaplanmıştır. [70] numaralı başka bir çalışmada ise proteinler arasındaki etkileşimler protein imzaları yöntemi ile belirlenmiştir. Çalışmada insan ve farede bulunan *helicobakter pilori* bakterisine ait veriler kullanılmıştır. Protein dizilimleri protein imzası yöntemi ile sayısallaştırılmıştır. Sınıflandırma işlemi için ise DVM (Destek Vektör Makineleri) kullanılmıştır. Geliştirilen yöntem üç farklı tür üzerinde test edilmiştir. Bu türler; *helicobakter pilori*, *koli basili* ve *ekmek mayası* türleridir. Önerilen yöntemin başarımı doğruluk, duyarlılık ve özgünlük değerlendirme ölçütleri ile belirlenmiştir. Protein dizilimlerinin çok sayıda olduğu durumlarda, makine öğrenmesi algoritmaları etkili olamamakta ve öznitelikler elde edilememektedir [50]. Bu nedenle, makine öğrenmesi algoritmalarının yanında derin öğrenme algoritmaları da bu alanda etkili bir şekilde kullanılmaktadır. [71] numaralı çalışmada araştırmacılar, proteinler arasındaki etkileşimleri ESA (Evrişimsel Sinir Ağı) ve LSTM (Long/Short Term Memory- Uzun/Kısa Vadeli Hafız Ağları) ile tahmin etmişlerdir. Çalışmada, proteinler arasındaki motifler, anlamsal ve uzun-kısa vadeli ilişkiler belirlenmiş ve bu ilişkiler öznitelik olarak değerlendirilmiştir. Kullanılan sınıflandırıcıların başarımları 5 katlı çapraz-doğrulama işlemi ile test edilmiş ve sınıflandırıcıların performansları sadece doğruluk skoru ile değerlendirilmiştir. ESA mimarisi ile %98,78 oranında doğruluk skoru elde edilmiştir. [72] numaralı çalışmada ise, proteinler arasındaki etkileşimler SSA (Siyam Sinir Ağı) ile tahmin edilmiş ve önerilen sınıflandırıcının performansı dört farklı veri seti üzerinde test edilmiştir. Çalışmada, protein dizilimleri hem protein imzaları hem de Prot2Vec yöntemi ile sayısallaştırılmıştır. Kullanılan SSA modelinin sınıflandırma performansı AUC (Area Under Curve – Eğri Altında Kalan Alan) skoru ile belirlenmiştir. Dört veri seti ile yapılan doğrulama işleminin sonunda, ortalama %83,25 oranında AUC skoru elde edilmiştir. [73] numaralı çalışmada araştırmacılar gliom evrelerini makine öğrenmesi ve protein-protein arasındaki etkileşim ağlarını kullanarak tahmin etmişlerdir. Çalışmada 527 adet gen ifadesi kullanılmış ve bu veriler insanların beyin dokularından elde edilmiştir. Elde edilen gen ifadeleri PseInf vektör yöntemi ile sayısallaştırılmış ve 1-boyutlu öznitelik vektörleri oluşturulmuştur. Ardından, öznitelik vektörleri sınıflandırılmak üzere SNB (Saf Naive Bayes), DVM, KNN (K En Yakın Komşu), RO (Rassal Orman) ve YSA algoritmalarında kullanılmıştır. Sınıflandırıcıların başarımları 10 katlı çapraz-doğrulama ile test edilmiş ve performansları özgünlük, duyarlılık ve doğruluk skorları ile belirlenmiştir. En başarılı sınıflandırma işlemi %97,1 doğruluk oranı ile RO algoritmasıyla elde edilmiştir. [74] numaralı çalışmada araştırmacılar, hibrit öznitelikleri kullanarak proteinler arasındaki etkileşimleri tahmin etmişlerdir. Çalışmada proteinlerin 12 adet fizikokimyasal özellikleri kullanılarak sayısallaştırma yapılmıştır. Bu öznitelikler; hidrofobiklik, esneklik, ulaşılabilirlik, dönüş ölçeği, yüzey alanı, polarite,

hidrofilik, antijenik eğilim, yan zincirlerin net yük endeksi, polarize edilebilirlik, solvent erişilebilir yüzey alanı ve yan zincir hacmidir. Bu 12 adet öznelikten, hidrofobiklik ve polarize edilebilirlik değerleri iki kere hesaplanmış ve bir protein dizilimi için 14 adet öznelik elde edilmiştir. Ardından bu öznelikler NB sınıflandırıcısı ile sınıflandırılmış ve proteinler arasındaki etkileşimler tahmin edilmiştir. Doğrulama işlemi için 10 katlı çapraz-doğrulama kullanılmış ve önerilen sınıflandırıcı ile %93,20 oranında doğruluk skoru hesaplanmıştır. [51] numaralı çalışmada ise, araştırmacılar proteinlerin fiziksel ve kimyasal özelliklerini kullanarak proteinler arasındaki etkileşimleri tahmin etmişlerdir. Çalışmada 10.000 adet pozitif ve 10.000 adet negatif olmak üzere, toplamda 20.000 adet protein etkileşimi değerlendirilmiştir. Proteinleri sınıflandırmak için YKE (Yüksek Kalite Endeksi) kullanılmış ve her bir protein için toplam 40 adet öznelik elde edilmiştir. Sınıflandırma sürecinde DVM algoritması kullanılmış ve sınıflandırıcının başarımı 5 katlı çapraz-doğrulama yöntemi ile belirlenmiştir. Sınıflandırıcının performansını hesaplamak için ise özgünlük, duyarlılık ve doğruluk skorları kullanılmıştır. Çalışmanın sonunda ortalama %75 oranında doğruluk gözlemlenmiştir. [75] numaralı çalışmada araştırmacılar protein-protein etkileşimlerini derin öğrenme kullanarak tahmin etmişlerdir. Çalışmada etkileşimler *maya* ve *helikobakter pilori* türleri üzerinden değerlendirilmiştir. Protein dizilimleri PSSM yöntemi ile sayısallaştırılmıştır. Sayısallaştırma işleminin ardından ESA modeli kullanılmış ve bu model aracılığıyla öznelikler elde edilmiştir. Sınıflandırma işlemi için RotOR (Rotasyon Ormanı) algoritması kullanılmış ve proteinler arasındaki etkileşimler tahmin edilmiştir. Sınıflandırıcının başarımı 5 katlı çapraz-doğrulama ile ölçülmüştür. Bunlara ek olarak sınıflandırıcının performansı ise doğruluk, özgünlük, duyarlılık, F-skor ve AUC skoru ile belirlenmiştir. Proteinler arasındaki etkileşimler ortalama %97,75 doğruluk ve %97,54 AUC skorları ile tahmin edilmiştir. [76] numaralı çalışmada araştırmacılar dizilim tabanlı yaklaşımlar kullanarak, proteinler arasındaki etkileşimleri tahmin etmişlerdir. Çalışmada toplamda 2.184 adet benzersiz protein dizilimleri kullanılmıştır. Protein dizilimleri OK (Oto Kovaryans) ve BT (Birleşik Triad) yöntemleri ile sayısallaştırılmıştır. Sınıflandırma işlemi için ise oto kodlayıcı kullanılmış ve sınıflandırıcının başarımı 10 katlı çapraz-doğrulama ile belirlenmiştir. Gerçekleştirilen çalışmanın ana amacı etkili protein sayısallaştırma yöntemini belirlemek olduğu için, bu iki yöntemin sınıflandırma performansları karşılaştırılmıştır. Çalışmanın sonunda OK sayısallaştırma yönteminin daha etkili olduğu gözlemlenmiştir.

4.2. Protein Ailelerinin Sınıflandırılması

Protein ailelerinin sınıflandırılması, biyomoleküler tanıma, biyoenformatik, protein kalıntılarının belirlenmesi ve akıllı ilaç tedavisi çalışmaları için önemli bir araştırma alanıdır. Protein aileleri, analog yapıları içeren çok sayıda proteinlerden oluşmaktadır. Sınıflandırma işlemi sırasında, benzer fonksiyonlara sahip olan protein aileleri bir grup olarak ifade edilmektedir. Ancak, yapısal bilgilere dayalı olarak bir protein ailesini belirlemek, karakterize edilmemiş protein

dizilimlerinin elde edilmesinin zor olması nedeniyle, sağlıklı olamamaktadır [77, 78]. Bunun en büyük nedeni, yapı bilgilerine dayalı gerçekleştirilen çalışmalarda, protein fonksiyonları bilgisi net olmayabilmektedir. Bundan dolayı, araştırmacılar alternatif yollar aramışlardır. Bu alternatif yol doğrultusunda kullanılan en popüler yaklaşım ise protein dizilimleriyle gerçekleştirilen çalışmalardır [79-81]. Bunlara ek olarak, protein fonksiyonlarını belirlemek için 3-boyutlu protein yapılarının kullanılması zaman ve maliyet açısından verimli olmamakta ve X-ışını, NMR gibi bazı karmaşık yöntemlerin kullanılmasını gerektirmektedir [82]. Bu dezavantajlardan dolayı protein dizilimlerine yönelik yaklaşımların önemi artmıştır.

Protein dizilimlerini analiz etmek ve protein dizilimlerinden anlamlı, ayrık ve sürekli öznitelikler toplamak için, öznitelik mühendisliğine başvurulmaktadır. Gerekli öznitelikler elde edildikten sonra ise, derin öğrenme ve makine öğrenmesi algoritmaları kullanılmakta ve protein aileleri sınıflandırılmaktadır. Ancak, bu sınıflandırma algoritmalarının kullanılabilmesi için, protein dizilimlerinin sayısallaştırılması gerekmektedir. Bu nedenle bu çalışmalarda protein sayısallaştırma yöntemleri sıklıkla kullanılmaktadır. [83] numaralı çalışmada araştırmacılar, protein ailelerini belirleyebilmek için derin öğrenme modellerine başvurmuşlardır. Çalışmada dört farklı derin öğrenme modeli kullanılmış ve modellerin başarımları çapraz-doğrulama ile ölçülmüştür. Protein dizilimleri n-gram ve keras vektörü yöntemleri ile sayısallaştırılmıştır. Çalışmada toplamda 84.753 adet protein dizilimi kullanılmıştır. Protein ailelerini tahmin etmek için DSA (Derin Sinir Ağları), LSTM, RNN (Recurrent Neural Network – Yinelemeli Sinir Ağları) ve ESA mimarileri kullanılmıştır. En iyi sınıflandırma LSTM derin öğrenme modeli ile elde edilmiş ve %91,24 oranında doğruluk skoru hesaplanmıştır. [84] numaralı çalışmada ise araştırmacılar, protein ailelerini sınıflandırmak için sinir ağlarını kullanmışlardır. Çalışmada, 10.345 adet protein ailesine ait 550.960 adet protein dizilimi analiz edilmiştir. Protein dizilimleri GloVe yöntemi ile sayısallaştırılmıştır. Sınıflandırma işlemi için GRU (Gated Recurrent Unit – Geçitli Tekrarlayan Birim), LSTM, BiLSTM (Çift Yönlü LSTM) ve ESA kullanılmıştır. Sınıflandırıcıların performansları F1-skor değeri ile ölçülmüş ve en yüksek skor GRU ile elde edilmiştir. Bu yöntem ile %94,84 oranında doğruluk gözlemlenmiştir. Bunun yanı sıra, araştırmacılar bu çalışmada, DVM yöntemi ile de sınıflandırma yapmışlar ve bu yöntemin sonucunu diğer derin öğrenme yöntemleri ile karşılaştırmışlardır. Derin öğrenme modellerinin DVM sınıflandırıcısından daha etkili olduğu belirtilmiştir. [82] numaralı çalışmada araştırmacılar RNN, LSTM ve GRU derin öğrenme modellerini kullanarak protein ailelerini sınıflandırmışlardır. Çalışmada 30 farklı protein ailesine ait veriler değerlendirilmiş ve bu doğrultuda 40.433 adet protein dizilimi analiz edilmiştir. Bu çalışmada da protein dizilimleri n-gram ve keras vektörü ile sayısallaştırılmıştır. Çalışmanın sonunda en başarılı yöntem LSTM derin öğrenme modeli olmuş ve bu model ile %78,4 oranında doğruluk skoru elde edilmiştir. Bunun yanı sıra, çalışmada KNN, NB, karar ağacı, DVM ve LR sınıflandırıcıları da kullanılmıştır. Sonuç olarak, derin öğrenme modellerinin, protein ailelerini

sınıflandırma sürecinde daha etkili oldukları gözlemlenmiştir. [85] numaralı çalışmada araştırmacılar GEA (Graf Evrişimsel Ağ) mimarisini kullanarak protein ailelilerini sınıflandırmışlardır. Çalışmada 22.725 adet protein dizilimi kullanılmıştır. Sınıflandırıcının performansı sadece doğruluk değerlendirme ölçütü ile hesaplanmış ve %74,33 oranında bir başarımla elde edilmiştir. Bunlara ek olarak, t-DSKG (t-Dağıtılmış Stokastik Komşu Gömme) yöntemi ile sınıflandırıcının performansı ölçülmüştür. [86] numaralı çalışmada araştırmacılar yeni bir sayısal haritama modeli oluşturmuşlar ve protein ailelerini modelleyip sınıflandırmışlardır. Çalışmada 4.655 protein ailesine ait 1,652.498 adet protein dizilimi kullanılmıştır. Protein dizilimleri önerilen yöntem, k-mer ve ProtVec yöntemleri ile sayısallaştırılmıştır. Sayısallaştırma işleminin ardından ESA derin öğrenme modeli tasarlanmış ve protein aileleri sınıflandırılmıştır. Araştırmada, ESA mimarisinin yanı sıra NB ve KNN olmak üzere 2 farklı makine öğrenmesi algoritmalarına da başvurulmuştur. Sınıflandırıcının performansları yalnızca doğruluk değerlendirme ölçütü ile belirlenmiştir. Çalışmanın sonunda, NB sınıflandırıcısı ile %77,29 doğruluk değeri elde edilirken, bu değer KNN sınıflandırıcısı ile %95,87'e kadar artış göstermiştir. Ancak, en etkili performans ESA tabanlı sınıflandırıcı ile elde edilmiştir. Önerilen derin öğrenme modeliyle %97,17 oranında doğruluk gözlemlenmiştir. [87] numaralı çalışmada ise araştırmacılar ESA mimarisi geliştirerek, protein ailelerini sınıflandırmışlardır. Çalışmada 60 protein ailesine ait protein dizilimleri değerlendirilmiştir. Bu doğrultuda, protein dizilimleri ilk olarak ikili-kodlama protein sayısallaştırma yöntemi ile sayısallaştırılmış ve sınıflandırıcılarda kullanılmak üzere hazır hale getirilmiştir. Çalışmada geliştirilen ESA mimarisine ek olarak LSTM, BiLSTM, GRU ve DVM sınıflandırıcıları da kullanılmıştır. Sınıflandırıcıların performansları doğruluk, kesinlik, duyarlılık ve F1-skorumları ile belirlenmiştir. Uygulama sonuçlarına göre DVM sınıflandırıcısı hariç, bütün sınıflandırıcılar %90'un üzerinde doğruluk sergilemiştir. DVM yöntemi ile %87,88 doğruluk skoru hesaplanmıştır. Ardından en düşük sınıflandırma performansı BiLSTM ve LSTM mimarileriyle elde edilmiş ve sırasıyla %92,22 ve %92,51 doğruluk skorumları hesaplanmıştır. GRU sınıflandırıcısıyla %94,84 oranında doğruluk skoru elde edilirken, tasarlanan ESA mimarisiyle ise %97,77 oranında doğruluk gözlemlenmiştir. En iyi sınıflandırma başarısına geliştirilen yöntem ile ulaşılmıştır. [88] numaralı çalışmada, araştırmacılar ESA mimarisini kullanarak protein ailelerini analiz etmişlerdir. Çalışmada 5 farklı protein ailesine ait toplamda 7.938 adet protein dizilimi ele alınmıştır. Proteinler Word2Vec yöntemi ile sayısallaştırılmış ve ESA mimarisinde sınıflandırma amacıyla kullanılmıştır. Sınıflandırıcının başarımla 10 katlı çapraz-doğrulama ile belirlenirken, performansı doğruluk, Matthews korelasyon katsayısı, kesinlik, duyarlılık ve F1-skor değerlendirme ölçütleri ile saptanmıştır. Çalışmanın sonunda %98,89 oranında doğruluk skoru hesaplanmıştır. [89] numaralı çalışmada ise, araştırmacılar DSA kullanarak hem protein etkileşimlerini tahmin etmişler hem de protein ailelerini sınıflandırmışlardır. Çalışma kapsamında 5 protein ailesine ait protein verileri ele alınmıştır. Protein dizilimlerini sayısallaştırmak için 4 farklı

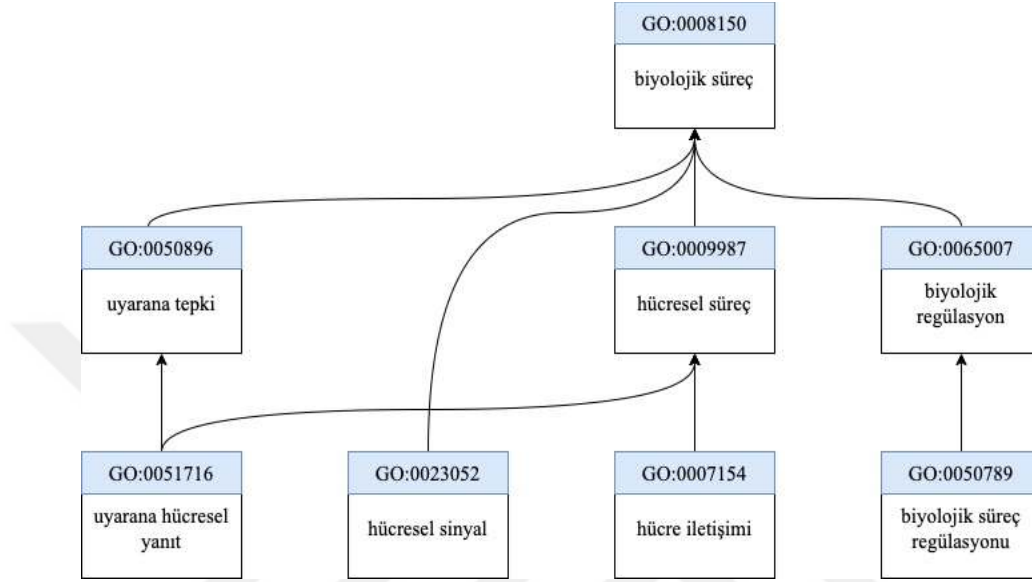
yöntem kullanılmıştır. Bu yöntemler; AAB (Amino Asit Bileşimi), GAAB (Gruplandırılmış Amino Asit Bileşimi), BT ve YDS (Yarı Dizilim Sıralaması) yöntemleridir. Sayısallaştırma işleminin ardından öznitelikler elde edilmiş ve bu öznitelikler sınıflandırılarak, protein aileleri kategorize edilmiştir. Kullanılan derin öğrenme modelinin başarımı 5 katlı çapraz-doğrulama ile test edilmiştir. Sınıflandırıcının performansı ise, kesinlik, doğruluk, duyarlılık ve AUC ile ölçülmüştür. Çalışmanın sonunda %100 oranında kesinlik değeri, %98,15 oranında duyarlılık değeri, 0,99 oranında AUC skoru ve %99,08 oranında doğruluk değeri gözlemlenmiştir. [90] numaralı çalışmada araştırmacılar protein ailelerini hiyerarşik bir şekilde sınıflandırmak için derin öğrenme modeli kullanmışlardır. Çalışmada toplamda 4.000 protein ailesi değerlendirilmiştir. Proteinlerin sayısallaştırılma adımında ise 2-boyutlu matris yapısı kullanılmıştır. Bu yapı doğrultusunda sayısallaştırılan proteinler, daha sonra ESA mimarisi ile sınıflandırılmıştır. Sınıflandırıcının performansı doğruluk, kesinlik, duyarlılık, F1-skor ve AUC değerlendirme ölçütleri ile belirlenmiştir. Çalışmanın sonunda, %95,53 oranında doğruluk, %97,24 oranında kesinlik, %94,05 oranında duyarlılık, 0,99 oranında AUC ve %95,62 oranında ise F1-skor elde edilmiştir.

4.3. Protein Fonksiyonlarının Belirlenmesi

Doğada görünen 20 farklı türdeki amino asit dizilimleri DNA tarafından oluşturulmaktadır. Amino asitler, hücreler arasındaki iletişimin sağlanması, hücrelerin düzenlenmesi, reaksiyonların hızlandırılması olmak üzere organizma üzerinde çeşitli rollere sahiptirler. Proteinlerin fonksiyonları belirlenirken, proteinlerin yapılarına bakılmaktadır. Protein fonksiyonlarının belirlenmesiyle, genlerde çeşitli hastalıklara yol açan değişiklikler saptanabilmekte, hastalıkların önüne geçilebilmekte ve bu sayede hastalıkların tanısı konabilmekte ve tedavisi için bir yol planı hazırlanabilmektedir [91]. Proteinlerin fonksiyonları belirlenirken çoğu zaman deneysel yaklaşımlar kullanılmaktadır. Fakat, deneysel çalışmalarda kimyasal işlemlerin zaman alması, laboratuvar ihtiyacının artması ve insan kaynağına ihtiyaç duyulması, deneysel yaklaşımların maliyetli olmasına sebebiyet vermektedir [92, 93]. Bunun yanı sıra, teknolojik gelişmelerin genetik alanında artış göstermesiyle, yeni protein dizilimleri elde edilmekte ve veri sayısında artış olduğu için analiz işlemleri yavaş olmaktadır [94]. Bu kısıtlamalardan dolayı, hesaplama-tabanlı analiz yöntemlerinin önemi artmış ve bu yöntemler protein fonksiyonlarının tahmini çalışmalarında sıklıkla kullanılmaya başlanmıştır.

Protein fonksiyonları üç farklı karmaşıklık seviyesinde değerlendirilmektedir. Bu seviyeler; biyokimyasal, fenotipik ve fizyolojik seviyelerdir. Bunun yanı sıra, hiyerarşik bir yapı kullanılarak da protein fonksiyonları analiz edilebilmektedir. Hesaplama-tabanlı yaklaşımlarda ise, karmaşıklık seviyelerinden ziyade, hiyerarşik bir yapı kullanılmaktadır. En çok kullanılan hiyerarşik yapı ise GO (Gen Ontolojisi) terimidir. GO terimleri, proteinlerin biyokimyasal veya moleküler fonksiyonlarını belirlemede etkili oldukları için, hesaplama-tabanlı yaklaşımlarda tercih

edilmektedir. GO terimleri, birbirine bağlı üç farklı seviyeden meydana gelmektedir. Bu seviyeler; BS (Biyolojik Süreç), MF (Moleküler Fonksiyon) ve HB (Hücresel Bileşen) seviyeleridir [95]. Şekil 4.1’de GO terimlerinin hiyerarşik yapısı verilmiştir.



Şekil 4.1. GO terimlerinin graf yapısı

Şekil 4.1’de verilen graf yapısına göre, her bir kare şekli hiyerarşik yapıyı ifade etmektedir. Bu şekillerin üst kısmında GO terimleri, alt kısmında ise o GO terimlerinin seviyeleri yer almaktadır. Örnek verilecek olursa, hiyerarşinin en üstünde GO:0008150 terimi bulunmaktadır. Bu terimin seviyesi ise biyolojik süreçtir. GO:0009987 terimi ise hiyerarşik olarak GO:0008150 teriminin altında yer almaktadır. Bu terimin ise seviyesi hücresel süreçtir.

Biyolojik süreç, moleküler seviyedeki fonksiyonların kümelenmesiyle ortaya çıkan geniş ölçekli fonksiyonlar olarak ifade edilmektedir. Hücresel bileşen ise, proteinlerin organizma ya da hücre içinde gerçekleştirdiği fonksiyon olarak tanımlanmaktadır. Son olarak ise, moleküler fonksiyon, moleküler seviyede görülen eylemlerdir. Teknolojik alanında görülen ilerlemeler, protein dizilimlerinin elde edilmesinde kullanılan yöntemlerin gelişmesine yol açarak günümüzde protein dizilimlerinin kolay bir şekilde elde edilmesine neden olmaktadır. Elde edilen dizilimlerinin kanıtlanması ve doğrulanması gerekmektedir. Bu amaç için ise deneysel yaklaşımlar kullanılmakta, ancak analiz işlemi uzun zaman almaktadır. Bu kısıtlamanın önüne geçebilmek için çeşitli sayıda algoritmalar geliştirilmiştir. Fakat geliştirilen algoritmaların çoğu manuel yaklaşımlara dayanmaktadır. Bundan dolayı bu yöntemlerden ziyade, hesaplama-tabanlı yaklaşımların kullanımı gerekmektedir.

Hesaplama-tabanlı yaklaşımlar ile protein fonksiyonlarının tahmin edilebilmesi için yapay zeka modellerine başvurulmaktadır. Bu amaç doğrultusunda, derin öğrenmeye ve makine öğrenmesine dayalı yöntemler kullanılmaktadır. Çalışmalar incelendiğinde makine öğrenmesi algoritmalarının daha sık kullanıldığı gözlemlenmiştir. Bu durumun en büyük nedeni, bu yöntemlerin, protein fonksiyonlarının tam olarak bilinmediği durumlarda bile etkili olmasıdır [96]. Ancak, bazı çalışmalar, makine öğrenmesinin yanlış hizalamalara yol açabildiğini kanıtlamıştır [97]. Bunun yanı sıra, makine öğrenmesi ile gerçekleştirilen çalışmalarda sınıflandırma performansı düşük olmaktadır [97]. Sınıflandırma performansının düşük olmasının en büyük nedeni, fonksiyonların çok sınıflı olmasıdır [98]. Makine öğrenmesi ile gerçekleştirilen çalışmalarda görülen problemleri çözmek için, araştırmacılar derin öğrenmeye başvurmuşlardır. Büyük verinin işlenmesinde derin öğrenmenin aktif ve etkili bir şekilde kullanılması, araştırmacıların derin öğrenme algoritmalarını bu alana uygulamalarına sebebiyet vermiştir [99, 100]. [101] numaralı çalışmada araştırmacılar derin öğrenme modeli kullanarak protein fonksiyonlarını sınıflandırmışlardır. Çalışmada maya ve insana ait olmak üzere iki türe ait protein fonksiyon verileri analiz edilmiştir. Çalışmanın ilk aşamasında, protein dizilimleri hem ProtVec hem de 3-gram protein sayısallaştırma yöntemleri olmak üzere iki farklı yöntem ile sayısallaştırılmıştır. Dizilimlerden öznitelikler elde edildikten sonra, ESA mimarisi tasarlanmış ve sınıflandırma işlemi yapılmıştır. Protein fonksiyonlarını sınıflandırma için GO terimleri kullanılmış ve BS, MF ve HB seviyeleri için ayrı ayrı tahmin işlemi gerçekleştirilmiştir. Sınıflandırıcının başarımı 5 katlı çapraz-doğrulama kullanılarak belirlenirken, performansı ise F1-skor ve AUC skoru ile saptanmıştır. Çalışmanın sonunda maya türü için en yüksek F1-skor MF seviyesinde elde edilmiş ve %61,1 oranında skor gözlemlenmiştir. Ardından en yüksek F1-skoru HB seviyesi için hesaplanmış ve bu seviyede skor %52,8 bulunmuştur. En başarısız sınıflandırma işlemi ise BS seviyesinde gözlemlenmiştir. Bu seviyede elde edilen F1-skor %41,5'tir. İnsana ait sınıflandırma işleminde ise en başarılı sınıflandırma MF seviyesinde elde edilmiş ve sonuç %65,3 olarak bulunmuştur. Ardından en başarılı tahmin işlemi HB seviyesinde gözlemlenmiş ve %61,7 oranında F1-skor hesaplanmıştır. En başarısız sınıflandırma ise BS seviyesinde görülmüştür. Bu seviyede hesaplanan F1-skor ise %50,7 olmuştur. [102] numaralı çalışmada ise araştırmacılar yinelemeli sinir ağı kullanarak protein fonksiyonlarını sınıflandırmışlardır. Çalışma kapsamında 42.918 GO terimi kullanılırken, 523.990 protein dizilimi analiz edilmiştir. Protein fonksiyonlarının GO terimleri GOLan sayısallaştırma yöntemi ile sayısallaştırılırken, elde edilen protein dizilimleri ise ProLan protein sayısallaştırma yöntemi ile sayısallaştırılmıştır. Protein ve GO terimlerinden gerekli öznitelikler elde edildikten sonra, RNN mimarisi kullanılarak sınıflandırma yapılmıştır. Tasarlanan sınıflandırıcının performansı sadece AUC skoru kullanılarak hesaplanmıştır. Çalışmanın sonucunda %33,3 oranında bir değer hesaplanmıştır. [103] numaralı çalışmada araştırmacılar, proteinlerin etkileşim ve sekans bilgilerini kullanarak protein fonksiyonlarını tahmin etmişlerdir.

Çalışmada protein fonksiyonları GO terimleri kullanılarak sınıflandırılmıştır. Çalışma kapsamında MF seviyesi için 6.221 adet, BS seviyesi için 19.181 adet ve HB seviyesi için ise 2.358 adet terim kullanılmıştır. Gerekli veriler elde edildikten sonra, araştırmacılar protein dizilimlerini sayısallaştırmak için ikili-kodlama yöntemine başvurmuşlardır. Sınıflandırma işlemi için ise ESA mimarisi geliştirilmiş ve kullanılmıştır. Geliştirilen sınıflandırıcının performansı F1-skor, MKK, AUC, kesinlik ve duyarlılık değerlendirme ölçütleri ile belirlenmiştir. Sınıflandırıcının başarımı BS, HB ve MF seviyeleri için ayrı ayrı test edilmiştir. MF seviyesi için 0,9 oranında AUC skoru elde edilirken, F1-skor ise %47 olarak hesaplanmıştır. HB seviyesine göre yapılan sınıflandırma işleminde ise, 0,95 oranında AUC skoru gözlemlenmiştir. Bu seviyede elde edilen F1-skor değeri ise %52 olarak bulunmuştur. Son olarak BS seviyesinde ise %34 oranında F1-skor hesaplanmıştır. Bu seviyede elde edilen AUC skoru ise 0,88'dir. [104] numaralı çalışmada araştırmacılar GO seviyelerini kullanarak protein fonksiyonlarını sınıflandırmışlardır. Sınıflandırma işlemi için öğrenilen sıralamayı öğrenmek tabanlı bir yaklaşım kullanılmıştır. Çalışmada protein dizilimleri ProFet, InterProt ve k-mer olmak üzere üç farklı protein sayısallaştırma yöntemi ile sayısallaştırılmıştır. Kullanılan sınıflandırıcının performansı hem AUC skoru hem de F1-skor ile belirlenmiştir. BS seviyesinde %37 oranında F1-skor gözlemlenirken, 0,23 oranında AUC skoru elde edilmiştir. En yüksek AUC skoru HB seviyesinde elde edilmiş ve 0,7 değerine ulaşmıştır. MF seviyesinde hesaplanan F1-skoru ise %69 olarak bulunmuştur. MF seviyesinde ise 0,55 AUC skoruna ulaşılırken, %69 F1-skoru hesaplanmıştır. [105] numaralı çalışmada ise protein fonksiyonları RO, ESA ve SA (Sinir Ağı) ile sınıflandırılmış ve tahmin edilmiştir. Çalışma doğrultusunda proteinleri sınıflandırmak için beş farklı yöntem kullanılmıştır. Sınıflandırma işlemi için hibrit model kullanılmış ve ESA, SA ve RO modelleri birleştirilmiştir. Geliştirilen hibrit sınıflandırıcının performansı F1-skor, kesinlik ve duyarlılık performans değerlendirme kriterleri ile saptanmıştır. Geliştirilen sınıflandırıcıyla, %49 oranında F1-skor, %55 oranında duyarlılık ve %45 oranında kesinlik elde edilmiştir. [106] numaralı çalışmada ise protein fonksiyonları graf teorisine dayanarak analiz edilmiş ve GO terimleri kullanılmıştır. Beş farklı türe ait protein fonksiyonları kullanılmıştır. Bu türler; insan, meyve sineği, zebra balığı, fare ve Avustralya kurbağasıdır. Çalışmada kapsamında, proteinlerin benzerlik oranına göre sayısallaştırma işlemi yapılmış ve proteinlerden öznetelikler elde edilmiştir. Uygulama sonuçlarına göre, BS seviyesi için %55,80, MF seviyesi için %71,63 ve HB seviyesi için %79,27 oranında doğruluk skorları elde edilmiştir. [107] numaralı çalışmada ise araştırmacılar KNN algoritmasını kullanarak protein fonksiyonlarını sınıflandırmışlardır. Çalışmada GO terimlerinden yalnızca BS seviyesi kullanılmış ve elde edilen protein dizilimlerini sayısallaştırmak için farklı benzerlik algoritmaları kullanılmıştır. Kullanılan sınıflandırma algoritmasının performansı ise sadece AUC değerlendirme ölçütü ile belirlenmiştir. Çalışmanın sonunda BS seviyesi için 0,77 oranında AUC skoru gözlemlenmiştir. [108] numaralı çalışmada araştırmacılar protein fonksiyonlarını sınıflandırmak için RY (Rastgele Yürüyüş)

sınıflandırıcısını uygulamışlardır. Maya ve insan olmak üzere iki farklı türe ait protein dizilimleri ve fonksiyonları kullanılmıştır. Geliştirilen sınıflandırıcının performansı ise sadece F1-skor ile ölçülmüştür. Çalışma kapsamında sınıflandırma işlemi için GO terimleri kullanılmış ve bu terimlerden sadece MF seviyesine başvurulmuştur. Çalışmanın sonunda insan için en yüksek F1-skor değeri %18 olarak hesaplanmıştır. Bu başarımların değeri maya için ise %27'ye yükselmiştir. [109] numaralı çalışmada araştırmacılar derin öğrenme modeli kullanarak protein dizilimlerinden protein fonksiyonlarını tahmin etmişlerdir. Sınıflandırıcı olarak ESA mimarisi kullanılmıştır. Çalışmaların çoğunda olduğu gibi, protein fonksiyonları GO terimlerine göre sınıflandırılmıştır. Başvurulan sınıflandırıcının performansı ise hem F1-skor hem de AUC skoru ile belirlenmiştir. Çalışmanın sonunda MF seviyesi için %58,5, BS için %47,4 ve HB için ise %69,9 oranında F1-skorlar elde edilmiştir. Bunun yanı sıra, en düşük AUC skoru BS seviyesinde gözlemlenmiştir. Bu seviyede elde edilen AUC skoru 0.407'dir. Ardından en yüksek AUC skoru MF seviyesinde elde edilmiş ve sonuç 0.536 olarak bulunmuştur. En başarılı AUC skoru ise HB seviyesinde görülmüş ve 0.726 oranında bir değer saptanmıştır. [110] numaralı çalışmada ise, araştırmacılar grafik çekirdeği yapısını kullanarak protein fonksiyonlarını sınıflandırmışlardır. Protein dizilimleri grafik teorisine göre sayısallaştırılmıştır. Sayısallaştırma adımı proteinlerin ikincil yapılarına bakılarak aralarındaki yollar belirlenmiştir. Bu sayede grafik yapısı oluşturulmuştur. Sınıflandırma adımı DVM kullanılmış ve protein fonksiyonları tahmin edilmiştir. Diğer çalışmaların aksine, bu çalışmada GO terimleri kullanılmamıştır. Bunun yerine enzimler ve enzim olmayan organik maddeler kullanılmıştır. Bu doğrultuda yapılan sınıflandırma işleminin performansı yalnızca doğruluk değerlendirme ölçütü ile belirlenmiştir. Çalışmanın sonunda %84,04 oranında doğruluk skoru elde edilmiştir.

4.4. İlaç-Hedef Etkileşimlerinin Tahmin Edilmesi

İlaç hedefleri, ilaç araştırma ve geliştirmenin temelinde yatmaktadır. Bu durumundan dolayı son birkaç yüzyılda insanlar ve araştırmacılar ilaç hedeflerine büyük ölçüde güvenmeye başlamıştır. Her ne kadar hedef proteinler ile etkileşime giren bilinen ilaçların sayısı artmaya devam etse de onaylanmış ilaç hedeflerinin sayısı halen insan proteomunun sadece küçük bir kısmını ifade etmektedir. İlaç-hedef etkileşimleri, ilaçların erken aşamada keşfedilmesinde çok önemli roller üstlenmektedir. Teknolojinin gelişmesiyle oluşan yeni imkanlar ile araştırmacılar bilinen ilaç hedeflerini kullanarak potansiyel ilaçları belirleyebilmektedir [111]. Yeni ilaçların geliştirilmesi, ilaçların takip edilmesi ve ilaçların sentezi, ilaçlar ve hedefleri arasındaki etkileşimlerin belirlenmesiyle mümkün olmaktadır [112]. Ancak, ilaç-hedef etkileşimlerinin belirlenmesi hem zaman açısından hem de bütçe açısından maliyetli olmaktadır. Çoğu zaman biyolojik ve kimyasal süreç ve deneylerden dolayı bu süre 2-3 yıl arasında değişiklik gösterebilmekte ve çok sayıda kaynak ile bütçeye ihtiyaç duyulabilmektedir [113-115]. Bundan dolayı ilaçlar, hedefler ve ilaç-

hedef etkileşimleri çeşitli hesaplama-tabanlı yöntemler ile daha hızlı bir şekilde belirlenebilmekte ve mümkün olabilecek ilaç-hedef etkileşimleri tahmin edilebilmektedir. Halka açık olan biyolojik ve kimyasal veri setlerindeki hızlı artış sayesinde, hesaplama-tabanlı yöntemlerin önemi artmış ve yeni verilerin sistematik bir biçimde öğrenilip analiz edilebilmesi mümkün olmuştur.

Hesaplama-tabanlı yöntemler üç farklı kategoride değerlendirilmektedir. Bu yöntemler yanalştırma yöntemi, özellik-tabanlı yöntem ve benzerlik-tabanlı yöntemlerdir. Benzerlik-tabanlı yöntemde, belirli bir ilacın veya bileşğin bilinen hedeflerinin kimyasal yapı benzerliği hesaplanmakta ve potansiyel hedefler belirlenmektedir. Benzerlik-tabanlı yöntemler, kimyasal yapı benzerliğinin yüksek olduğu durumlarda etkili olmaktadır. Uygulaması ve kullanılması kolay bir yöntemdir. Ancak, verilerin çok olduğu durumlarda, bu yöntemin doğruluğu azalmaktadır [112]. Buna ek olarak, benzerlik-tabanlı yöntemler, belirli bir protein ailesine ait ilaç-hedef etkileşimlerinin belirlenmesinde etkiliyken, diğer protein ailelerine ait etkileşimlerde etkili olamamaktadır [116]. Bu kısıtlamalardan dolayı, son zamanlarda bu yaklaşım araştırmacılar tarafından tercih edilmemektedir. Yanalştırma yönteminde ise, üç boyutlu yapılar üzerinde ilaçlar ve potansiyel hedeflerin yapıları ve elektriksel uyumları hesaplanmakta ve mümkün olabilecek hedefler belirlenmektedir. Yanalştırma yönteminde, hedef proteinin 3-boyutlu yapı bilgisine ihtiyaç duyulmaktadır. Bu nedenden dolayı, bu yaklaşımın tahmin doğruluğu düşük olabilmektedir [112]. Yanalştırma yönteminde görülen diğer bir problem ise, 3-boyutlu yapı bilgisi bilinmeyen proteinlere uygulanamamasıdır. Günümüzde 3-boyutlu yapı bilgisi bilinmeyen proteinler mevcuttur. Bu da yanalştırma yaklaşımının limitli olmasına ve aktif bir şekilde kullanılamamasına neden olmaktadır. Benzerlik-tabanlı ve yanalştırma yönteminin yeteri kadar etkili olmamaları, özellik-tabanlı yaklaşımın ortaya çıkmasına sebebiyet vermiştir [117, 118]. Özellik-tabanlı yöntemlerde, ilaç-hedefleri belirlenirken iki senaryo ön plana çıkmaktadır; ilaç ve hedefler arasında etkileşim ya vardır ya da yoktur. Bu yöntemde, bilinen etkileşimlere ait örüntüler kullanılmakta ve yapay zeka algoritmalarına dayalı yöntemler ile tahmin modeli geliştirilmektedir. Geliştirilen tahmin modeli ile potansiyel ilaç-hedef etkileşimleri belirlenmektedir. Ancak, makine öğrenmesi ile gerçekleştirilen çalışmalarda bilgi kaybı gözlemlenmektedir [119]. Kaybolan bilginin tekrar elde edilmesi geleneksel makine öğrenmesi yöntemleri ile zor olmaktadır. Bundan dolayı, makine öğrenmesinin neden olduğu problemin önüne geçebilmek için, derin öğrenme algoritmaları kullanılmaya başlanmıştır. [112] numaralı çalışmada araştırmacılar LSTM derin öğrenme modelini kullanarak ilaç-hedef etkileşimlerini tahmin etmişlerdir. Çalışmada protein dizilimleri PSSM ve LA (Legendre Anları) olmak üzere iki farklı protein sayısallaştırma yöntemi ile sayısallaştırılmıştır. Sayısallaştırma işleminin ardından 1842-boyutlu öznitelik vektörü elde edilmiş ve bu vektör üzerinde öznitelik seçimi yapılmıştır. Bu doğrultuda TBA (Temel Bileşen Analizi) yöntemi kullanılmış ve öznitelik vektörünün boyutu küçültülmüştür. Ardından LSTM derin öğrenme modeli tasarlanmış ve sınıflandırma işlemi gerçekleştirilmiştir. Sınıflandırıcının performansı doğruluk,

özgünlük, MKK ve AUC değerlendirme ölçütleri ile belirlenmiştir. Çalışmanın sonunda ortalama doğruluk skoru %91,95 olarak elde edilmiştir. Bunun yanı sıra ortalama AUC skoru 0,97 olarak hesaplanırken, %85,04 oranında ise MKK skoru gözlemlenmiştir. [120] numaralı çalışmada araştırmacılar ağaç-tabanlı kolektif öğrenme kullanarak ilaç ve hedefleri arasındaki etkileşimleri tahmin etmişlerdir. Çalışmada dört farklı veri setine ait dört farklı etiket değerlendirilmiştir. Bu etiketler; nükleer reseptör, G-proteini kenetli reseptör, iyon kanalları ve enzimlerdir. İlaç bileşenleri kimyasal yapılarındaki benzerliklere göre sayısallaştırılmıştır. Bunun yanı sıra hedef proteinleri ise dizilim hizalama kullanılarak sayısal hale getirilmiştir. Bu işlem için ise Smith-Waterman skoru kullanılmıştır. Sayısallaştırma işleminin ardından kolektif öğrenme tabanlı bir mimari geliştirilmiş ve sınıflandırma gerçekleştirilmiştir. Geliştirilen modelin performansı yalnızca AUC değerlendirme ölçütü ile belirlenmiştir. Çalışmanın sonunda, nükleer reseptör için 0,91, G-proteinli kenetli reseptör için 0,95, iyon kanalları ve enzimler için ise 0,97 oranında AUC skoru hesaplanmıştır. [116] numaralı çalışmada ise, araştırmacılar ilaç-hedef etkileşimlerini tahmin edebilmek için derin öğrenme modeli önermişlerdir. Çalışmada 1.412 adet ilaç verisi kullanılırken, 1.520 adet ise hedef proteini değerlendirmeye alınmıştır. Sınıflandırma işlemi için DİA (Derin İnanç Ağları) kullanılmıştır. Bunun yanı sıra önerilen modelin performansı doğruluk ve AUC skorları ile belirlenmiş ve modelin başarımı ise NB, RO ve KA (Karar Ağacı) sınıflandırıcılarıyla karşılaştırılmıştır. Çalışmanın sonunda en düşük doğruluk skoru NB sınıflandırıcısıyla elde edilmiş ve %72,72 oranında değer gözlemlenmiştir. KA ile %76,84 oranında doğruluk hesaplanırken, bu değer RO algoritmasıyla %83,42'ye kadar yükselmiştir. En yüksek doğruluk skoru ise DİA ile elde edilmiş ve sonuç %85,88 olmuştur. AUC skorlarında da benzer bir çıkarım gerçekleşmiştir. NB ile 0,75 oranında AUC skoru elde edilirken, bu skor DT ile 0,77'ye kadar yükselmiştir. En etkili sınıflandırma işlemini ise RO ve DİA algoritmaları gerçekleştirmiştir. RO algoritması ile 0,91 AUC skor, DİA sınıflandırıcısıyla ise 0,92 AUC skoru hesaplanmıştır. [119] numaralı çalışmada protein dizilimleri kullanarak ilaç-hedef etkileşimleri tahmin edilmiştir. Hedef proteinleri ve ilaç bileşenleri üç farklı veri setinden elde edilmiştir. Çalışmada 11.950 adet ilaç bileşeni ve 3.675 adet protein dizilim göz önünde bulundurulmuştur. İlaç bileşenlerini ve hedef proteinlerini sayısallaştırmak için parmak izi yöntemi kullanılmış ve verilerden öznitelikler elde edilmiştir. Sınıflandırma adımı ise ESA mimarisi kullanılmış ve öznitelikler sınıflandırıcıya girdi olarak verilmiştir. Sınıflandırıcının performansı, duyarlılık, kesinlik, doğruluk, AUC ve F1-skor değerlendirme ölçütleri ile belirlenmiştir. Bütün kriterler %80'in üzerinde sonuçlar üretmiştir. [121] numaralı çalışmada araştırmacılar hedef proteinlerin derin sinir özniteliklerini kullanarak ilaç-hedef tahminini gerçekleştirmişlerdir. Çalışmada protein dizilimlerinden gerekli öznitelikler LSTM derin öğrenme modeli ile elde edilmiştir. İlaç bileşenleri ise ESA mimarisi kullanılarak analiz edilmiş ve bu mimari aracılığıyla bu ilaçlardan öznitelikler elde edilmiştir. Elde edilen öznitelikler SSA sınıflandırıcısı ile sınıflandırılmış ve tahmin işlemi gerçekleşmiştir. Çalışmada

kullanılan sınıflandırıcının performansı AUC ve doğruluk değerlendirme ölçütleri ile belirlenmiştir. Uygulama sonuçlarına göre, önerilen sınıflandırıcının AUC skoru 0,81 olarak gözlemlenmiştir. [122] numaralı çalışmada ise moleküllerin imzaları kullanılarak ilaç-hedefleri arasındaki etkileşimler tahmin edilmiştir. Protein dizilimleri hem moleküler imzalar hem de reaksiyon imzaları ile sayısallaştırılmış ve sınıflandırıcıda sınıflandırılmak üzere hazır hale getirilmiştir. Sınıflandırma adımı DVM sınıflandırıcısına başvurulmuştur. Kullanılan sınıflandırma algoritmasının başarımı 5 katlı çapraz-doğrulama ile test edilirken, performansı ise doğruluk, kesinlik, duyarlılık, özgünlük ve AUC skorları ile belirlenmiştir. Çalışmanın sonunda gerçekleştirilen uygulama sonuçlarına göre, tasarlanan DVM modeli ile %94,2 oranında doğruluk skoru gözlemlenmiştir. %93,6 değerinde kesinlik skoru hesaplanırken, %95,2 oranında ise duyarlılık performansı sergilenmiştir. Kullanılan sınıflandırıcının AUC skoru ise 0,96 olarak bulunmuştur. [123] numaralı çalışmada araştırmacılar ilaç-hedef etkileşimlerini tahmin etmek için derin öğrenmeye ve transkriptom veri sınıflandırmasına başvurmuşlardır. İki farklı veri seti kullanarak hem ilaç bileşenleri hem de hedef protein dizilimleri değerlendirilmiştir. Sınıflandırma işlemi için DSA mimarisi tasarlanmış ve tasarlanan model LR, RO, OS (Oylama Sınıflandırıcı) ve GAKA (Gradyan Artırma Karar Ağacı) sınıflandırıcılarıyla karşılaştırılmıştır. Sınıflandırıcıların performansları F1-skor ve doğruluk değerlendirme ölçütleri ile belirlenmiştir. En düşük doğruluk skoru LR ile elde edilmiş ve sonuç %76,84 olarak bulunmuştur. Ardından, RO algoritması ile %87,12 oranında doğruluk elde edilmiştir. Geri kalan bütün sınıflandırıcılar %90'ın üzerinde başarımlar sergilemişler ve birbirine yakın değerler üretmişlerdir. F1-skorlarına göre kıyaslama yapıldığında en düşük sınıflandırma performansı LR ile elde edilmiş ve sonuç %68,88 olarak hesaplanmıştır. Bu skor RO sınıflandırıcı ile %77,57'ye kadar yükselmiştir. OS, GAKA ve DSA sınıflandırıcıları ise %80'in üzerinde performans sergilemişler ve en etkili sınıflandırma işlemi %86,38 F1-skoru ile DSA mimarisinden elde edilmiştir. [124] numaralı çalışmada araştırmacılar çoklu bilgi entegrasyonu yoluyla ilaç-hedefleri arasındaki etkileşimleri tahmin etmişlerdir. Çalışmada ilaç bileşenleri parmak izi yöntemi ile sayısallaştırılmıştır. Ardından öznelik vektörü oluşturulmuş ve en etkili öznelikler belirlenerek, vektör uzayının boyutu indirgenmiştir. Sınıflandırma işlemi için ise DVM kullanılmış ve sınıflandırıcının performansı yalnızca AUC değerlendirme skoru ile ölçülmüştür. Çalışmada kullanılan ilaç-hedef etkileşimleri iyon kanalları, enzimler, nükleer reseptör ve G-proteini kenetli reseptör olmak üzere dört farklı tür üzerinde test edilmiştir. Kullanılan sınıflandırıcı ve yöntemler ile, enzimler için 0,98, iyon kanalları için 0,99, G-proteini kenetli reseptör için 0,96 ve nükleer reseptör için ise 0,92 oranında AUC skorları hesaplanmıştır. [125] numaralı çalışmada ise araştırmacılar proteinlerin kimyasal, biyolojik ve ağ özelliklerini kullanarak, ilaç-hedef etkileşimlerini tahmin etmişlerdir. İlaç bileşenleri parmak izi yöntemi ile sayısal hale getirilmiştir. Bunlara ek olarak, protein dizilimleri ise biyolojik, kimyasal ve ağ özelliklerine göre sayısallaştırılmıştır. Sayısallaştırma işleminin ardından elde edilen

öznitelik vektörü kullanılarak RO sınıflandırıcısı ile sınıflandırma yapılmıştır. Dört farklı tür üzerinde gerçekleştirilen etkileşim sürecinde sadece AUC değerlendirme kriteri kullanılmış ve sınıflandırıcının performansı belirlenmiştir. Uygulama sonuçlarından elde edilen bulgulara göre, kullanılan sınıflandırıcı ile enzim için 0,98, G-proteini kenetli reseptör için 0,95, iyon kanalları için 0,99 ve nükleer reseptör için 0,92 AUC başarımları elde edilmiştir.

4.5. Kodlayan ve Kodlamayan RNA Dizilimlerinin Saptanması

RNA genetik bilgiyi saklayan bir makro molekül olup, bu genetik bilgiyi bir nesilden diğer nesile yüksek doğruluk ve esneklik ile aktarmaktadır. RNA'nın türleri merkezi dogma sürecinde önemli roller üstlenmekte ve aktivasyon, modifikasyon ve transportasyon gibi çeşitli özelliklere dayanarak karakterize edilmektedir. RNA'lar gerçekleştirmiş oldukları fonksiyonlara göre üç farklı türe ayrılmaktadır. Bunlar; mRNA, tRNA (transfer RNA) ve rRNA (ribozomal RNA)'dır. Bununla beraber, genetik materyal, kodlayan RNA ve kodlamayan RNA olarak sınıflandırılmaktadır. Kodlamayan RNA'lar biyolojik reaksiyonların katalizlenmesi, hücrel savunma sağlanması, gelişimsel süreçlerin kontrol edilmesi ve hücrel cevapların verilmesi olmak üzere pek çok görevlerde etkili olmaktadır [126-128]. Bunlara ek olarak, bazı çalışmalar kodlamayan RNA'ların hastalıkların gelişiminde de etkili olduğunu göstermiştir [129].

Kodlayan RNA'lar ile kodlamayan RNA'lar arasındaki ayrımı belirleyebilmek için kullanılan hızlı ve doğru yaklaşımlar, birçok transkriptin analiz edilmesinde kritik roller üstlenmektedir. Genellikle yeni-nesil hizalama algoritmaları kullanılmaktadır. Her ne kadar bu algoritmalar kodlayan ve kodlamayan RNA'ları etkili bir şekilde belirleyebilse de küçük kodlayan ve küçük kodlamayan RNA'ların ayrımını yapamamaktadır [130]. Araştırmalar, kodlamayan RNA'ların mikropeptitlerinin küçük açık okuma çerçeveleri içerdiğini göstermektedir [131, 132]. Bu nedenler kodlayan ve kodlamayan RNA'ların ayrıştırılması büyük bir önem arz etmektedir. Hücrel aktivitelere ve hastalık gelişiminde kodlamayan RNA'ların önemi büyük olduğu için, kodlamayan RNA'ları ve onların işlevlerini tahmin edebilmek için birçok deneysel yaklaşımlar geliştirilmiştir. Bu deneysel yaklaşımların başında mikro dizi analizi, SELEX genomik hizalama ve kimyasal RNA dizileme yöntemleri gelmektedir [133]. Ancak, deneysel yöntemlerle gerçekleştirilen çalışmalar hem maliyetli olmakta hem de zaman almaktadır [134]. Bu dezavantajlardan dolayı araştırmacılar hesaplama-tabanlı yaklaşımlar kullanarak, bu kısıtlamaları azaltmışlardır.

Literatürdeki çalışmalar incelendiğinde, kodlayan RNA ve kodlamayan RNA'ların arasındaki farkların belirlenebilmesi için yapay zeka tabanlı yaklaşımların etkili bir şekilde kullanıldığı gözlemlenmiştir. [135] numaralı çalışmada araştırmacılar derin öğrenme modeli kullanarak kodlamayan RNA'ları sınıflandırmışlardır. Çalışmada toplamda 187.609 adet RNA verisi değerlendirilmiştir. RNA dizilimleri k-mer yöntemi ile sayısal hale getirilmiş ve

sınıflandırma sürecinde kullanılmak üzere hazırlanmıştır. Bu adımdan sonra, derin öğrenme modeli tasarlanmış ve bu doğrultuda DSA sınıflandırıcısı kullanılmıştır. Sınıflandırıcının başarımı 10 katlı çapraz-doğrulama ile test edilmiştir. Bunun yanı sıra, performansı ise doğruluk, duyarlılık, özgünlük ve AUC değerlendirme ölçütleri ile belirlenmiştir. Çalışmanın sonunda ortalama %98,07 doğruluk skoruna ulaşılmıştır. Ortalama AUC skor ise 0,99 olarak hesaplanmıştır. [136] numaralı çalışmada ise araştırmacılar DVM sınıflandırıcısını kullanarak kodlayan ve kodlamayan RNA'ların ayrımını yapmışlardır. Çalışma kapsamında toplamda 5.610 adet RNA dizilimi kullanılmıştır. Kodlayan ve kodlamayan RNA'lar BLASTCLUST uygulaması ile analiz edilmiş ve DNA dizilimlerine dönüştürülmüştür. Ardından DNA dizilimlerinden öznitelikler elde edilmiştir. Her bir DNA dizilimden peptit uzunluğu, amino asit bileşimi, ortalama hidrofobiklik, ikinci yapı içeriği, kalıntı yüzdesi, dizilim bileşim entropisi, homolog sayısı ve dizilim entropisi olmak üzere 8 adet öznitelik elde edilmiştir. Öznitelik vektörünün oluşturulmasının ardından DVM sınıflandırıcısı kullanılmış ve sınıflandırma işlemi gerçekleştirilmiştir. Sınıflandırıcının performansı F1-skor, duyarlılık ve özgünlük başarımları ölçütleri ile belirlenmiştir. Uygulama sonuçlarına göre kodlayan RNA'lar %97,4 F1-skoru ile sınıflandırılmıştır. Kodlamayan RNA'larda ise bu oran %94,5'e kadar inmiştir. [137] numaralı çalışmada kodlayan ve kodlamayan RNA'ların ayrımı yapılmış ve bu doğrultuda ise DVM kullanılmıştır. RNA dizilimlerinden C-terminali, açık okuma çerçeve uzunluğu, açık okuma çerçeve kaplaması, heksamer skoru, Fickett skoru ve açık okuma çerçeve bütünlüğü olmak üzere 6 adet öznitelik elde edilmiştir. Bunlara ek olarak, RNA dizilimleri protein dizilimlerine dönüştürülmüş ve bu protein dizilimlerinden ise 3 adet öznitelik toplanmıştır. Sınıflandırıcının performansı doğruluk, özgünlük, duyarlılık, F1-skor, AUC ve MKK değerlendirme ölçütleri ile saptanmıştır. Çalışmanın sonunda %87,77 oranında doğruluk skoru hesaplanmıştır. Bunun yanı sıra, AUC skoru 0,96 olarak, MKK skoru ise 0,77 olarak hesaplanmıştır. [138] numaralı çalışmada araştırmacılar kodlanmayan RNA'ların yapı bilgilerinden yararlanarak, kodlayan ve kodlamayan RNA'ları sınıflandırmışlardır. Sınıflandırmak için DVM algoritması kullanılmıştır. RNA dizilimleri üçlü-nükleotid yöntemi kullanılarak sayısallaştırılmış ve sınıflandırılmak üzere hazırlanmıştır. Çalışmada 444.417 adet kodlanmayan RNA verisi kullanılırken, 97.386 adet ise kodlayan RNA dizilimleri değerlendirilmiştir. Tasarlanan DVM modelinin performansı MKK, doğruluk, özgünlük ve duyarlılık başarımları değerlendirme ölçütleri ile belirlenmiştir. Uygulama sonuçlarına göre ise %99,46 oranında doğruluk elde edilmiştir. [139] numaralı çalışmada kodlayan ve kodlamayan RNA'lar YTÖ (Yinelemeli Topluluk Öğrenmesi) yöntemi ile sınıflandırılmıştır. RNA dizilimleri spektrum profili, uyumsuzluk profili, k-mer profili, sözde nükleotit bileşimi, çapraz kovaryans ve transkript olmak üzere 6 farklı yöntem ile sayısallaştırılmıştır. Çalışmada tasarlanan sınıflandırıcının performansı duyarlılık, özgünlük ve doğruluk başarımları ölçütleri ile belirlenmiştir. Çalışmanın sonunda %91,21 doğruluk skoru gözlemlenmiştir. [140] numaralı çalışmada araştırmacılar kodlamayan RNA dizilimlerini

saptayacak yeni bir araç geliştirmişlerdir. Çalışmada hem insana hem de fareye ait RNA dizilimleri kullanılmıştır. RNA dizilimleri k-mer yöntemi ile haritalanmış ve sınıflandırılmak üzere hazırlanmıştır. Sınıflandırma işlemi RO ile yapılmış ve sınıflandırıcının performansı doğruluk, kesinlik, duyarlılık, özgünlük, F1-skor ve MKK değerleri ile belirlenmiştir. İnsana ait RNA'lar üzerinde gerçekleştirilen sınıflandırma sonucunda, %91,9 oranında doğruluk skoru hesaplanırken, 0,84 oranında ise MKK elde edilmiştir. [141] numaralı çalışmada ise araştırmacılar gözetimli öğrenme kullanarak kodlayan RNA'ları saptamışlardır. Çalışmada insan, fare, sinek, solucan ve bitkiye ait RNA dizilimleri kullanılmıştır. RNA dizilimleri çeşitli kimyasal yöntemler ile sayısal hale getirilmiş ve sınıflandırılmıştır. Sınıflandırıcının performansı doğruluk, AUC, F1-skor, duyarlılık ve özgünlük başarımları ölçütleri kullanılarak değerlendirilmiştir. İnsana ait kodlayan RNA dizilimleri %94,7 doğruluk, %89,3 F1-skor ve 0,98 AUC skoru ile sınıflandırılmıştır. Fare için ise %94,6 doğruluk değeri elde edilirken, %86,5 F1-skor ve 0,98 AUC skoru hesaplanmıştır. En yüksek doğruluk değeri ise solucan RNA'larında görülmüştür. Solucana ait kodlayan RNA'lar %99,1 doğruluk skoru ile sınıflandırılmıştır. Bu tür için AUC skoru ve F1-skoru ise sırasıyla 0,98 ve %47,3 olarak gözlemlenmiştir. Sinek için %98,3 oranında doğruluk skoru elde edilirken, bu oran bitki için %98,6'ya kadar yükselmiştir. AUC skorlarına bakıldığında ise sineğin bitkiden daha iyi olduğu gözlemlenmiştir. Sinek ile 0,99 AUC skoru elde edilirken, bitki için bu skor 0,98 olmuştur. F1-skorlarına göre ise sineğin daha iyi sınıflandırıldığı tespit edilmiştir. Bitki için %51,2 F1-skoru hesaplanırken, sinek için bu skor %61,3 olmuştur. [142] numaralı çalışmada araştırmacılar DVM algoritmasını kullanarak uzun kodlamayan RNA'ların ayrımını yapmışlardır. Çalışmada insan, fare ve zebra balığına ait RNA dizilimleri kullanılmıştır. Ardından RNA'ların nükleotit örüntü frekansları hesaplanmış ve dizilimlerinin sayısallaştırılması sağlanmıştır. Sayısallaştırma adımından sonra eğitim ve test olmak üzere iki farklı veri seti oluşturulmuştur. Eğitim veri seti PBA ile analiz edilmiş ve veri setinin boyutu indirgenmiştir. Son kısımda ise DVM sınıflandırıcısı kullanılmış ve kodlamayan RNA'ların ayrımı gerçekleştirilmiştir. Sınıflandırıcının performansı doğruluk, kesinlik ve duyarlılık değerlendirme kriterleri ile belirlenmiştir. İnsan verileri için %91,94, fare için %92,21 ve zebra balığı için ise %97,71 oranında doğruluk skoru elde edilmiştir. [143] numaralı çalışmada araştırmacılar RO kullanarak kodlamayan RNA'ları sınıflandırmışlardır. Veriler bu çalışmada insan, fare ve diğer türler olmak üzere üç farklı kategoride değerlendirilmiştir. Eğitim veri seti oluşturulurken kümeleme yapılmış ve SOM (Self Organizing Map - Kendi Kendini Düzenleyen Harita) kullanılmıştır. Sinyal gürültü oranı kullanılarak RNA dizilimleri sayısallaştırılmış ve sınıflandırıcıda kullanılmak üzere vektör haline getirilmiştir. Kullanılan RO sınıflandırıcısının performansı duyarlılık, özgünlük, doğruluk ve MKK değerlendirme ölçütleri ile belirlenmiştir. Uygulama sonucunda elde edilen bulgular doğrultusunda, insan için %92,96 doğruluk ve 0,86 MKK skoru elde edilmiştir. Fareye ait kodlamayan RNA'ların analizi sonucunda ise, %94,3 doğruluk değerine ve 0,89 MKK skoruna ulaşılmıştır. Diğer türler için doğruluk oranı

%97,78 iken, MKK skoru ise 0'dır. [144] numaralı çalışmada koli basiline ait kodlamayan RNA'lar genetik programlama ile tahmin edilmiştir. Genetik programlama tabanlı regresyon kullanılmış ve bir RNA diziliminin kodlayan ya da kodlamayan olup olduğu belirlenmiştir. Sınıflandırıcının performansı MKK ve duyarlılık değerlendirme skorları ile saptanmıştır. Sınıflandırıcının başarımı ise 10 katlı çapraz-doğrulama ile test edilmiştir. Uygulama sonucunda, önerilen yöntem ile %54 duyarlılık değeri elde edilmiştir.

4.6. Kanser İçeren Protein Dizilimlerinin Belirlenmesi

Amino asit varyasyonlarının ve mutasyonlarının insan hastalığındaki etkileri büyüktür. Mutasyonlar, protein katlanması ve stabilizesini [145, 146], protein fonksiyonlarını [147], protein-proteine etkileşimlerini [148] ve protein ifadeleri ile alt hücre lokalizasyonunu [149] etkilemektedir. Proteinlerde ortaya çıkan mutasyonlar, kanserin başlangıcında ve gelişmesinde büyük rol oynamaktadır [150]. Mutasyonlar, onkogenlerde olduğu gibi protein fonksiyonunu aktive ederek ya da tümör baskılayıcı genlerde olduğu gibi fonksiyonu inaktive ederek kanserin ortaya çıkmasına neden olmaktadır. Proteinlerde görülen mutasyonların, kanserdeki etkisinin farkına varan ve dizileme teknolojisindeki son gelişmelerden yararlanan araştırmacılar, kanser örneklerini daha detaylı incelemeye başlamışlardır. Yeni dizileme teknolojileri, yılda binlerce kanserle ilişkili protein mutasyonlarını bulabilmektedir. Ancak bu yöntemin en büyük sorunu, kanserli hücreleri erken teşhis edememesidir [151]. Kanserlin erken teşhis edilmesi kanser tedavisindeki en önemli unsurlardan birisidir. Bu sayede hastanın hayat kalitesi artmakta ve hayat süresi uzayabilmektedir. Kanseri erken teşhis edebilmek için çeşitli yöntemler kullanılmaktadır. En çok kullanılan yöntemler röntgen teknolojisi, endoskopi, histolojik analiz, kan testleri, genetik görüntüleme ve manyetik rezonanstır. Fakat, tüm bu yöntemler zaman almakta, laboratuvar ihtiyacı gerektirmekte ve uzman görüşüne ihtiyaç duymaktadır. Bu kısıtlamalardan dolayı, araştırmacılar kanseri ve kansere neden olan mutasyonları erken aşamada teşhis edebilmek için yeni yaklaşımlar aramaktadır. Mutasyonun protein fonksiyonu üzerindeki etkilerini değerlendirmek için yıllarca çeşitli yöntemler geliştirilmiştir. Mutasyonun etkilerini değerlendirebilmek için genel olarak amino asitlerin fizikokimyasal özellikleri kullanılmakta ve amino asitlerin yan zincirlerinin protein yapısındaki rolü göz önünde bulundurulmaktadır. Bu tür yöntemler genellikle hesaplama-tabanlı yaklaşımların kullanılmasına vesile olmuş ve bu yaklaşımların önemini arttırmıştır. Hesaplama-tabanlı yaklaşımlarda makine öğrenmesi ve derin öğrenme sıklıkla kullanılmaktadır. [152] numaralı çalışmada kanserli genomları belirleyebilmek için derin öğrenme modeli kullanılmış ve 12 farklı kanser türü sınıflandırılmıştır. Çalışma kapsamında genler üzerinde yer alan nokta mutasyonlardan faydalanılmıştır. Sınıflandırılmış olan kanser türleri gliyom, yumurtalık, akciğer, pankreas, rahim, prostat, kolon, göğüs, mesane, beyin tümörü, böbrek ve mide kanseridir. Araştırma kapsamında kanser türlerinden protein dizilimleri elde edilmiş ve sınıflandırılmak üzere kullanılmıştır.

Çalışmada göz önünde bulundurulmuş sağlıklı dizilim sayısı 1.991 adetken, toplamda 6.083 adet kanserli dizilim değerlendirilmiştir. Sınıflandırma işlemi için DSA mimarisi tasarlanmıştır. Sınıflandırma işleminden önce protein dizilimleri ikili-kodlama yöntemi ile sayısal hale getirilmiştir. Sınıflandırıcının performansı doğruluk, duyarlılık ve özgünlük başarımları değerlendirme ölçütleri ile belirlenmiştir. Uygulama sonuçlarından elde edilen bulgulara dayanarak, %94,70 doğruluk, %97,30 duyarlılık ve %85,54 özgünlük değerleri hesaplanmıştır. [153] numaralı çalışmada araştırmacılar kanserli genleri belirleyebilmek için protein-protein etkileşimlerini kullanmışlardır. Mesane, tiroit, böbrek ve kolon olmak üzere dört farklı kansere ait protein dizilimleri değerlendirilmiştir. Çalışmada toplamda 9.856 adet gen ifadesi kullanılırken, 7.039 adet protein etkileşim verileri ele alınmıştır. Uygulama sonuçlarına göre geliştirilen yöntemin etkili olduğu vurgulanmıştır. [154] numaralı çalışmada araştırmacılar akciğer kanserine neden olan genomları belirleyebilmek için farklı makine öğrenmesi algoritmalarını kullanmışlardır. Sınıflandırma işlemi için proteinler arasındaki etkileşimlerden, alan etkileşim ve alan frekanslarından faydalanmışlardır. Sınıflandırma işlemi için RO, LR ve NB sınıflandırıcıları kullanılmıştır. Sınıflandırıcıların performansı doğruluk, duyarlılık, özgünlük, F1-skor, MKK ve AUC başarımları değerlendirme kriterleri ile belirlenmiştir. RO algoritması ile %76 doğruluk, %77,7 özgünlük, %74,4 duyarlılık, %52,4 MKK ve 0,76 AUC skoru elde edilmiştir. LR algoritmasıyla ise %66,9 doğruluk, %65,2 özgünlük, %69,2 duyarlılık, %34,2 MKK ve 0,67 AUC skorları hesaplanmıştır. En düşük sınıflandırma performansı ise NB algoritmasıyla elde edilmiştir. NB sınıflandırıcısıyla %59,8 doğruluk skoru gözlemlenmiştir. Bu sınıflandırıcının özgünlük ve duyarlılık değerleri ise sırasıyla %57,1 ve %66 olarak belirlenmiştir. MKK ve AUC skorları ise sırasıyla %21,4 ve 0,62 olarak hesaplanmıştır. [155] numaralı çalışmada ise akciğer kanserine neden olan proteinler ESA mimarisi kullanılarak sınıflandırılmıştır. Çalışmada protein-protein etkileşimleri ve gen ifadeleri kullanılmıştır. Çalışmanın ilk aşamasında veri seti oluşturulmuş ve 487 adet kanserli genomlar kullanılırken, 152 adet kanser içermeyen genomlar değerlendirilmiştir. Proteinler arasındaki etkileşim verileri için ise toplamda 47.358 adet etkileşim kullanılmıştır. Veri seti oluşturulduktan sonra ise sınıflandırma yapılmış ve bu amaç için ESA mimarisi tasarlanmış ve kullanılmıştır. Geliştirilen sınıflandırıcının performansı doğruluk, duyarlılık, özgünlük ve kesinlik skorları ile belirlenmiştir. Buna ek olarak, başarımları ise 10 katlı çapraz-doğrulama ile test edilmiştir. Uygulama sonuçlarından elde edilen bulgular ışığında, ortalama doğruluk skoru %81 olarak hesaplanmıştır. Bununla beraber, ortalama özgünlük değeri %74, ortalama kesinlik değeri %78 ve ortalama duyarlılık değeri ise %88 olarak elde edilmiştir. [156] numaralı çalışmada akciğer kanserine neden olan protein dizilimleri makine öğrenmesi algoritmaları ile sınıflandırılarak saptanmıştır. Çalışmada UHCK (Ufak Hücre Ciğer Kanseri), UOHCK (Ufak Olmayan Hücre Ciğer Kanseri) ve ortak olmak üzere 3 farklı sınıfa ait protein dizilimleri kullanılmıştır. Sınıflandırma adımında DVM, ANN ve NB algoritmaları kullanılmıştır. Sınıflandırıcıların performansı yalnızca

doğruluk skoru ile ölçülmüştür. Çalışmanın sonunda DVM ile %71,9 doğruluk ile sınıflandırma gerçekleştirilmiştir. ANN ile doğruluk skoru %85,9'a kadar yükselmiştir. En başarısız doğruluk skoru ise NB ile elde edilmiş ve doğruluk skoru %69,3 olarak hesaplanmıştır. [157] numaralı çalışmada gen ifadeleri kullanılarak kanserli genler saptanmış ve bu doğrultuda derin öğrenme algoritması kullanılmıştır. Çalışmada 32 farklı kanser türü göz önünde bulundurulmuştur. Bu amaçla toplamda 10.267 tane protein dizilimi değerlendirilmiştir. Sınıflandırma adımıyla ise 2-boyutlu ESA mimarisi tasarlanmış ve kullanılmıştır. Sınıflandırıcının performansı doğruluk, F1-skor, duyarlılık ve kesinlik değerlendirme kriterleri ile ölçülmüştür. Geliştirilen ESA mimarisinin doğruluk skoru %95,65 olarak elde edilirken, F1-skoru ise %94,56 olarak hesaplanmıştır. Bunun yanı sıra, sınıflandırıcı RO, DVM ve KNN algoritmalarıyla da kıyaslanmıştır. En düşük sınıflandırma performansı KNN ile gözlemlenmiş ve bu sınıflandırıcı ile %90 doğruluk skoruna ulaşılmıştır. Bunun yanı sıra, RO ile %92,5 doğruluk, DVM ile ise %94,5 doğruluk skoruna erişilmiştir. En başarılı sınıflandırma ise önerilen ESA mimarisi ile elde edilmiş ve %95,65 oranında doğruluk skoru gözlemlenmiştir. [158] numaralı çalışmada ise kanser molekülleri derin öğrenme algoritması ile sınıflandırılmış ve bu moleküllerin alt tipleri saptanmıştır. Çalışma kapsamında koleraktal kanseri için analiz yapılmış ve bu doğrultuda RNA dizilimleri ile mikro dizi verileri kullanılmıştır. Hem RNA hem de mikro dizi verileri çeşitli araçlarla çözümlenmiş ve spektrum öznelikleri elde edilerek bu dizilimler sayısal hale getirilmiştir. Sayısallaştırma işleminin ardından YSA sınıflandırıcısı geliştirilmiş ve öznelıklar sınıflandırılmak üzere değerlendirilmiştir. YSA sınıflandırıcısının başarımı çeşitli makine öğrenmesi algoritmaları ile de kıyaslanmış ve sınıflandırıcıların performansları değerlendirilmiştir. Geliştirilen YSA modelinin performansı duyarlılık, özgünlük ve doğruluk skoru ile ölçülmüştür. Uygulama sonucuna göre YSA ile %92 doğruluk oranı ile moleküllerin alt tipleri saptanmıştır. [159] numaralı çalışmada araştırmacılar derin öğrenme mimarisi kullanarak meme tümörünü tahmin etmişlerdir. Çalışma doğrultusunda gen ifadeleri kullanılmış ve toplamda 30.612 adet veri değerlendirilmiştir. Öznelıkların elde edilmesi aşamasında gen ifadeleri çeşitli yöntemlerle görüntülere dönüştürülmüştür. Ardından ESA modeli ile bu görüntüler sınıflandırılmış ve meme türü kanserleri saptanmıştır. Çalışmanın sonunda, sınıflandırıcının performansı korelasyon skoru ile ölçülmüş ve sonuç %48 olarak elde edilmiştir.

Literatürdeki tüm çalışmalar incelendiğine, bu çalışmaların karşılaşmış olduğu problemler ve bunlar için önerilen çözümler aşağıda verilmiştir:

- Proteinler arasındaki etkileşimler belirlenirken genellikle in-vivo ve in-vitro olmak üzere çeşitli deneysel yaklaşımlar kullanılmaktadır. In-vitro ile gerçekleştirilen çalışmaların sonucunda, organizmanın tamamındaki etkileşimler net ve tam olarak belirlenmeyebilmektedir [25-28]. Bu yöntemde görülen başka bir dezavantaj ise, organizmanın sahip olduğu hassas hücresel durumlar oluşturulamamaktadır [29].

Buna ek olarak, in-vivo yöntemlerde ise çok sayıda yanlış pozitif etkileşimler elde edilmektedir [48]. Bu yöntemlerle gerçekleştirilen deneysel uygulamalarda çok sayıda veri oluştuğu için, verilerin işlenmesi hem zaman almakta hem de bu verilerin işlenmesi için laboratuvar ihtiyacı gerekmektedir [50]. Bu nedenlerden dolayı araştırmacılar in-silico tabanlı yaklaşımlara yönelmiş ve bu yaklaşımların popülaritesi artmıştır [52]. In-silico tabanlı yöntemlerin hem ekipman hem bütçe hem de laboratuvar ihtiyacı bakımından daha az maliyetli olduğu gözlemlenmiştir [68]. Bundan dolayı in-silico yaklaşımlarıyla gerçekleştirilen çalışmalar, in-vivo ve in-vitro yöntemlerinde karşılaşılan problemlere çözüm olmuştur. In-silico tabanlı yaklaşımlarda makine öğrenmesi ve derin öğrenme sıklıkla kullanılmakta ve protein dizilimleri analiz edilmektedir. Proteinlerin analiz edilebilmesi için protein dizilimlerinin sayısallaştırılması gerekmektedir. Tez çalışmasında ise, yeni protein sayısallaştırma yöntemleri önerilmiş ve in-silico yaklaşımlarında kullanılması sağlanmıştır.

- Protein aileleri belirlenirken, benzer fonksiyonlara sahip olan protein aileleri bir grup olarak ifade edilmektedir. Ancak, protein ailelerinin, yapı bilgilerine dayalı olarak belirlenmesi, karakterize edilmemiş protein dizilimlerinin elde edilmesinin zor olması nedeniyle, sağlıklı olmamaktadır [77]. Ayrıca, 3-boyutlu protein yapılarının kullanılması zaman ve maliyet açısından verimli olmamaktadır [82]. Bu nedenle, araştırmacılar alternatif yollar arayarak, protein dizilimleriyle gerçekleştirilen çalışmalara yoğunlaşmışlardır [80]. Protein dizilimlerini analiz etmek ve protein dizilimlerinden anlamlı, ayık ve sürekli öznitelikler toplamak için, öznitelik mühendisliğine başvurulmaktadır. Gerekli öznitelikler elde edildikten sonra ise, derin öğrenme ve makine öğrenmesi algoritmaları kullanılmakta ve protein aileleri sınıflandırılmaktadır. Ancak, bu sınıflandırma algoritmalarının kullanılabilmesi için, protein dizilimlerinin sayısallaştırılması gerekmektedir. Bu nedenle bu çalışmalarda protein sayısallaştırma yöntemleri sıklıkla kullanılmaktadır. Tez kapsamında önerilen yöntemlerden birisi bu alana uygulanmış ve önerilen yöntemin protein dizilimleri üzerindeki etkisi araştırılmıştır.
- Protein fonksiyonları belirlenirken deneysel yaklaşımlar kullanılmaktadır. Fakat, deneysel yaklaşımlarda kimyasal işlemlerin zaman alması, laboratuvar ihtiyacının artması ve insan kaynağına ihtiyaç duyulması, bu yaklaşımların maliyetli olmasına neden olmaktadır [92, 93]. Bunun yanı sıra, teknolojik gelişmelerin genetik alanında artış göstermesiyle, yeni protein dizilimleri elde edilmekte ve veri sayısında artış olduğu için analiz işlemleri yavaş olmaktadır [94]. Bu kısıtlamalardan dolayı, hesaplama-tabanlı analiz yöntemlerinin önemi artmış ve bu yöntemler protein

fonksiyonlarının tahmini çalışmalarında sıklıkla kullanılmaya başlanmıştır. Hesaplama-tabanlı yaklaşımlarda protein dizilimlerinin analize edilebilmesi için sayısal ifadelerle dönüştürülmesi gerekmektedir. Bundan dolayı çalışmalarda çeşitli kategorilere ait protein sayısallaştırma yöntemleri kullanılmaktadır. Tez çalışmasında önerilen protein sayısallaştırma yöntemleri ileriki zamanlarda bu alanlara da uygulanacaktır.

- Teknolojinin gelişmesiyle oluşan yeni imkanlarla bilinen ilaç hedefleri kullanılarak potansiyel ilaçlar belirlenebilmektedir. Yeni ilaçların geliştirilmesi, ilaçlar ve hedefleri arasındaki etkileşimlerin belirlenmesiyle mümkün olmaktadır [112]. Ancak, bu etkileşimlerin belirlenmesi hem zaman açısından hem de bütçe açısından maliyetli olmaktadır. Çoğu zaman, kimyasal süreç ve deneylerden dolayı bu süre 2-3 yıl arasında değişiklik göstermekte ve çok sayıda kaynak ile bütçeye ihtiyaç duyulmaktadır [113-115]. Bundan dolayı ilaçlar, hedefler ve ilaç-hedef etkileşimleri çeşitli hesaplama-tabanlı yöntemler ile daha hızlı bir şekilde belirlenerek, ilaç-hedef etkileşimleri saptanmaktadır. Diğer çalışmalarda da olduğu gibi, ilaç—hedef etkileşimlerinin hesaplama-tabanlı yaklaşımlar ile analiz edilebilmesi için protein dizilimlerinin sayısal hale getirilmesi gerekmektedir. Tez çalışmasında literatüre üç farklı protein sayısallaştırma yöntemi önerilmiş ancak bu yöntemler bu alana uygulanmamıştır. İleride gerçekleştirilecek çalışmalarda değerlendirilecektir.
- Kodlayan RNA'lar ile kodlamayan RNA'lar arasındaki ayrımı belirleyebilmek için genellikle yeni-nesil hizalama algoritmaları kullanılmaktadır. Her ne kadar bu algoritmalar kodlayan ve kodlamayan RNA'ları etkili bir şekilde saptayabilse de küçük kodlayan ve küçük kodlamayan RNA'ların ayrımını yapamamaktadır [130]. Hücrel aktivitelere ve hastalık gelişiminde kodlamayan RNA'ların önemi büyük olduğu için, kodlamayan RNA'ları ve onların işlevlerini tahmin edebilmek için birçok deneysel yaklaşımlar geliştirilmiştir. Bu deneysel yaklaşımların başında mikro dizi analizi, SELEX genomik hizalama ve kimyasal RNA dizileme yöntemleri gelmektedir [133]. Ancak, deneysel yöntemlerle gerçekleştirilen çalışmalar hem maliyetli olmakta hem de zaman almaktadır [134]. Bu dezavantajlardan dolayı araştırmacılar hesaplama-tabanlı yaklaşımlar kullanarak, bu kısıtlamaları azaltmışlardır. Literatürde incelenen diğer çalışmalarda da olduğu gibi, hesaplama-tabanlı analiz işlemlerinde protein dizilimlerinin sayısal ifadelerle dönüştürülmesi gerekmektedir. Araştırmacılar çeşitli kategorilere ait protein sayısallaştırma yöntemlerini sıklıkla bu alana uygulayarak, kodlayan ve kodlamayan RNA'ların ayrımını yapmaktadırlar. Tez çalışmasında önerilen protein sayısallaştırma

yöntemleri bu alana uygulanmamıştır. İlerleyen zamanlarda bu alana da uygulanacak ve yöntemlerin başarımı değerlendirilecektir.

- Proteinlerde görülen mutasyonlar, kanserin gelişmesinde büyük rol oynamaktadır. Yeni dizilime teknolojileri ile bu protein mutasyonları belirlenmektedir. Ancak bu yöntemin en büyük sorunu, kanserli hücreleri erken teşhis edememesidir [151]. Kanserlin erken teşhis edilmesi kanser tedavisindeki en önemli unsurlardan birisidir. Bu sayede hastanın hayat kalitesi artmakta ve hayat süresi uzayabilmektedir. Kanseri erken teşhis edebilmek için çeşitli yöntemler kullanılmaktadır. En çok kullanılan yöntemler röntgen teknolojisi, endoskopi, histolojik analiz, kan testleri, genetik görüntüleme ve manyetik rezonanstır. Fakat, tüm bu yöntemler zaman almakta, laboratuvar ihtiyacı gerektirmekte ve uzman görüşüne ihtiyaç duymaktadır. Mutasyonun protein fonksiyonu üzerindeki etkilerini değerlendirmek için yıllarca çeşitli yöntemler geliştirilmiştir. Mutasyonun etkilerini değerlendirebilmek için genel olarak amino asitlerin fizikokimyasal özellikleri kullanılmakta ve amino asitlerin yan zincirlerinin protein yapısındaki rolü göz önünde bulundurulmaktadır. Bu tür yöntemler genellikle hesaplama-tabanlı yaklaşımların kullanılmasına vesile olmuş ve bu yaklaşımların önemini arttırmıştır. Bu yaklaşımlar ile daha hızlı bir analiz işlemi gerçekleştirilmekte ve hem ekipman hem de insan gücü açısından verim elde edilmektedir. Hesaplama-tabanlı yaklaşımlarda ilk olarak protein dizilimleri sayısal hale getirilmektedir. Tez çalışmasında önerilen protein sayısallaştırma yöntemleri ise ileriki çalışmalarda bu alana da uygulanacaktır.

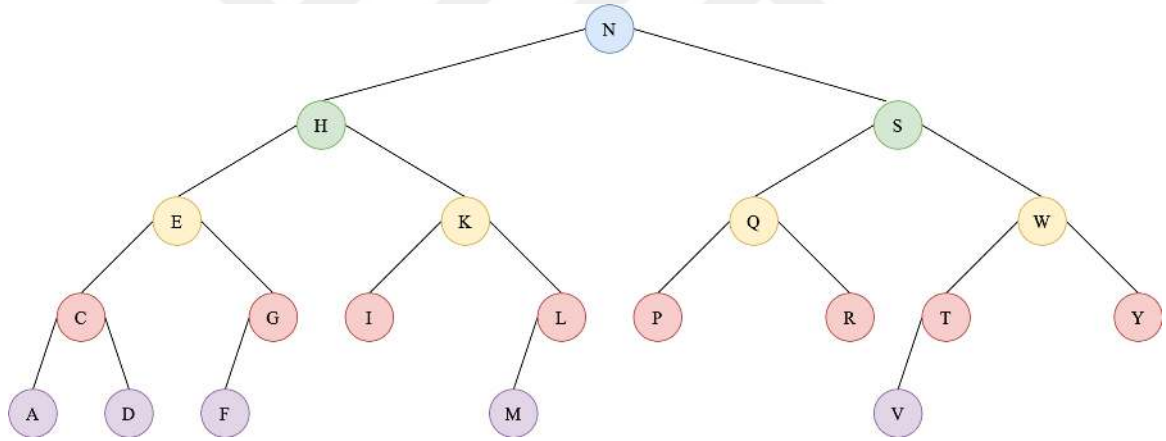
5. AVL AĞACI TABANLI YENİ BİR SAYISALLAŞTIRMA YÖNTEMİ

Tezin bu kısmında, mevcut protein sayısallaştırma yöntemlerine ek olarak, AVL-ağacı tabanlı yeni bir protein sayısallaştırma yöntemi önerilmiştir. Önerilen bu yöntem, proteinler arasındaki etkileşimleri tahmin etmek için kullanılmış ve yöntemin literatürdeki diğer protein sayısallaştırma yöntemleri ile kıyaslanması yapılarak başarımı değerlendirilmiştir. Önerilen AVL ağacı protein sayısallaştırma yöntemi, içeriğinde bulunan ikili arama ağacı yapısı nedeniyle yeni bir kategorisel yaklaşım olan algoritmik-tabanlı protein sayısallaştırma kategorisinde değerlendirilmiştir.

5.1. Önerilen Sayısallaştırma Yöntemi

Protein dizilimlerinin bilgisayar bilimlerinde algoritmik süreçlerle değerlendirilebilmesi için sayısallaştırılması yani rakamlara dönüştürülmesin gerekmektedir. Algoritma, bir problemi çözmek için izlenecek yollar dizisi olarak ifade edilmektedir. Belirli bir problem için bir algoritma tanımlanıp modellendikten sonra, algoritmanın ne kadar sürede çalışacağı ya da ne kadar alana ihtiyacı olacağı belirlenmektedir. Bir problemi yıllar içinde çözen ya da binlerce gigabaytlık hafıza kullanan bir algoritma işe yaramamaktadır. Bu nedenle, zaman ve hafıza bakımından verimli bir program için düzgün bir algoritmanın modellenmesi gerekmektedir. Algoritmaların performansı zaman karmaşıklığı analizi ile belirlenmektedir [160]. Analiz süreci, işletim sistemi, işlemci ve donanım gibi birden çok çevresel ve dahili koşullara bağlı olmaktadır. Ancak algoritmaların analizinde bu durum söz konusu değildir. Girdi boyutu arttıkça, en iyi performansın sabit $O(c)$ zamanından elde edildiği bilinmektedir [160, 161]. Bunu sırasıyla $O(\log N)$, $O(N)$, $O(N \log N)$, $O(N^c)$ ve $O(c^N)$ takip etmektedir. Biyoenformatikte olduğu gibi büyük miktarda veriler için doğrusal bir karmaşıklık elde etmek zor olmaktadır. Bu nedenle, çoğu durumda ortalama çalışma süresi $O(\log N)$ olduğu için bir ağaç yapısına ihtiyaç duyulmaktadır. Bilgisayar biliminde ikili arama ağacı, AVL ağacı, kırmızı-siyah ağacı, N-ary ağacı olmak üzere çeşitli sayıda ağaç yapısı bulunmaktadır. Bu ve buna benzer ağaçlar, veri yapıları ve algoritma analizi açısından önem arz etmektedir. Günümüzde gerçekleştirilen uygulamaların çoğunda ağaç yapısı kullanılmaktadır. Ağaç yapısında veriler genellikle hiyerarşik bir şekilde saklanmaktadır. Her bir ağacın kendine göre avantajları ve dezavantajları bulunmaktadır. Önerilen yöntemde AVL ağacına dayalı bir protein sayısallaştırma yöntemi geliştirilmiştir. AVL ağacının kullanılmasının nedeni çoğu durumda zaman karmaşıklığının $O(\log N)$ olması ve işletim süresinin hızlı olmasıdır. İkili arama ağacı ve yayılma ağacı gibi $O(\log N)$ zamanda çalışan başka ağaçların da olduğu bilinmektedir. İkili arama ağacında bir denge durumu söz konusu değildir. Yayılma ağacında ise denge durumu vardır, ancak silme ve

ekleme işlemleri $O(N \log N)$ zamanda gerçekleştiği için daha çok işletim süresine gerek duymaktadır [162]. Ayrıca, AVL ağacı mantıksal bir yapıya sahipken, yayılma ağacında ise sezgisel bir yapı bulunmaktadır [163]. AVL ağacının diğer ağaçlardan en büyük üstünlüğü arama işlemini hızlı yapması ve dengeli bir yapıya sahip olmasıdır [164]. Dengesiz olan ağaçlarda işlemler uzun sürmekte bu da uygulamaların verimsiz olmasına sebebiyet vermektedir. AVL ağacı denge koşulu olan bir tür ikili ağaçtır. Bilgisayar biliminde, ikili ağaç, her düğümün sol ve sağına en fazla iki düğüm olan ağaç türüdür. AVL ağacının denge koşulu, düğümlerin yükseklikleri aracılığıyla belirlenmektedir. Her düğümün sağ ve sol tarafındaki düğümlerin yükseklikleri arasındaki fark 1'e eşit olmalı ya da 1'den küçük olmalıdır. Doğada 20 farklı türde amino asit bulunmaktadır. Bu amino asitler AVL ağacına alfabetik olarak yerleştirilmiştir. Örnek verilecek olursa, alfabetik sıraya göre AVL ağacına ilk olarak Alanin (A) amino asidi eklenmiştir. Ardından Arjinin (R) ve Asparajin (N) yerleştirilmiştir. Tüm amino asitler AVL ağacına bu şekilde eklenmiştir. AVL ağacının son durumu ise Şekil 5.1'de verilmiştir.



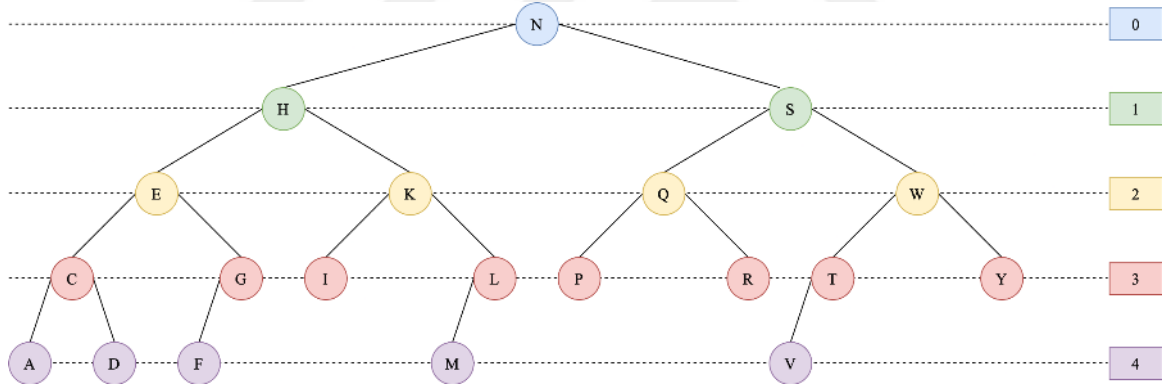
Şekil 5.1. Amino asit kodlarının AVL ağacına eklenmesi

Şekil 5.1'de verilen AVL ağacına göre kök düğümü Asparajin (N) düğümüdür. Amino asit kodları ağaca eklendikçe, ağacın dengesi bozulmaktadır. Ancak dengeleme işlemi AVL ağacının ekleme kurallarına göre yapılmaktadır. AVL ağacında ekleme işlemi için aşağıda verilen kurallar uygulanmaktadır;

- Sağ alt ağacın sol çocuğunda bir dengesizlik olması durumunda, sol-sağ döndürme işlemi yapılmaktadır.
- Sol alt ağacın sol çocuğunda bir dengesizlik olması durumunda, sadece sağa döndürme işlemi yapılmaktadır.
- Sağ alt ağacın sağ çocuğunda dengesizlik olması durumunda, sadece sola döndürme işlemi uygulanmaktadır.

- Sol alt ağacın sağ çocuğunda dengesizlik olması durumunda, sağ-sol döndürme işlemine başvurulmaktadır.

Her bir amino asit tek tek eklendikten sonra eğer denge bozulursa, verilen kurallar doğrultusunda denge yeniden sağlanmaktadır. Histidin (H) ve Serin (S) proteinleri N düğümünün çocuklarıdır. H düğümünün sol çocuğu Glutamik asitken (E), sağ çocuğu ise Lizin (K) amino asididir. S düğümünün sol alt ağacında Glutamin (Q) bulunurken, sağ alt ağacında ise Triptofan (W) yer almaktadır. E düğümünün sol çocuğu Sistein (C) ise, Alanin (A) ve Aspartik asit (D) olmak üzere iki adet çocuğa sahiptir. E düğümünün sağ çocuğu olan Glisin (G) ise Fenilalanin (F) amino asidi dışında başka bir çocuğa sahip değildir. İzolösin (I) ve Lösin (L), kardeş olan bu iki düğüm, K düğümünün çocuklarıdır. Metionin (M), K düğümünün torunu ve L düğümünün ise çocuğudur. Q ve W düğümlerinin ikiye tane çocuğu bulunmaktadır. Prolin (P) ve Arjinin (R) amino asitleri Q düğümünün çocuğu iken, Treoin (T) ve Tirozin (Y) ise W düğümünün çocuklarıdır. T düğümünün ise tek çocuğu Valin (V) amino asididir. AVL ağacına yerleştirildikten sonra, her bir düğümün derinlik değeri hesaplanmıştır. Şekil 5.2’de düğümlerin derinlik değerleri gösterilmiştir.



Şekil 5.2. Her bir amino asidin derinlik değerleri

Sadece kök düğümünün yani N düğümünün derinlik değeri 0’dır. Diğer bütün düğümler farklı derinlik değerlerine sahiptir. Her bir amino asidin/düğümünün derinlik değerleri belirlendikten sonra, amino asit dizilimleri bu değerlere göre sayısal hale getirilmiştir. Tablo 5.1’de proteinlerin AVL değerleri verilmiştir.

Tablo 5.1. Amino asitlerin AVL değerleri

Amino Asit Kodu	AVL Değeri	Amino Asit Kodu	AVL Değeri
A	4	L	3
R	3	K	2
N	0	M	4
D	4	F	4
C	3	P	3
Q	2	S	1
E	2	T	3
G	3	W	2
H	1	Y	3
I	3	V	4

Tablo 5.2’de verilen değerler doğrultusunda, $P(N) = [M, W, F, Y, P, C, \dots]$ şeklindeki bir protein dizilimi, AVL yöntemi ile $M(N) = [4, 2, 4, 3, 3, 3, \dots]$ biçiminde sayısallaştırılır.

5.2. COVID-19 Protein Dizilimlerine Uygulanması

İlk korona virüs vakası Aralık 2019’da Çin’in Wuhan kentinde meydana gelmiş ve hızla dünyanın her bölgesine yayılmıştır. SARS-CoV-2 virüsünün neden olduğu bu hastalığa COVID-19 ismi verilmiştir. Korona virüslerin çoğu hayvanları etkilemektedir, ancak insanların genomik yapısı nedeniyle bu virüsler insanlara da bulaşmaktadır [165]. 2002 ve 2003 yıllarında ortaya çıkan SARS-CoV virüsü 8.000 insanı etkilemiş ve yaklaşık olarak %10’luk bir ölüm oranına erişmiştir. Buna ek olarak, MERS-CoV virüsü ise 2012’de ortaya çıkmış ve en az 2.500 kişiyi etkilemiştir [166]. Bu tür korona virüsler, akciğerlerde enfeksiyona neden olarak ARDS’yi (Akut Solunum Sıkıntısı Sendromu) ortaya çıkarmakta ve ölüme sebebiyet vermektedirler. Yeni tip korona virüsün daha hızlı yayıldığı, kontrol altına alınmasının daha zor olduğu ve pandemi potansiyelinin daha yüksek olduğu gözlemlenmiştir [167, 168]. Her ne kadar enfekte olmuş kişilerin %80’i bu hastalığı hafif atlatsa da COVID-19 diyabet, kanser, kronik rahatsızlıkları ve kalp yetmezliği olan kişiler için ölümcül olabilmektedir. Bu nedenlerden dolayı, SARS-CoV-2 virüsüne karşı yeni stratejiler belirlemek, enfeksiyon aşamasında virüsün konakçı ile nasıl temas ettiği hakkında bilgi sahibi olmak ve yeni ilaçlar geliştirmek ya da mevcut ilaçları yeniden kullanmak, COVID-19 hastalığını anlamada hayati önem taşımaktadır.

Şimdiye kadar SARS-CoV-2 virüsü için herhangi bir ilaç geliştirilmemiştir. Ancak, her türlü tedavi için klinik araştırmalar yapılmakta ve hem RNA hem de protein dizilimleri etkin bir şekilde incelenmektedir [169, 170]. Bu araştırmalardan birisi ise protein-protein etkileşimlerine dayanmaktadır. Proteinler arasında görülen etkileşimler sayesinde ilaç hedefleri belirlenebilmekte

ve uygun tedaviyi sağlayabilecek ilaçlar geliştirilebilmektedir [171, 172]. Bununla beraber, protein etkileşimleri protein fonksiyonlarını belirlemede [173], kanserli hücreleri saptamada [174] ve filogenetik analizinde [175] etkili olabilmektedir. Proteinler arasındaki etkileşimler belirlenirken genellikle deneysel ve hesaplama tabanlı yaklaşımlar kullanılmaktadır. Literatürde proteinler arasındaki etkileşimleri tahmin etmek için deneysel yöntemleri tamamlayıcı birçok hesaplama tabanlı yaklaşımlar bulunmaktadır. Bu yaklaşımlar genel olarak ikili sınıflandırma yaparak tahminde bulunmaktadır. Tahmin işlemi çoğunlukla protein alan bilgilerine, gen ifadelerine, gen komşuluğuna, protein yapı bilgilerine ve filogenetik profillere dayanmaktadır. Fakat, belirli bir ön bilgi mevcut değilse, bu tür yaklaşımlar uygulanamamaktadır. Hesaplama-tabanlı yöntemlerde tahmin işleminin gerçekleştirilebilmesi için, protein çiftlerinin etkileşim bilgilerinin önceden bilinmesi gerekmektedir. Hesaplama-tabanlı yaklaşımlarda, önce genomik dizilimler sayısal hale getirilmekte ve daha sonra makine öğrenmesi ya da derin öğrenme yöntemleriyle sınıflandırılarak protein etkileşimleri tahmin edilmektedir. Deneysel yaklaşımlarla kıyaslandığında, hesaplama-tabanlı yöntemlerin zaman açısından verimli olduğu gözlemlenmiş ve daha az ekipmanla protein etkileşimlerini analiz ettiği belirtilmiştir. Ayrıca, teknolojinin son yıllarda gelişmesiyle birlikte protein dizilim bilgileri de kolaylıkla elde edilebilmektedir. Bu nedenle hesaplama-tabanlı yaklaşımların uygulanması kolay olmaktadır.

Gerçekleştirilen bu çalışmada, SARS-CoV-2 virüsü ile insan proteinleri arasındaki etkileşimleri tahmin etmek için yeni bir sayısallaştırma yöntemi önerilmiştir. Önerilen yöntemin performansı COVID-19 hastalığına ait proteinler üzerinde test edilmiş ve doğrulanmıştır. Çalışmanın ilk aşamasında hem önerilen protein sayısallaştırma yöntemi hem de çeşitli protein sayısallaştırma yöntemleri ile SARS-CoV-2 virüsüne ve insana ait protein dizilimleri sayısal hale getirilmiştir. Ardından, sayısal hale getirilen genomik veriler normalleştirilmiş ve sınıflandırılmak üzere değerlendirilmiştir. Sınıflandırma adımı ise BiRNN (Çift Yönlü RNN) kullanılmıştır. Protein sayısallaştırma yöntemlerinin başarımları doğruluk, kesinlik, duyarlılık, özgünlük, F1-skor ve AUC değerlendirme ölçütleri ile belirlenmiştir. Uygulama sonuçlarından elde edilen bulgular doğrultusunda, önerilen AVL-ağacı protein sayısallaştırma yöntemi ile protein etkileşimleri %97,76 doğruluk ve 0,89 AUC skoru ile tahmin edilmiştir. Önerilen protein sayısallaştırma yönteminin etkili olduğu belirlenmiştir. Protein-protein etkileşimlerini analiz etmek, organizmaların biyolojik aktivitelerini anlamak için büyük önem arz etmektedir. Deneysel yöntemlerden ziyade, hesaplama-tabanlı yöntemlerin kullanılması, protein fonksiyon bilgilerinin, protein ailesi türlerinin ve proteinler arasındaki etkileşimlerin daha hızlı elde edilmesini sağlamaktadır. Bunlara ek olarak, çalışma kapsamında tasarlanmış olan yeni protein sayısallaştırma yöntemi ile literatüre alternatif bir yöntem getirilmiştir. Önerilen yöntem algoritmik-tabanlı bir yöntemdir ve ilk kez protein sayısallaştırma çalışmalarında kullanılmıştır. Gerçekleştirilen çalışmanın ana katkıları şu şekilde ifade edilmektedir;

- Protein dizilimlerini sayısal hale getirmek için yeni bir protein sayısallaştırma yöntemi önerilmiştir.
- Bu çalışma ile ilk kez SARS-CoV-2 virüsü ve insan proteinleri arasındaki etkileşimler hesaplama-tabanlı yaklaşımlar ile tahmin edilmiştir.
- Hesaplama-tabanlı yaklaşımların da en az deneysel yaklaşımlar kadar etkili olduğu gözlemlenmiştir.
- Önerilen algoritmik-tabanlı protein sayısallaştırma yönteminin etkili olduğu test edilmiş ve literatüre kazandırılmıştır.

COVID-19'a neden olan SARS-CoV-2 virüsü incelenmiş ve virüsün proteinleri belirlenmiştir. Virüs membran proteinleri (M), kılıf proteinleri (E), nükleokapsid proteinleri (N) ve spike glukoproteinleri (S) olmak üzere dört yapısal ve altı adet yapısal olmayan (orf3a, orf3b, orf6, orf7a, orf7b ve orf8) proteinlerden oluşmaktadır [176]. Şekil 5.3'te virüsün genom yapısı verilmiştir.

5'UTR	orf1ab	S	orf3a	E	M	orf6a	orf7a	orf7b	orf8	N	orf10	3'UTR
Kodlanmayan bölge	Poliprotein	Glukoprotein	orf3a protein	Kılıf protein	Membran protein	orf6a protein	orf7a protein	orf7b protein	orf8 protein	Nükleokapsid protein	orf10 protein	Kodlanmayan bölge

Şekil 5.3. SARS-CoV-2 virüsünün genomik yapısı

Çalışmada yapısal olmayan protein dizilimleri kullanılmıştır. Çalışma kapsamında COVID-19 ve insan proteinleri arasındaki etkileşimler BioGRID veri setinden elde edilmiştir. Yapısal olmayan proteinlerin kullanılmasının nedeni, bu proteinlerin, virüs genomlarının replikasyonu için gerekli olmasıdır [177]. Buna ek olarak, yapısal olmayan proteinler viral RNA sentezi için gerekli olmaktadır [178]. Bundan dolayı etkileşim ağını tahmin etmek ya da belirlemek için yapısal olmayan proteinlerin analizi büyük önem arz etmektedir. Tablo 5.2'de etkileşim halinde olan SARS-CoV-2 virüsünün proteinleri ve insan proteinleri verilmiştir.

Tablo 5.2. Etkileşim halinde olan proteinler ve toplam sayıları

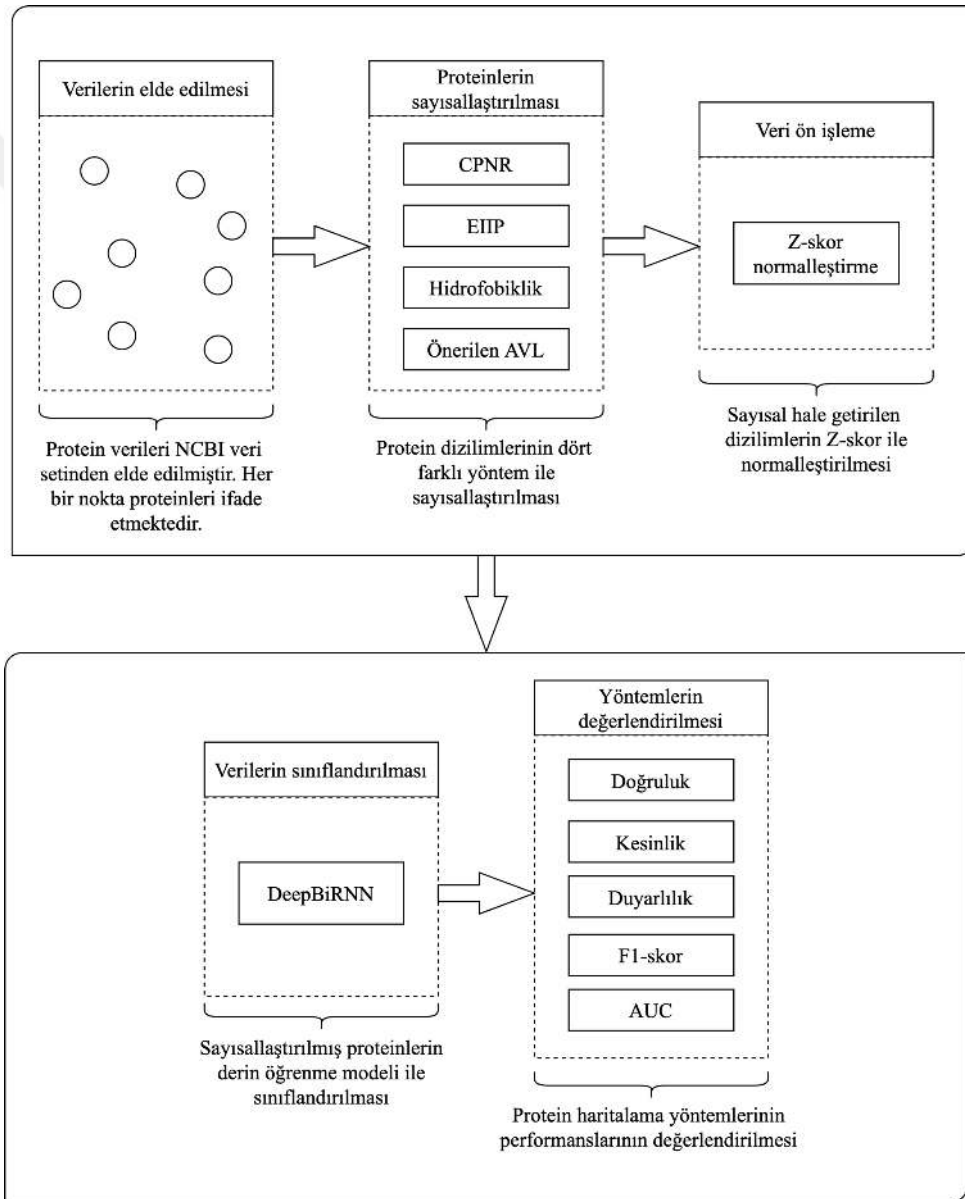
SARS-CoV-2 Proteinleri	Etkileşim Halindeki İnsan Proteinleri	Toplam Protein Sayısı
orf3a	ALG5, ARL6IP6, CLCC1, HMOX1, SUN2, TRIM59, VPS11, VPS39	8
orf3b	STOML2	1
orf6	MTCH1, NUO98, RAE1	3
orf7a	HEATR3, MDN1	2
orf8	ADAM9, ADAMTS1, CHPF2, CISD3, COL6A1, DNMT1, EDEM3, EMC1, ERLEC1, ERO1LB, FBXL12, FKBP7, FKBP10, FOXRED2, GDF15, GGH, HS6ST2, HYOU1, IL17RA, INHBE, ITGB1, KDELC1, KDELC2, LOX, MFGE8, NEU1, NGLY1, NPC2, NPTX1, OS9, PCSK6, PLAT, PLD3, PLEKHF2, PLOD2, POFUT1, PUSL1, PVR, SDF2, SIL1, SMOC1, STC2, TM2D3, TOR1A, UGGT2	47

Orf7b proteini herhangi bir proteinle etkileşim halinde olmadığı için çalışmada kullanılmamıştır. Protein dizilimleri NCBI veri setinden elde edilmiş ve toplamda 3.201 adet protein dizilimi kullanılmıştır. Veri setinin az olmasının nedeni COVID-19 hastalığının yeni bir hastalık olmasıdır. Bu da gerçekleştirilen çalışmanın karşılaştığı olduğu en büyük sorundur. Bu problemin üstesinden gelebilmek için BLAST algoritması kullanılmıştır. Bu algoritma ile protein benzerlik oranı %90 ve üzerinde olan protein dizilimleri değerlendirilmiş ve veri setinin boyutu yükseltilmiştir.

Çalışmada geliştirilen yeni protein sayısallaştırma yöntemi CPNR, EIIP ve hidrofobiklik yöntemleriyle kıyaslanmıştır. Önerilen yöntemin bu yöntemlerden en büyük farkı sayısallaştırma sürecidir. Özetle, kategorik olarak farklılık göstermektedir. EIIP ve hidrofobiklik yöntemleri fizikokimyasal-tabanlı yöntemlerken, CPNR karakter-tabanlı bir yöntemdir. EIIP ve hidrofobiklik yöntemlerinde, protein dizilimleri kimyasal bilgilere dayanarak sayısal hale getirilir. CPNR yönteminde ise belirli bir bilgiye ihtiyaç duyulmadan yalnızca amino asit kodlarına göre sayısallaştırma yapılır. Ancak, önerilen yöntem ise algoritmik-tabanlı bir yöntemdir. Bildiğimiz kadarıyla literatürde daha önceden bu kategoriye ait başka bir protein sayısallaştırma yöntemi bulunmamaktadır. Önerilen yöntem de tıpkı CPNR yöntemi gibi belirli bir spesifik bilgiye dayandırmadan sayısallaştırma yapmaktadır. Ancak, CPNR yönteminin aksine sayısallaştırma süreci belirli bir algoritmik yapıya dayanmaktadır.

Çalışmanın ilk aşamasında insana ait protein dizilimleri NCBI veri setinden elde edilmiştir. Hangi insan proteinin hangi protein ile etkileşim halinde olduğu bilindiği için, virüse ait protein dizilimleri kullanılmamıştır. Örneğin, orf3b proteinin sadece STOML2 proteini ile etkileşim halindedir. STOML2 proteini pozitif etkileşim olarak ifade edilmişken, diğer proteinler negatif etkileşim olarak kabul edilmiştir. Bu yaklaşım SARS-CoV-2 virüsüne ait diğer proteinler için de uygulanmıştır. Ardından, proteinler çeşitli protein sayısallaştırma yöntemleri ile sayısallaştırılmıştır. Sayısallaştırılan proteinler normalleştirilmiş ve bu adımda Z-skor normalleştirme uygulanmıştır. Normalleştirme sürecinin ardından derin öğrenme modeli tasarlanmış

ve tasarlanan derin öğrenme modeli ile sınıflandırma yapılarak, protein etkileşimleri tahmin edilmiştir. Protein etkileşimlerinin tahmin edilmesi genellikle iki senaryoya dayanmaktadır; etkileşim ya vardır ya da yoktur. Özetle ikili-sınıflandırma yapılmaktadır. Virüse ait proteinler ile etkileşim halinde olan insan proteinleri belli olduğu için etkileşen proteinlerin çıktısı 1, diğerleri 0 olarak kabul edilmektedir. Sınıflandırma adımından sonra, protein sayısallaştırma yöntemlerinin başarımları doğruluk, kesinlik, duyarlılık, F1-skor ve AUC değerlendirme ölçütleri ile belirlenmiştir. Geliştirilen uygulamanın akış şeması Şekil 5.4'te verilmiştir.



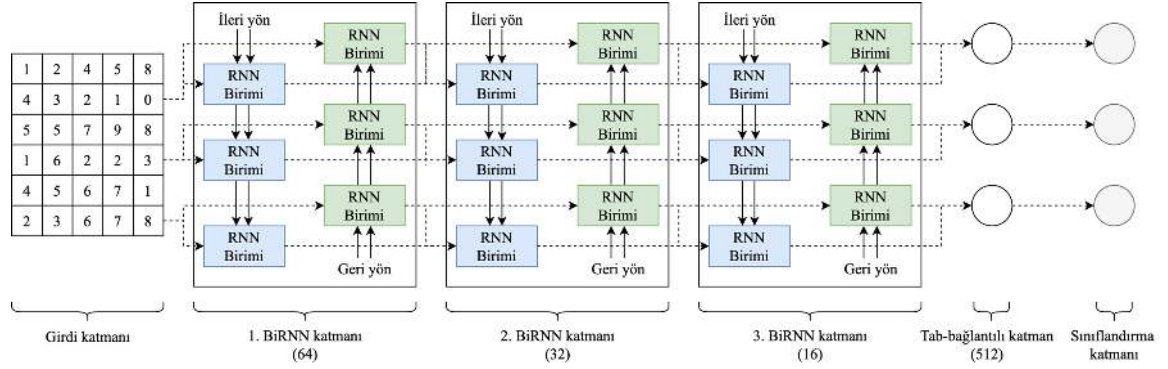
Şekil 5.4. Önerilen AVL yönteminin yapay zekayla değerlendirilme süreci

5.2.1. Derin Öğrenme Kullanılarak Protein Etkileşimlerinin Tahmin Edilmesi

COVID-19 hastalığının teşhisi için bugüne kadar çok sayıda yöntem önerilmiş ve literatürde pek çok sayıda çalışma yapılmıştır. Literatürdeki çalışmalar detaylı bir şekilde incelendiğinde, çoğu çalışmada X-ray görüntülerinin kullanıldığı gözlemlenmiştir. Derin öğrenme algoritmaları bu çalışmalarda sıklıkla kullanılmış ve etkili bir yöntem olarak ön plana çıkmıştır. Bu çalışmalardan elde edilen başarımlar ve derin öğrenmenin popülaritesi, bizi bu çalışmada COVID-19 hastalığının analizi için derin öğrenme yöntemini kullanmaya teşvik etmiştir. Ancak, bu çalışmada hastalığın teşhisinden ziyade virüsün iç yapısı ele alınmış ve virüsün proteinleri analiz edilmiştir. Bu doğrultuda protein-protein etkileşimlerinden yararlanılmış ve yeni bir sayısallaştırma yöntemi önerilmiştir. Derin öğrenme mimarisi olarak ise BiRNN kullanılmıştır. Geliştirilen BiRNN modelinin parametreleri deneme yanılma yaklaşımı ile belirlenmiş ve aşağıda modelin parametreleri verilmiştir;

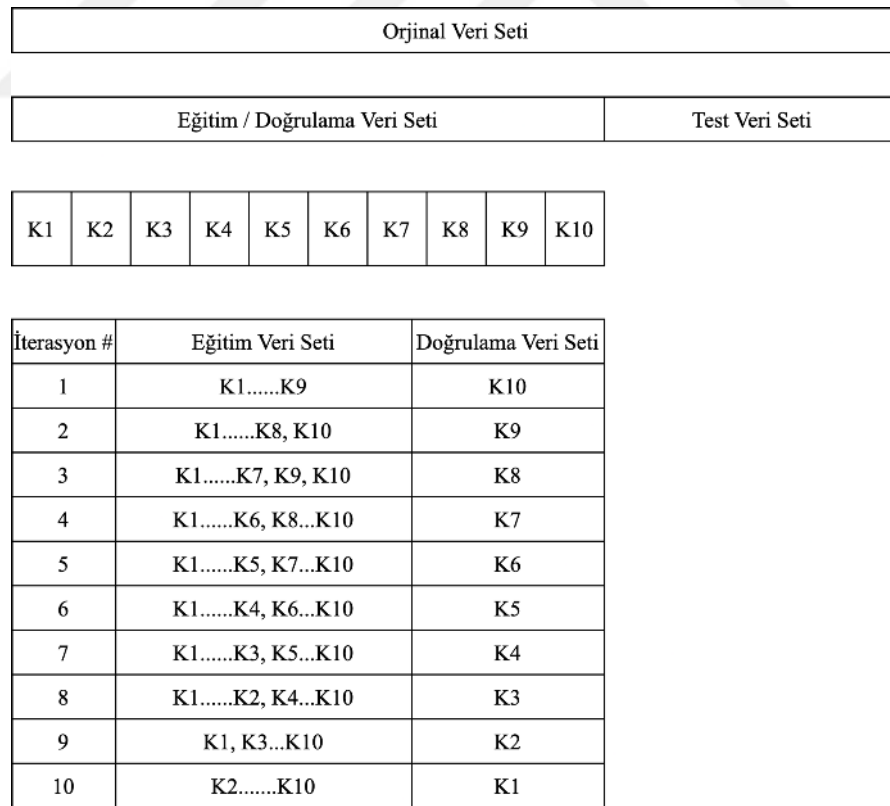
- Birinci katmanda 64 adet BiRNN birimi kullanılmıştır. Çıktı ise ReLU (Rectified Linear Unit – Düzleştirilmiş Doğrusal Birim) aktivasyon fonksiyonu ile hesaplanmıştır.
- İkinci katmanda 32 adet BiRNN birimi değerlendirilmiştir. Bu katmanda ReLU aktivasyon fonksiyonu kullanılmıştır.
- Diğer iki katmanda olduğu gibi bu katmanda da BiRNN birimi kullanılmış ve toplamda 16 adet BiRNN birimi değerlendirmeye alınmıştır. Bu katmanın da çıktısını hesaplamak için ReLU aktivasyon fonksiyonuna başvurulmuştur.
- Ardından düzleştirme işlemi yapılmış ve veriler matris formatına çevrilmiştir.
- Veriler matris formatına geldikten sonra yığın normalleştirme yapılmış ve eş zamanlı öğrenme gerçekleştirilmiştir.
- Düzleştirme ve yığın normalleştirme adımından sonra seyreltme yapılmış ve verilerin %25'i unutulmuştur.
- Daha sonra tam-bağlantılı katman kullanılmış ve bu katmanda 512 adet nöron değerlendirilmiştir.
- Ardından sınıflandırma katmanı oluşturulmuş ve bu katmanda sınıflandırma yapılmıştır.
- En iyileme işlemi için ise SGD (Stochastic Gradient Descent – Skotastik Gradyan İnişi) kullanılmıştır.
- Öğrenme katsayısı ve momentum değeri sırasıyla 0,0001 ve 0,9 olarak belirlenmiştir.
- Modelin kaybı ikili çapraz-entropi ile belirlenmiş ve eğitim işlemi 500 iterasyon ile gerçekleştirilmiştir.

Şekil 5.5'te geliştirilen derin öğrenme modelinin mimarisi verilmiştir.



Şekil 5.5. Geliştirilen BiRNN modelinin yapısı

Çalışmada kullanılan veri seti eğitim, test ve doğrulama olmak üzere 3'e ayrılmıştır. Test veri seti kör veri seti olarak kullanılmış ve protein sayısallaştırma yöntemlerinin başarımları değerlendirilmiştir. Orijinal veri setinin sadece %15'i test veri seti olarak kullanılmıştır. Verilerin geri kalan %85'i (%70 eğitim ve %15 doğrulama) eğitim ve doğrulama için kullanılmış ve 10 katlı çapraz-doğrulamaya başvurulmuştur. Şekil 5.6'da yapılan test işlemi verilmiştir.



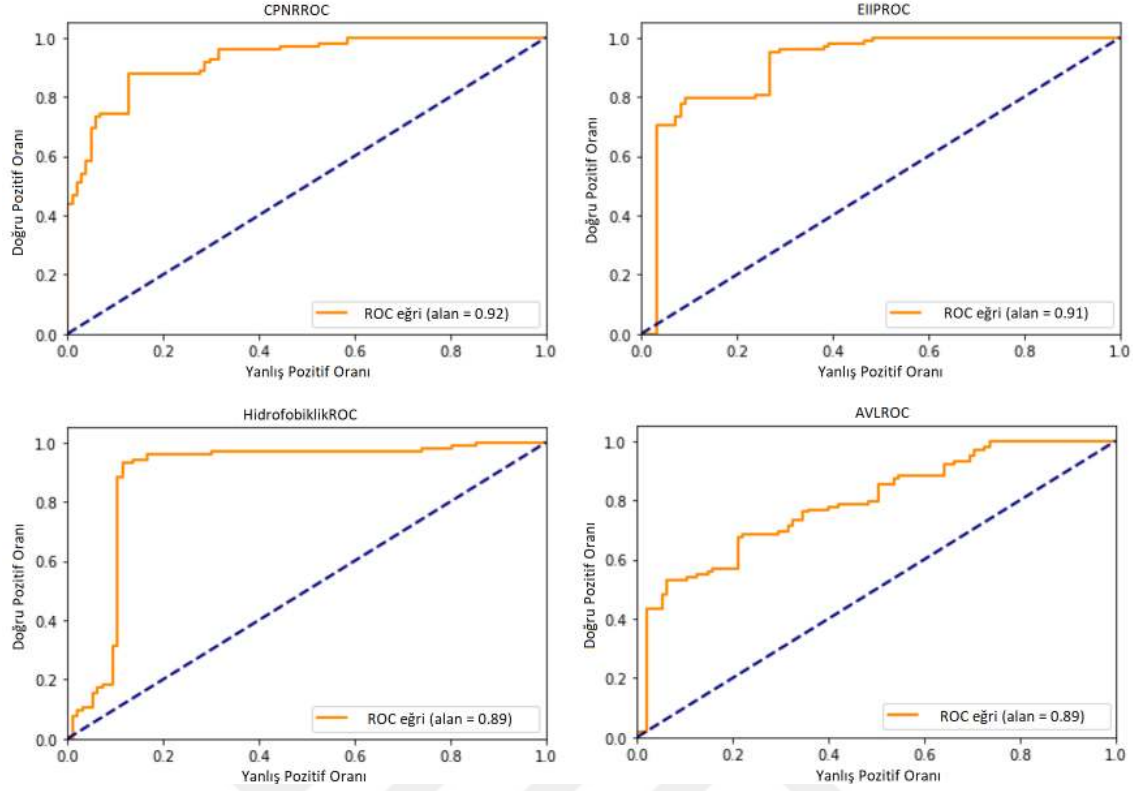
Şekil 5.6. 10 katlı çapraz-doğrulama işleminin adımları

Protein dizilimleri hem önerilen hem de farklı yöntemler ile sayısal hale getirilmiştir. Çalışmada yapısal olmayan protein dizilimleri kullanılmış ve SARS-CoV-2 virüsü ile insan proteinleri arasındaki etkileşimler tahmin edilmiştir. Tablo 5.3'te orf3a proteinin etkileşim sonuçları verilmiştir.

Tablo 5.3. Orf3a proteininin etkileşim sonuçları

Sayısallaştırma Yöntemleri	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1-Skor (%)	AUC
CPNR	83,98	75,41	89,11	81,06	0,92
EIIP	81,44	72,00	83,94	77,37	0,91
Hidrofobiklik	80,19	72,63	84,88	77,88	0,89
Önerilen AVL Yöntemi	81,07	81,77	81,69	81,63	0,89

Tablo 5.3'teki uygulama sonuçlarına göre, bütün protein sayısallaştırma yöntemleri %80'in üzerinde bir doğruluk sergilemiştir. En iyi sınıflandırma işlemi CPNR yöntemi ile gerçekleştirilmiş ve orf3a ile insan proteinleri arasındaki etkileşim ortalama %83,98 doğruluk oranı ile tahmin edilmiştir. Ardından en yüksek doğruluk skoru EIIP protein sayısallaştırma yöntemi ile elde edilmiş ve ortalama doğruluk skoru %81,44 olarak elde edilmiştir. Geliştirilmiş olan AVL protein sayısallaştırma yöntemi ise en az CPNR ve EIIP yöntemleri kadar etkili olmuştur. Önerilen yöntem ile proteinler arasındaki etkileşimler ortalama %81,07 doğruluk skoru ile tahmin edilmiştir. En düşük sınıflandırma performansı ise hidrofobiklik protein sayısallaştırma yöntemi ile elde edilmiş ve ortalama %80,19 doğruluk değeri hesaplanmıştır. Şekil 5.7'de protein sayısallaştırma yöntemlerinin AUC grafikleri verilmiştir.



Şekil 5.7. Orf3a proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri

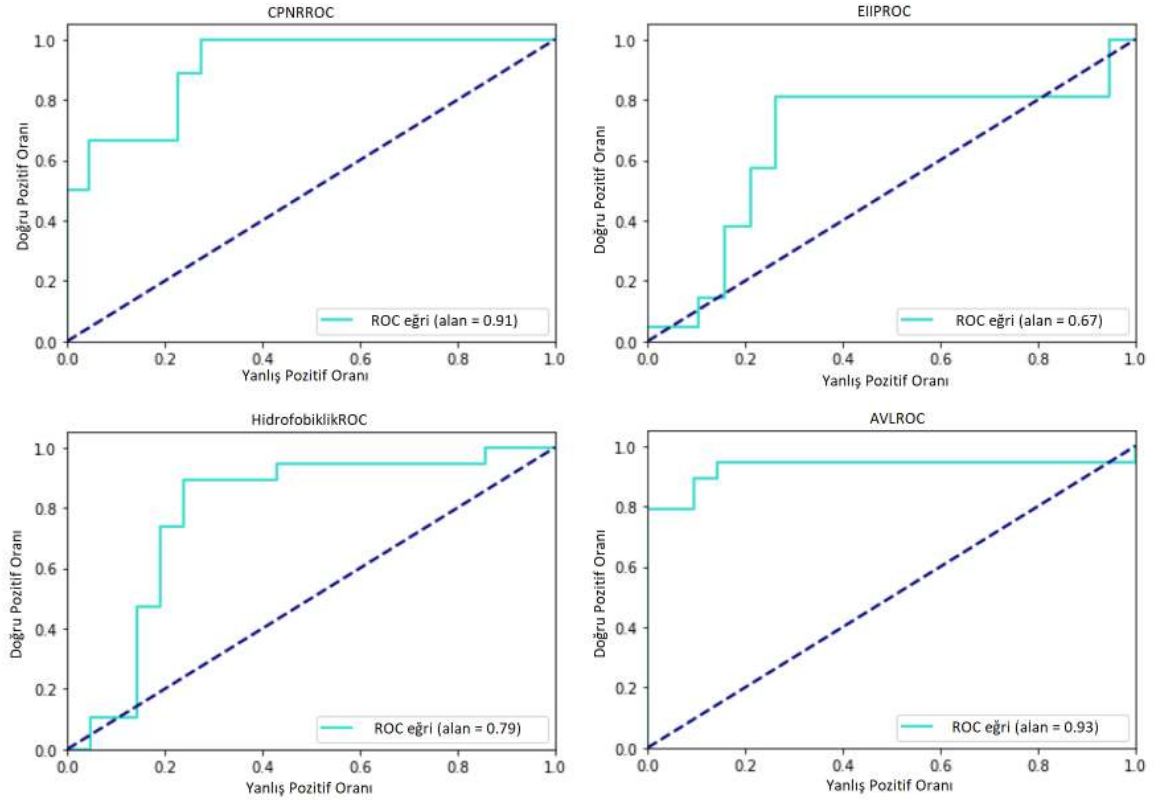
İnsan proteinleri ile etkileşim halinde olan orf3b proteinine ait sınıflandırma sonuçları Tablo 5.4'te verilmiştir.

Tablo 5.4. Orf3b proteininin etkileşim sonuçları

Sayısallaştırma Yöntemleri	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1-Skor (%)	AUC
CPNR	80,65	64,29	69,53	66,52	0,91
EIIP	67,96	47,70	49,18	48,43	0,67
Hidrofobiklik	74,72	79,26	64,06	70,64	0,79
Önerilen AVL Yöntemi	90,65	84,29	89,53	86,52	0,93

Tablo 5.4'te verilen bulgular doğrultusunda, en yüksek sınıflandırma işlemi önerilen AVL yöntemi ile elde edilmiş ve ortalama %90,65 doğruluk elde edilmiştir. Bu sonuç AUC skoruyla da pekiştirilmiş ve en yüksek skor gene önerilen AVL yöntemi ile hesaplanmıştır. Önerilen yöntem en yakın tahmin işlemi ise CPNR protein sayısallaştırma yöntemi ile gerçekleştirilmiştir. Bu yöntem ile ortalama %80,65 doğruluk ve ortalama 0,91 AUC gözlemlenmiştir. Hidrofobiklik yöntemi ise en iyi sınıflandırma yapan üçüncü yöntem olmuştur. Bu yöntem ile proteinler arasındaki etkileşimler ortalama %74,72 doğruluk oranı ile tespit edilmiştir. Ortalama AUC skoru

ise 0,79'dur. En başarısız sınıflandırma işlemi ise EIIP protein sayısallaştırma yöntemi ile yapılmıştır. Bu yöntemin ortalama doğruluk skoru %67,96'yken, AUC skor ise ortalama 0,67 olarak hesaplanmıştır. Şekil 5.8'de protein sayısallaştırma yöntemlerinin AUC grafikleri verilmiştir.



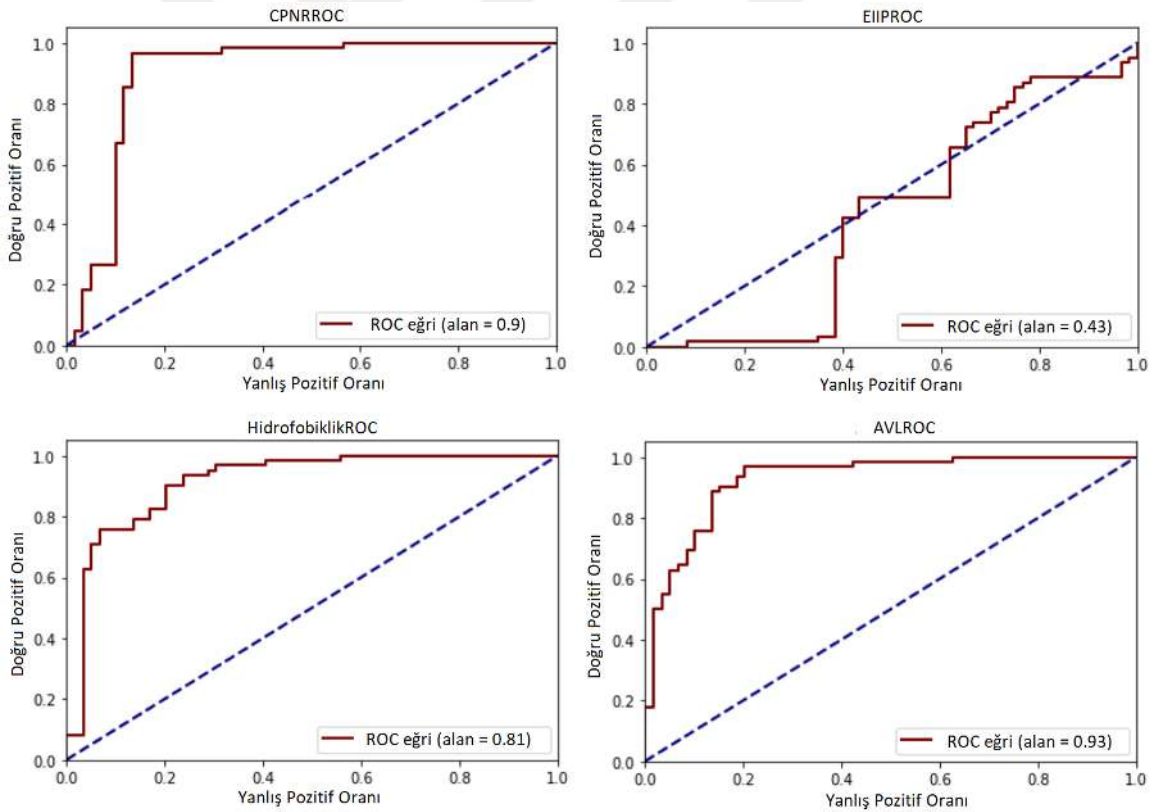
Şekil 5.8. Orf3b proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri

Orf6 proteini ile etkileşim halinde olan insan proteinlerinin sınıflandırma sonuçları Tablo 5.5'te verilmiştir.

Tablo 5.5. Orf6 proteininin etkileşim sonuçları

Sayısallaştırma Yöntemleri	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1-Skor (%)	AUC
CPNR	77,90	66,29	76,78	70,81	0,90
EIIP	41,09	46,80	48,88	47,81	0,43
Hidrofobiklik	69,21	72,25	65,49	68,27	0,81
Önerilen AVL Yöntemi	85,23	72,91	72,31	71,76	0,93

Tablo 5.5’te verilen sonuçlar doğrultusunda, en iyi sınıflandırma işlemi önerilen AVL yöntemi ile elde edilmiş ve ortalama %85,23 doğruluk skoruna ulaşılmıştır. AUC skorlarında ise tekrar en başarılı sınıflandırma işleminin önerilen AVL ile gerçekleştirildiği gözlemlenmiştir. Önerilen yöntem ile en yüksek AUC skoru elde edilmiş ve 0,93 skoruna ulaşılmıştır. İkinci en iyi sınıflandırma işlemi ise CPNR protein sayısallaştırma yöntemi ile yapılmış ve orf6 ile insan proteinleri arasındaki etkileşimler %77,90 doğruluk skoru ile tahmin edilmiştir. Doğruluk skorunda olduğu gibi en yüksek ikinci AUC skoru da CPNR yöntemiyle hesaplanmış ve sonuç 0,90 bulunmuştur. Hidrofobiklik yöntemi ile proteinler arasındaki etkileşimler %69,21 doğruluk ile tahmin edilirken bu oran AUC skorunda 0,81’e kadar yükselmiştir. En başarısız sınıflandırma işlemi ise EIIP protein sayısallaştırma yöntemi ile gerçekleştirilmiştir. Bu yöntem ile protein etkileşimleri ortalama %41,09 doğruluk ve ortalama 0,43 AUC skoru ile sınıflandırılmıştır. Şekil 5.9’da protein sayısallaştırma yöntemlerinin AUC grafikleri verilmiştir.



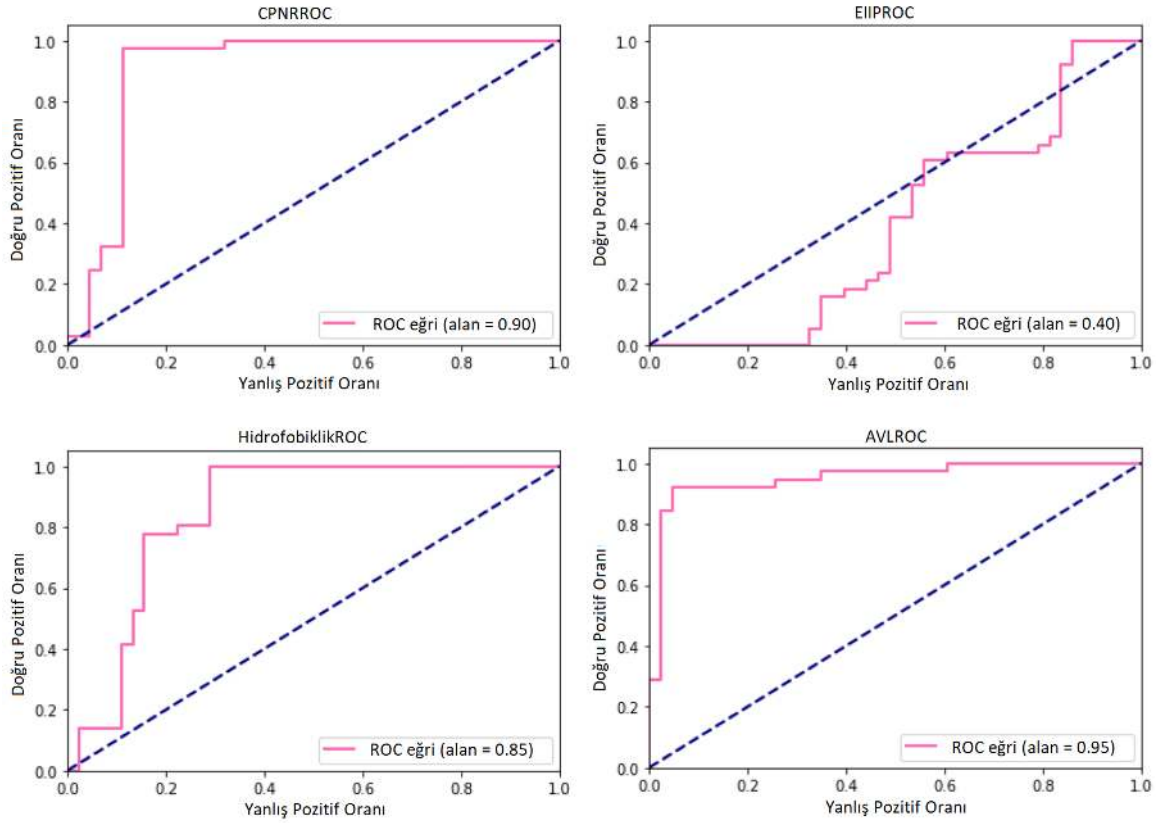
Şekil 5.9. Orf6 proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısallaştırma yöntemlerinin AUC grafikleri

İnsan proteinleri ile etkileşim halinde olan orf7a proteinine ait sınıflandırma sonuçları Tablo 5.6’da verilmiştir.

Tablo 5.6. Orf7a proteininin etkileşim sonuçları

Sayısallaştırma Yöntemleri	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1-Skor (%)	AUC
CPNR	84,57	81,82	81,21	81,33	0,90
EIIP	38,40	40,52	43,85	41,98	0,40
Hidrofobiklik	81,45	79,39	74,02	71,47	0,85
Önerilen AVL Yöntemi	88,91	81,81	93,10	87,09	0,95

Tablo 5.6’da verilen tahmin sonuçlarına göre, önerilen AVL yönteminin en başarılı yöntem olduğu gözlemlenmiştir. AVL yöntemi ile orf7a proteini ve diğer proteinler arasındaki etkileşim ortalama %88,91 doğruluk ve ortalama 0,95 AUC skoru ile tahmin edilmiştir. Diğer protein sayısallaştırma yöntemlerinden ise, en iyi ikinci sınıflandırma işlemi CPNR protein sayısallaştırma yöntemi ile gerçekleştirilmiştir. CPNR yöntemi ile proteinler arasındaki etkileşim ortalama %84,57 doğruluk skoru ile belirlenmiştir. AUC skorlarında ise sonuç ortalama 0,90 olmuştur. Diğer sonuçlarda olduğu gibi, hidrofobiklik yöntemi üçüncü en etkili yöntem olmuştur. Bu yöntem ile ortalama %81,45 doğruluk skoru elde edilirken, ortalama 0,85 AUC skoru hesaplanmıştır. En başarısız sınıflandırma işlemi ise EIIP protein sayısallaştırma yöntemi ile gerçekleştirilmiştir. EIIP protein sayısallaştırma ile ortalama %38,40 doğruluk skoru ile protein etkileşimleri sınıflandırılmıştır. En düşük AUC skoru da EIIP yöntemi ile hesaplanmış ve sonuç ortalama 0,40 olarak hesaplanmıştır. Şekil 5.10’da protein sayısallaştırma yöntemlerinin AUC grafikleri verilmiştir.



Şekil 5.10. Orf7a proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayısalştırma yöntemlerinin AUC grafikleri

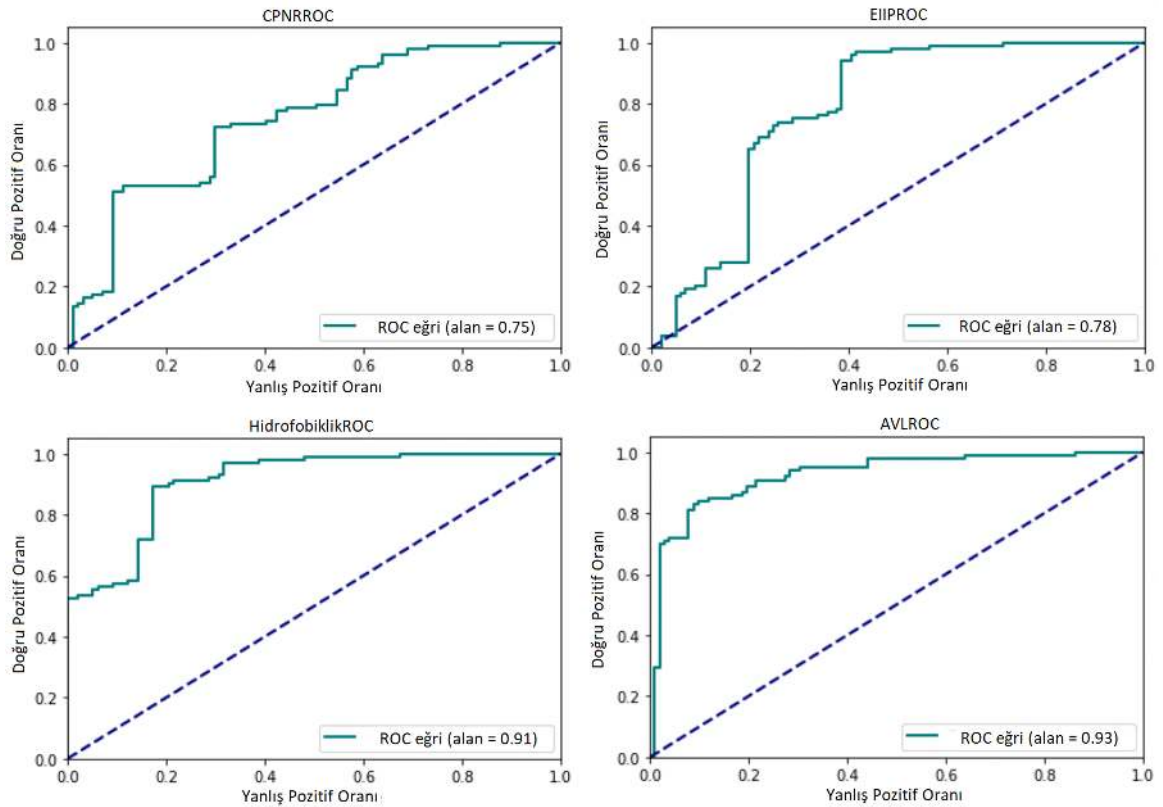
Tablo 5.7’de ise orf8 proteinin insan proteinleri ile kurmuş olduğu etkileşimin skorları verilmiştir.

Tablo 5.7. Orf8 proteininin etkileşim sonuçları

Sayısalştırma Yöntemleri	Doğruluk (%)	Kesinlik (%)	Duyarlılık (%)	F1-Skor (%)	AUC
CPNR	64,09	62,03	59,27	60,15	0,75
EIIP	60,85	57,29	43,29	49,14	0,78
Hidrofobiklik	80,02	73,77	65,28	69,26	0,91
Önerilen AVL Yöntemi	80,78	75,67	66,05	70,11	0,93

Tablo 5.7’de verilen etkileşim skorları incelendiğinde, en başarılı sınıflandırma sürecinin önerilen AVL yöntemi ile gerçekleştirildiği gözlemlenmiştir. Önerilen yöntem ile orf8a proteini ve insan proteinleri arasındaki etkileşimler ortalama %80,78 doğruluk oranı ile tahmin edilmiştir. AUC skorlarına göre sınıflandırma değerlendirildiğinde ise bu oranında en yüksek önerilen yöntem ile elde edildiği belirlenmiştir. Önerilen yöntem ile ortalama 0,93 AUC skoruna ulaşılmıştır. Diğer etkileşimlerin aksine, orf8 proteinin etkileşimlerinde en yüksek ikinci performans ise hidrofobiklik

protein sayıslaştırma yöntemi ile elde edilmiştir. Bu yöntem ile proteinler arasındaki etkileşimler ortalama %80,02 doğruluk ile tahmin edilmiştir. AUC skoruna göre yapılan değerlendirme işleminde ise önerilen AVL yöntemine yakın bir performans sergilemiş ve ortalama 0,91 AUC skoruna erişilmiştir. Yapılan uygulama sonuçlarına göre ise en etkili üçüncü sınıflandırma işlemi CPNR protein sayıslaştırma yöntemi ile elde edilmiştir. CPNR protein sayıslaştırma yöntemi ile orf8 proteini ve insan proteinleri arasındaki etkileşim ortalama %64,09 doğruluk ile tahmin edilmiştir. Bu yöntemin AUC skoru ise ortalama 0,75 olarak bulunmuştur. En etkisiz tahmin işlemi ise EIIP protein sayıslaştırma yöntemi ile yapılmıştır. EIIP protein sayıslaştırma yöntemi ile proteinler arasındaki etkileşimler ortalama %60,85 doğruluk oranı ile sınıflandırılmıştır. AUC skorlarına göre kıyaslama yapıldığında ise EIIP yönteminin CPNR yönteminden daha etkili olduğu gözlemlenmiştir. EIIP yöntemi ile ortalama 0,78 AUC skoru hesaplanmıştır. Şekil 5.11’de protein sayıslaştırma yöntemlerinin AUC grafikleri verilmiştir.



Şekil 5.11. Orf8 proteini ve insan proteinleri arasındaki etkileşimleri tahmin eden protein sayıslaştırma yöntemlerinin AUC grafikleri

Tablo 5.8’de bütün protein sayısallaştırma yöntemlerinin ortalama doğruluk skorları verilmiştir.

Tablo 5.8. Protein sayısallaştırma yöntemlerinin ortalama etkileşim doğruluk skorları

Sayısallaştırma Yöntemleri	Etkileşim Doğruluğu (%)
CPNR	78,24
EIIP	57,95
Hidrofobiklik	77,12
Önerilen AVL Yöntemi	85,33

Tablo 5.8’de verilen bilgiler ışığında, en başarılı sınıflandırma işlemini yapan yöntemin önerilen AVL yöntemi olduğu görülmüştür. Bütün protein etkileşimlerinin AVL yöntemi ile etkili bir şekilde sınıflandırıldığı belirlenmiştir. Ortalama doğruluk skorlarına bakıldığında önerilen AVL yönteminin %85,33 doğruluk değerine ulaştığı gözlemlenmiştir. Ardından CPNR ve hidrofobiklik yöntemleri gelmektedir. İki yöntem de birbirlerine yakın doğruluk skorları üretmiştir. CPNR yöntemi ortalama %78,24 doğruluk ile sınıflandırma yaparken, hidrofobiklik yöntemi ortalama %77,12 doğruluk skoru ile tahmin yapmıştır. En başarısız sınıflandırma işlemi ise EIIP yöntemi ile yapılmış ve genel ortalamaya bakıldığında bu yöntem ile %57,95 doğruluk değeri elde edilmiştir.

5.3. Uygulama Sonuçlarının Değerlendirilmesi

Uygulama sonuçları genel olarak incelendiğinde, bütün protein sayısallaştırma tekniklerinin başarılı olduğu söylenebilir. Ancak, tahmin doğruluğu kullanılan yönteme göre değişiklik göstermiştir. EIIP protein sayısallaştırma yöntemi genellikle aynı fonksiyon bilgisine sahip proteinlerin analizinde etkili olmaktadır [179]. Bunlara ek olarak EIIP yönteminde, fonksiyonel olarak birbirinden farklı olsalar bile amino asitler benzer sayısal ifadeler almaktadır. Bu durumda sayısallaştırma yönteminin dejenerasyona uğramasına sebebiyet vermektedir [180]. COVID-19 yeni bir hastalık olduğu için, proteinlerinin fonksiyonları tam olarak bilinmemektedir. Protein fonksiyonları net olarak bilinemediği için EIIP protein sayısallaştırma yöntemi en başarısız sınıflandırma işlemini gerçekleştirmiş olabilir. Ancak buna rağmen dikkate değer bir tahmin işlemi gerçekleştirilmiştir. Hidrofobiklik protein sayısallaştırma yöntemi fizikokimyasal-tabanlı bir yöntem olduğu için, fizikokimyasal bilgilerin ön planda olduğu çalışmalarda etkili olmaktadır [181]. Fizikokimyasal özellikler protein katlanması çalışmalarında önem arz etmektedir. Ancak, protein katlanmasına dayalı kimyasal özelliklerin tamamı bilinemediğinden, etkili bir fizikokimyasal yöntem geliştirmek problem teşkil etmektedir. Çalışmada fizikokimyasal bilgilerden ziyade yapısal-olmayan genlerin kullanılması hidrofobiklik yönteminin başarısız

olmasına sebebiyet vermiş olabilir. Fizikokimyasal özelliklerin elde edilmesi, yöntemin performansını arttırabilir. En etkili ikinci yöntem ise CPNR protein sayısallaştırma yöntemi olmuştur. Bunun nedeni CPNR yönteminin nokta-mutasyonlarından etkilenmemesi olabilir [182]. Bu başarının bir başka nedeni, CPNR yönteminde protein dizilimlerinin, önerilen yöntemde olduğu gibi belirli bir yapı, protein ve fizikokimyasal bilgilere dayanmadan sayısallaştırılması olabilir. EIIP, CPNR ve hidrofobiklik yöntemlerinden farklı olarak, önerilen AVL protein sayısallaştırma yönteminin en iyi sonucu vermesi şaşırtıcı değildir. Bunun en büyük nedenlerinden biri önerilen yöntemdeki sayısallaştırma işleminin CPNR yönteminde olduğu gibi, belirli bir yapı, kimyasal veya fonksiyon bilgisine dayalı olmamasıdır. CPNR yönteminin de benzer bir yaklaşımla haritalanması ve en iyi ikinci sonucu vermesi bu başarıyı desteklemektedir. Önerilen yöntemin CPNR yönteminden başarılı olmasının nedeni, karakter-tabanlı bir yaklaşım yerine algoritma-tabanlı bir yaklaşım kullanması olabilir. Algoritmik-tabanlı bir yöntem önerilerek, amino asitlerin gen ifadesinden ziyade bilgisayar bilimlerinde sıklıkla kullanılan ağaç yapısına göre sayısallaştırılması amaçlanmıştır. Bu amaçla proteinler hakkında herhangi bir spesifik bilgiye ihtiyaç duymadan kullanılabilir etkili bir yöntem önerilmiştir. Sınıflandırma sonuçları önerilen yöntemin etkili olduğunu göstermiştir. Bunun yanı sıra, önerilen yöntemde de farklı amino asit kodlarının aynı değerleri aldığı gözlemlenmiştir. Tıpkı EIIP ve hidrofobiklik yöntemlerin de olduğu gibi, önerilen yöntemde de bir dejenerasyon söz konusudur. Ancak, önerilen yöntemin geliştirilme amacı hızlı bir analiz sürecini sağlamaktır. Biyoenformatik alanında olduğu gibi, veri sayısının çok olduğu alanlarda, sabit bir $O(c)$ çalışma zamanı elde edilememektedir. Bu durumda en hızlı işlemi yapan $O(\log N)$ zamanında çalışabilecek algoritmik yapılara ihtiyaç duyulmaktadır. Bu nedenle önerilen yöntemde ortalama çalışma zamanı $O(\log N)$ olan AVL ağacı kullanılmıştır. Önerilen yöntem ile diğer yöntemlerin aksine hızlı bir sayısallaştırma işlemi gerçekleştirilmiş ve istenilen hedefe ulaşılmıştır.

Gerçekleştirilen çalışmanın avantajları ve dezavantajları aşağıda verilmiştir;

- Veri sayısı az olsa da hem önerilen sayısal sayısallaştırma yöntemi hem de diğer sayısallaştırma yöntemleri başarılı bir tahmin işlemi gerçekleştirmiştir. Derin öğrenme ile yapılan çalışmalarda veri sayısı önemlidir. Ancak küçük bir veri seti ile de başarı sağlanmıştır.
- SARS-CoV-2 virüsünün protein etkileşimlerinin tahmin edilmesiyle ilaç çalışmaları ivme kazanabilir. Protein etkileşimleri potansiyel ilaçların keşfinde ya da mevcut ilaçların kullanımında önemli bir faktördür.
- Benzer şekilde etkileşimlerin tahmin edilmesiyle, konak hücreler hakkında ayrıntılı bilgiler elde edilebilir.
- Bunlara ek olarak, etkileşimlerin tahmin edilmesiyle, SARS-CoV-2'ye ait proteinlerin fonksiyonları hakkında bilgi sahibi olunabilir.

- Çalışma sırasında, beş farklı yapısal olmayan gen için protein etkileşimleri belirlenmiştir. Gelecekteki çalışmalarda, SARS-CoV-2 virüsüne ait proteinler ile insan proteinleri arasındaki etkileşimler artabilir veya azalabilir. Buna göre, bu çalışmada elde edilen sonuçlar farklılık gösterebilir.
- Bazı durumlarda, makine öğrenmesi algoritmaları, derin öğrenme algoritmalarından daha iyi sonuçlar üretebilmektedir. Çalışmanın sağlıklı olabilmesi için sınıflandırma işleminin makine öğrenmesi algoritmaları ile de gerçekleştirilmesi gerekmektedir.
- Bu çalışmada optimizasyon işlemi yapılmamıştır. Elde edilen sonuçlar belirli optimizasyon algoritmaları ile geliştirilebilir. Bu bağlamda belirli optimizasyon algoritmalarının incelenmesi ve değerlendirilmesi gerekmektedir.
- Protein etkileşimlerini yapay zeka ile tahmin edebilmek için belli bir ön bilgi gerekmektedir. Bu ön bilgiler de deneysel yöntemlerle elde edilmektedir. Günümüzde protein etkileşimlerinin önceden bilgi sahibi olmadan belirlenememesi bu alandaki en büyük sorunlardan birisidir.

Çalışmanın sonuna önerilen yöntemin etkili olduğu gözlemlenmiştir. Bu başarımın ana nedenlerinden birisi, protein dizilimlerindeki amino asit kodlarının ağaç üzerine dinamik bir şekilde yerleştirilmesi ve algoritmik bir şekilde ifade edilmesi olabilir. Algoritmanın sözlük yapısını ve AVL ağacının denge unsurunun dikkate alınması yöntemin daha kararlı ve sağlam olmasına neden olmuştur. Bu sayede algoritmik-tabanlı sayısallaştırma yönteminin diğer kategorilerdeki yöntemlerden daha başarılı olabileceği gösterilmiştir. Uygulama sonuçlardan da anlaşılacağı gibi, önerilen yaklaşım ile protein etkileşimleri öznitelik çıkarımı yapılmadan başarılı bir şekilde tahmin edilmiştir. Ayrıca, önerilen sayısallaştırma yönteminin potansiyel protein-protein etkileşimlerini doğru etkili bir şekilde tahmin ettiği gözlemlenmiştir. Sonuçlar, önerilen algoritmik-tabanlı protein sayısallaştırma yönteminin PPE (Protein-Protein Etkileşimi) çalışmalarında uygulanabilirliğini ve üstünlüğünü göstermiştir. Gelecek çalışmalarda farklı algoritmik-tabanlı yöntemler önerilebilir ve bu yöntemlerin diğer kategorilere ait yöntemlerle kıyaslanması algoritmik-tabanlı yaklaşımların performansı daha ayrıntılı bir şekilde ortaya koyabilir. Ayrıca, farklı derin öğrenme yaklaşımları kullanmak ya da farklı alanlarda protein analizi yapmak, önerilen protein sayısallaştırma yönteminin performansı hakkında daha sağlam ve sağlıklı bilgiler sağlayabilir. Önerilen yöntem,

- İlaç-hedef etkileşimlerinin belirlenmesi,
- İlaç terapi ve ilaç geliştirme çalışmaları,
- Protein ailelerinin sınıflandırılması,
- Filogenetik analizi çalışmaları,
- Viral-konak protein etkileşimlerinin tahmin edilmesi,
- Kanseri-protein etkileşimlerini belirlenmesi,
- Protein fonksiyonlarının tahmin edilmesi,

- Protein yapı ailelerinin sınıflandırılması gibi çeşitli farklı alanlarda kullanılabilir.

Gelecekteki çalışmalarda önerilen yöntem, bahsedilen farklı protein çalışmaları üzerinde kullanılacak ve test edilecektir.



6. ENTROPİ TABANLI YENİ BİR SAYISALLAŞTIRMA YÖNTEMİ

Tezin bu bölümünde, mevcut protein sayısallaştırma yöntemlerine ek olarak, Shannon entropisine dayalı yeni bir protein sayısallaştırma yöntemi önerilmiştir. Önerilen bu yöntem, viral genomlardaki protein etkileşimlerini sınıflandırmak için kullanılmış ve yöntemin literatürdeki diğer protein sayısallaştırma yöntemleri ile kıyaslanması gerçekleştirilerek başarımı belirlenmiştir. Önerilen protein sayısallaştırma yöntemi Shannon entropisine dayanmaktadır. Shannon entropisi günümüzde verilerin sıkıştırılması ve yorumlanması için bilgisayar bilimlerinde sıklıkla kullanılmaktadır. Önerilen yöntem bu algoritmik yapısı nedeniyle yeni bir kategorisel yaklaşım olarak belirlenen algoritmik-tabanlı protein sayısallaştırma kategorisinde değerlendirilmiştir.

6.1. Önerilen Sayısallaştırma Yöntemi

Entropi kavramı ilk kez 19.yüzyılda ortaya çıkmıştır. Entropi, düzensizliği, dağınıklığı ya da bir sistemin belirsizliğini ifade etmektedir. İlk başlarda matematik ve fizik alanlarında sıklıkla kullanılan entropi, zamanla diğer fen bilimleri ve sosyal bilimlerde de kullanılmıştır. Günümüzde istatistik, jeopolitik, mekanik, siyaset, sosyoloji, görüntü işleme, sinyal işleme ve genomik işaret işleme alanlarında da değerlendirilmektedir. Entropi, herhangi bir frekansta bulunan zaman serilerindeki tutarsızlıkları hesaplamak için kullanılmaktadır. Veri düzgün ve geniş bir olasılığa sahip olduğunda entropi değeri yüksek olurken, veri dağınık ve dar bir olasılığı sahipse entropi değeri düşük olmaktadır. Günümüzde, Shannon, eşik, norm, log enerji ve Sure olmak üzere çeşitli türlerde entropi bulunmaktadır. Önerilen bu yöntem Shannon entropisine dayanmaktadır. Shannon entropisi genellikle bir sinyalde görülen bozuklukları ölçmek ve sinyal hakkında genel bilgiler elde etmek için kullanılmaktadır. Denklem 6.1’de verilen formüle göre hesaplanmaktadır;

$$S = - \sum_i p(x_i) \log(p(x_i)) \quad (6.1)$$

Denklem 6.1’de x değişkeni zaman serisini ifade ederken, p ise zaman serisinde bulunan her bir değişkenin olasılığını göstermektedir. x_i ise zaman serisinde bulunan i . değeri belirtmektedir. Entropi uygulamaları ve bilgi teorisi çalışmaları, entropinin genetik dizilimlerin analizi ve evrimi için etkili bir yol olduğunu kanıtlamıştır [183]. Genel olarak ise Shannon entropi kullanılmakta ve bu entropi ile genom dizilimlerinin bağlanma sahaları tahmin edilmekte ve tanımlanmaktadır [184, 185]. Buna ek olarak Shannon entropi ile türler arasında genom bilgileri de elde edilmekte [184], DNA’da bulunan ekzon ve intron bölgeleri belirlenmekte [185], temizlik genleri ve belirli genler arasındaki farklar etkili bir şekilde saptanmaktadır [186]. Bunlara ek olarak, Shannon entropisi ile biyomedikal sinyallerden etkili bilgiler elde edilebilmektedir [187]. Biyomedikal ve genomik sinyal işleme alanlarında elde edilen bu başarımlardan dolayı, çalışma kapsamında Shannon

entropisine dayalı protein sayısallaştırma yöntemi önerilmiş ve kullanılmıştır. Shannon entropisi etkili bir yöntem olmasına rağmen bazı durumlarda kullanımı uygun olmamaktadır. Bazı çalışmalar, bilgi kaybı olabileceğinden entropi ile sayısallaştırmanın yetersiz olabileceğini belirtmektedir [188, 189]. Entropinin etkili olduğu çalışmalar genellikle DNA ve genom analizine dayanmaktadır. Protein sayısallaştırma çalışmalarında bu etki kaybolmaktadır. Bu nedenle, önerilen protein sayısallaştırma yöntemi, orijinal Shannon entropisi yerine, [190] numaralı çalışmada verilen değiştirilmiş Shannon entropiye dayanarak gerçekleştirilmiştir. Değiştirilmiş Shannon entropisi, Shannon entropisinin kesirli bir türevidir. Denklem 6.2’de değiştirilmiş Shannon entropisinin formülü verilmiştir [190];

$$S_f = - \sum_i p((x_i))^{a_i} p(x_i) \log(p(x_i)) \quad (6.2)$$

Denklem 6.2’de, i değeri amino asidin dizilimdeki konumunu ifade etmektedir. $p(x_i)$ ise her bir amino asidin tekrarlanma frekansını göstermektedir. Örnek olarak, $S(N) = [MAKDYY]$ şeklinde verilen bir protein diziliminde M amino asidinin i değeri 1’dir. Çünkü M dizilimdeki ilk amino asittir. A amino asidinin i değeri 2, K amino asidinin i değeri ise 3’tür. D amino asidinin i değeri 4’ken Y amino asidinin iki farklı i değeri vardır. Bu değerler 5 ve 6’dır. M amino asidinin tekrarlanma frekansı $i=1$ için $p(x_1)$ hesabı ile gerçekleştirilir çünkü dizilimdeki ilk amino asit M amino asidir. D amino asidinin tekrarlanma frekansı ise $i=4$ için $p(x_4)$ hesabı yapılarak elde edilir. Bu verilen hesaplamalar doğrultusunda, verilen örnekteki amino asitlerin tekrarlanma frekansları aşağıda verilmiştir;

$$i = 1 \text{ için, } p(x_1) = P(M) = \frac{1}{6} = 0,1667$$

$$i = 2 \text{ için, } p(x_2) = P(A) = \frac{1}{6} = 0,1667$$

$$i = 3 \text{ için, } p(x_3) = P(K) = \frac{1}{6} = 0,1667$$

$$i = 4 \text{ için, } p(x_4) = P(D) = \frac{1}{6} = 0,1667$$

$$i = 5 \text{ için, } p(x_5) = P(Y) = \frac{2}{6} = 0,3333$$

$$i = 6 \text{ için, } p(x_6) = P(Y) = \frac{2}{6} = 0,3333$$

Tekrarlanma frekansının hesaplanması adımından da anlaşılacağı üzere, aynı amino asitlerin hepsinin tekrarlanma sıklığının hesaplanmasına gerek olmamaktadır. Beşinci ve altıncı sıradaki amino asitler aynı oldukları için, tekrarlanma frekansları da aynı olmaktadır. Bundan dolayı her bir amino asit için tekrarlanma frekansını bir kere hesaplamak yeterli olmaktadır. Denklem 6.3’te α

değerinin hesaplanması verilmiştir. Bu değer 1 sayısının her amino asidin tekrarlanma frekansının logaritmasına oranı ile belirlenmektedir [191].

$$\alpha_i = \frac{1}{\log(p(x_i))} \quad (6.3)$$

Denklem 6.3'ten de anlaşılacağı üzere her bir amino asidin α değeri bulunmaktadır. Bu değer amino asitlerin tekrarlanma frekansına göre değişiklik göstermektedir. Bunlara ek olarak, α değeri dizilimden dizilime göre de değişiklik gösterebilmektedir. Bu bilgiler doğrultusunda önerilen protein sayısalştırma yönteminin çalışma adımları aşağıda verilmiştir;

1. İlk olarak elde edilen protein diziliminin uzunluğu belirlenir. $P(S) = [\text{MAKATGRNNLVSPKK}]$ protein dizilimi örnek olarak verilsin. Bu protein diziliminin uzunluğu 15'tir.
2. Ardından her bir protein diziliminin tekrarlanma frekansı belirlenir. Söz konusu örnek dikkate alındığında her bir amino asidinin tekrarlanma frekansı şu şekilde elde edilmektedir;

$$i = 1 \text{ için, } p(x_1) = P(M) = \frac{1}{15} = 0,066$$

$$i = 2 \text{ için, } p(x_2) = P(A) = \frac{2}{15} = 0,133$$

$$i = 3 \text{ için, } p(x_3) = P(K) = \frac{3}{15} = 0,200$$

$$i = 4 \text{ için, } p(x_4) = P(A) = \frac{2}{15} = 0,133$$

$$i = 5 \text{ için, } p(x_5) = P(T) = \frac{1}{15} = 0,066$$

$$i = 6 \text{ için, } p(x_6) = P(G) = \frac{1}{15} = 0,066$$

$$i = 7 \text{ için, } p(x_7) = P(R) = \frac{1}{15} = 0,066$$

$$i = 8 \text{ için, } p(x_8) = P(N) = \frac{2}{15} = 0,133$$

$$i = 9 \text{ için, } p(x_9) = P(N) = \frac{2}{15} = 0,133$$

$$i = 10 \text{ için, } p(x_{10}) = P(L) = \frac{1}{15} = 0,066$$

$$i = 11 \text{ için, } p(x_{11}) = P(V) = \frac{1}{15} = 0,066$$

$$i = 12 \text{ için, } p(x_{12}) = P(S) = \frac{1}{15} = 0,066$$

$$i = 13 \text{ için, } p(x_{13}) = P(P) = \frac{1}{15} = 0,066$$

$$i = 14 \text{ için, } p(x_{14}) = P(K) = \frac{3}{15} = 0,200$$

$$i = 15 \text{ için, } p(x_{15}) = P(K) = \frac{3}{15} = 0,200$$

3. Ardından Denklem 6.3'te verilen formül doğrultusunda her bir amino asidin α değeri hesaplanır. Amino asitlerin α değerleri aşağıda verilmiştir;

$$i = 1 \text{ için, } \alpha_1 = \frac{1}{\log(p(x_1))} = \frac{1}{\log(p(M))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 2 \text{ için, } \alpha_2 = \frac{1}{\log(p(x_2))} = \frac{1}{\log(p(A))} = \frac{1}{\log(0,133)} = 0,496$$

$$i = 3 \text{ için, } \alpha_3 = \frac{1}{\log(p(x_3))} = \frac{1}{\log(p(K))} = \frac{1}{\log(0,200)} = 0,621$$

$$i = 4 \text{ için, } \alpha_4 = \frac{1}{\log(p(x_4))} = \frac{1}{\log(p(A))} = \frac{1}{\log(0,133)} = 0,496$$

$$i = 5 \text{ için, } \alpha_5 = \frac{1}{\log(p(x_5))} = \frac{1}{\log(p(T))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 6 \text{ için, } \alpha_6 = \frac{1}{\log(p(x_6))} = \frac{1}{\log(p(G))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 7 \text{ için, } \alpha_7 = \frac{1}{\log(p(x_7))} = \frac{1}{\log(p(R))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 8 \text{ için, } \alpha_8 = \frac{1}{\log(p(x_8))} = \frac{1}{\log(p(N))} = \frac{1}{\log(0,133)} = 0,496$$

$$i = 9 \text{ için, } \alpha_9 = \frac{1}{\log(p(x_9))} = \frac{1}{\log(p(N))} = \frac{1}{\log(0,133)} = 0,496$$

$$i = 10 \text{ için, } \alpha_{10} = \frac{1}{\log(p(x_{10}))} = \frac{1}{\log(p(L))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 11 \text{ için, } \alpha_{11} = \frac{1}{\log(p(x_{11}))} = \frac{1}{\log(p(V))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 12 \text{ için, } \alpha_{12} = \frac{1}{\log(p(x_{12}))} = \frac{1}{\log(p(S))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 13 \text{ için, } \alpha_{13} = \frac{1}{\log(p(x_{13}))} = \frac{1}{\log(p(P))} = \frac{1}{\log(0,066)} = 0,368$$

$$i = 14 \text{ için, } \alpha_{14} = \frac{1}{\log(p(x_{14}))} = \frac{1}{\log(p(K))} = \frac{1}{\log(0,200)} = 0,621$$

$$i = 15 \text{ için, } \alpha_{15} = \frac{1}{\log(p(x_{15}))} = \frac{1}{\log(p(K))} = \frac{1}{\log(0,200)} = 0,621$$

4. Hem tekrarlanma frekansları hem de α değerleri hesaplandıktan sonra Denklem 6.1'de verilen formül doğrultusunda amino asitler sayısal hale getirilir. Amino asitlerin sayısal ifadeleri aşağıda verilmiştir;

$$i = 1, - \sum_1 p((x_1))^{a_1} p(x_1) \log(p(x_1)) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 2, - \sum_2 p((x_2))^{a_2} p(x_2) \log(p(x_2)) = (0,133)^{0,496} (0,133) \log(0,133) = 0,09$$

$$i = 3, - \sum_3 p((x_3))^{a_3} p(x_3) \log(p(x_3)) = (0,200)^{0,621} (0,200) \log(0,200) = 0,12$$

$$i = 4, - \sum_4 p((x_4))^{a_4} p(x_4) \log(p(x_4)) = (0,133)^{0,496} (0,133) \log(0,133) = 0,09$$

$$i = 5, - \sum_5 p((x_5))^{a_5} p(x_5) \log(p(x_5)) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 6, - \sum_6 p((x_6))^{a_6} p(x_6) \log(p(x_6)) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 7, - \sum_7 p((x_7))^{a_7} p(x_7) \log(p(x_7)) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 8, - \sum_8 p((x_8))^{a_8} p(x_8) \log(p(x_8)) = (0,133)^{0,496} (0,133) \log(0,133) = 0,09$$

$$i = 9, - \sum_9 p((x_9))^{a_9} p(x_9) \log(p(x_9)) = (0,133)^{0,496} (0,133) \log(0,133) = 0,09$$

$$i = 10, - \sum_{10} p((x_{10}))^{a_{10}} p(x_{10}) \log(p(x_{10})) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 11, - \sum_{11} p((x_{11}))^{a_{11}} p(x_{11}) \log(p(x_{11})) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 12, - \sum_{12} p((x_{12}))^{a_{12}} p(x_{12}) \log(p(x_{12})) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 13, - \sum_{13} p((x_{13}))^{a_{13}} p(x_{13}) \log(p(x_{13})) = (0,066)^{0,368} (0,066) \log(0,066) = 0,07$$

$$i = 14, - \sum_{14} p((x_{14}))^{a_{14}} p(x_{14}) \log(p(x_{14})) = (0,200)^{0,621} (0,200) \log(0,200) = 0,12$$

$$i = 15, - \sum_{15} p((x_{15}))^{a_{15}} p(x_{15}) \log(p(x_{15})) = (0,200)^{0,621} (0,200) \log(0,200) = 0,12$$

5. Gerekli işlemlerin ardından sayısallaştırma tamamlanır. Aşağıda önerilen protein sayısallaştırma yöntem ile sayısal hale getirilmiş protein dizilimi verilmiştir;

$$P(S) = [0,07 \ 0,09 \ 0,12 \ 0,09 \ 0,07 \ 0,07 \ 0,07 \ 0,09 \ 0,09 \ 0,07 \ 0,07 \ 0,07 \ 0,07 \ 0,12 \ 0,12]$$

Proteinler sayısal hale getirildikten sonra, Z-skor normalizasyonu uygulanmış ve sayısallaştırma işlemi sonuçlandırılmıştır. Sayısallaştırma ve normalizasyon adımının ardından

protein dizilimlerine AFD (Ayrık Fourier Dönüşümü) uygulanmıştır. AFD, sinyal işleme uygulamalarında sıklıkla kullanılmakta ve bir sinyalin özniteklerini elde etmek için sinyali zaman uzayından frekans uzayına dönüştürmektedir. Protein sayısallaştırma yöntemlerinin dağılımını analiz edebilmek için protein dizilimleri boyunca kayan bir pencere kullanılmakta ve her bir pencere için AFD uygulanmaktadır. AFD, DNA dizilimlerinin analizi [192], DNA ve protein kodlama bölgelerinin tahmini [193, 194], protein fonksiyonlarının belirlenmesi [195], protein dizilimlerinin filogenetik analizi [196] başta olmak üzere çeşitli genomik sinyal işleme çalışmalarında etkili bir şekilde kullanılmaktadır. Orijinal protein dizilimleri ters Fourier dönüşümü ile tekrar oluşturulabildiği için, protein dizilimlerine ait bilgiler Fourier katsayıları ile elde edilebilmektedir. Bunlara ek olarak, orijinal protein dizilimi ile dönüşüme uğramış dizilim arasında büyük benzerlikler görülmektedir. Bu nedenlerden dolayı, AFD kullanılmış ve birlikte evrim analizi ile proteinler arasındaki etkileşimler değerlendirilmiştir. Dönüştürme işleminin ardından proteinler arasındaki Öklid uzaklığı hesaplanmış ve uzaklık matrisi oluşturulmuştur. İki protein diziliminin AFD değeri dizilim uzunluğuna bağlı olduğu için, protein dizilimlerinin Öklid mesafesini ya da Fourier katsayılarını doğrudan hesaplamak ve kıyaslamak zor olmaktadır. Sinyal işleme çalışmalarında bu problemi ortadan kaldırmak için genellikle kısa olan sinyalin sonuna sıfır eklenmektedir [197]. Bu durum genomik dizilimler içinde uygulanmakta ve bütün dizilimlerin uzunluğu eşitlenmektedir. Dizilimlerin hepsinin aynı uzunlukta olması uzaklık matrisinin kolay bir şekilde oluşturulmasını sağlamaktadır. Gerçekleştirilen çalışmada da bu yaklaşım kullanılmış ve kısa olan dizilimlerin sonuna sıfır eklenmiştir. Her bir protein dizilimi için Öklid uzaklık matrisi oluşturulduktan sonra, Denklem 6.4'te verilen formül ile Pearson korelasyon katsayıları hesaplanmıştır;

$$P_{kor} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (6.4)$$

Denklem 6.4'te verilen, X ve Y protein dizilimlerini ifade ederken, $\bar{X}$ ve $\bar{Y}$ ise ortalama değerleri belirtmektedir. Pearson korelasyonu iki nokta ya da değişken arasındaki istatistiksel ilişkiyi ölçmek için kullanılmaktadır. Yaygın olarak kullanılan bir yöntemdir. Hesaplama işlemi kovaryansa bağlı olduğu için değişkenler arasındaki etkileşimi değerlendirmek için kullanılan en iyi yöntemdir [198]. Korelasyon katsayıları -1 ve +1 arasında değerler almaktadır. 1 değeri pozitif ilişkiyi ifade ederken, -1 değeri negatif ilişkiyi temsil etmektedir. Değerin 0 olması ise ilişkinin olmadığını göstermektedir. Pearson korelasyon katsayılarının ilişkileri şu şekilde ifade edilmektedir [199];

- Korelasyon yok: 0
- Çok düşük derece korelasyon: 0.10 – 0.19
- Düşük derece korelasyon: 0.20 – 0.39
- Orta derece korelasyon: 0.40 – 0.59

- Yüksek derece korelasyon: 0.60 – 0.79
- Mükemmel derece korelasyon: 0.80 – 1.00

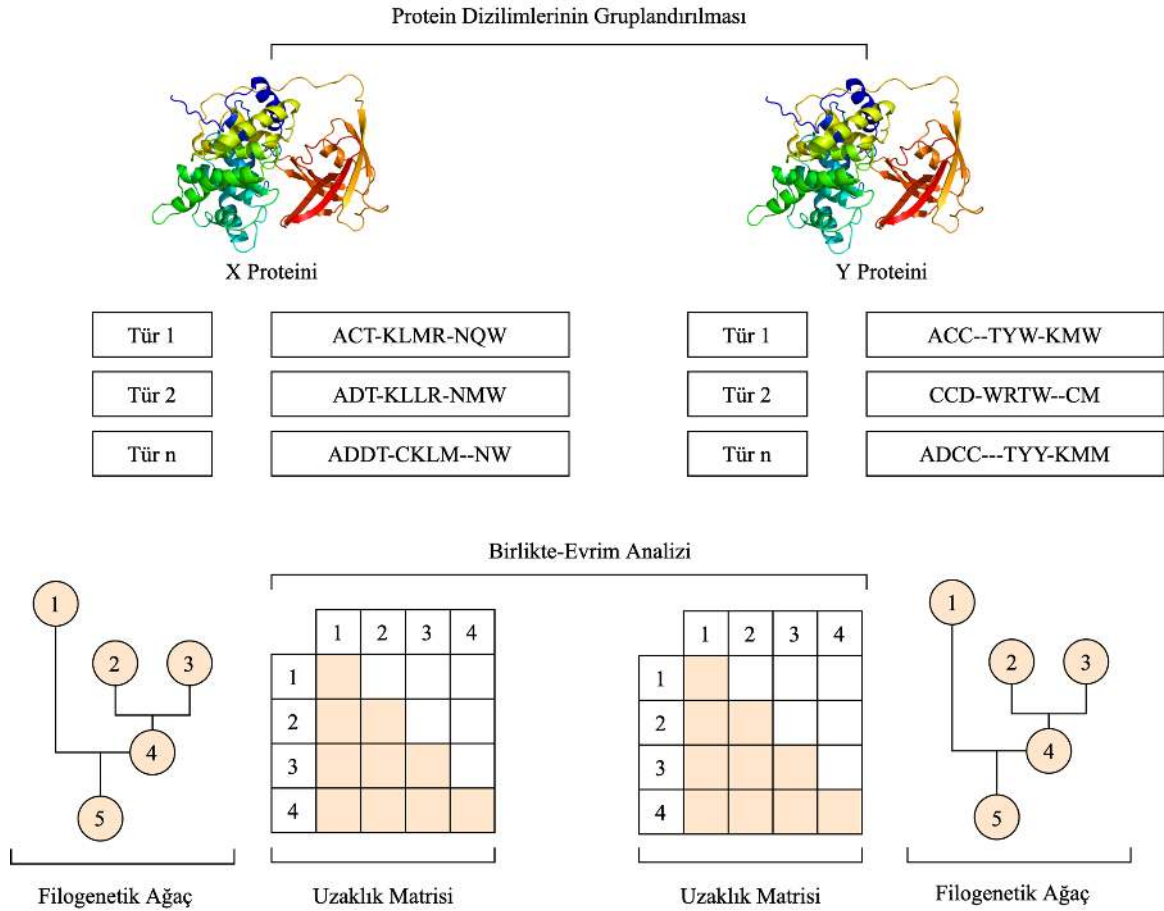
Çalışmada bu katsayılara bakarak birlikte evrim analizi yapılarak viral genomlar arasındaki protein etkileşimleri tahmin edilmiştir. Önerilen Shannon entropisine dayalı protein sayısallaştırma yönteminin diğer protein sayısallaştırma yöntemlerinden farkı şu şekilde ifade edilebilir;

- Önerilen protein sayısallaştırma yöntemi, mevcut protein sayısallaştırma yöntemlerinden farklı bir kategoriye aittir. Mevcut yöntemlerden bazıları yapı-tabanlı, evrim-tabanlı, fizikokimyasal-tabanlı vb. yöntemlerdir. Ancak önerilen yöntem algoritmik-tabanlı bir yöntemdir.
- Mevcut yöntemlerde, protein sayısallaştırma işlemi belirli bir protein bilgisine dayanmaktadır. Örneğin, hidrofobiklik yönteminde proteinlerin fizikokimyasal bilgilerine ihtiyaç duyulmaktadır. Benzer şekilde, Miyazawa enerjileri proteinlerin yapı bilgilerini kullanmaktadır. Ancak önerilen Shannon entropisine dayalı protein sayısallaştırma yönteminde ise belirli bir protein bilgisi gerekmemektedir. Yalnızca amino asitlerin tekrarlanma frekansına ve entropi hesaplamasına dayanmaktadır. Bu sayede protein dizilimleri belirli bir bilgiye dayanmadan sayısal hale getirilmektedir.
- Mevcut protein sayısallaştırma yöntemlerinin çoğunda amino asitlerin bazıları aynı sayısal değerleri almaktadır. Örneğin protein diziliminin uzunluğuna bakılmaksızın, amino asitlerin sayısal değerleri aynı olabilmektedir. Hidrofobiklik yönteminde proteinler farklı bir türe ait olsa ya da farklı uzunluklara sahip olsa bile M amino asidi her zaman aynı değeri almaktadır. Ancak önerilen yöntemde böyle bir durum söz konusu değildir. Önerilen yöntemde amino asitlerin tekrarlanma sıklığı ve dizilimin uzunluğu kullanıldığı için dinamik bir yapı bulunmaktadır. Kısacası amino asitlerin sayısal ifadeleri sabit değil, değişkendir. Bu da yöntemin adaptif olmasına neden olmaktadır.

6.2. Viral Proteinler Arasındaki Etkileşimlere Uygulanması

Proteinler arasındaki etkileşimleri belirleyebilmek için birlikte evrim analizi sıklıkla kullanılmaktadır. Birlikte evrim, belirli etkileşimler sayesinde birbirine bağlı türler arasında gerçekleşen bir evrim olarak tanımlanabilir. Birlikte evrim teorisi, protein etkileşimi çalışmalarında hem proteinlerdeki temas bölgelerini göstermek hem de proteinler arasındaki etkileşimleri tanımlamak için sıklıkla kullanılmaktadır. KMA (Korelasyonlu Mutasyon Analizi) ve KEO (Korelasyonlu Evrimsel Oran) protein-protein etkileşimlerinde kullanılan en yaygın birlikte evrim analiz yöntemleridir [200]. KMA yönteminde protein etkileşimleri amino asit dizilimlerinin kalıntılarına göre gerçekleştirilmektedir. Bu yöneme göre benzer kalıntı gösteren proteinler, birbirleriyle daha güçlü bir bağ oluşturmaktadır [201]. KEO yönteminde ise protein etkileşimleri,

benzer filogenetik ağaçlardan toplanan birlikte evrim oranlarına dayalı olarak tespit edilmektedir. Bu yöntemde protein etkileşimleri filogenetik ağacın uzaklık matrisine göre belirlenmekte ve protein dizilimlerinin birlikte evrim oranları kullanılmaktadır. Şekil 6.1’de birlikte-evrim analizinin genel şeması verilmiştir.



Şekil 6.1. Birlikte evrim analizinin genel şeması

Birlikte evrim analizinde ilk olarak çeşitli türlere ait protein dizilimleri elde edilmektedir. Ardından ÇDH (Çoklu Dizilim Hizalama) uygulanmakta ve türlerin protein dizilimlerinin filogenetik ağaçları oluşturulmaktadır. Son adımda ise, filogenetik ağaç yapılarına göre uzaklık matrisleri oluşturulmakta ve birlikte evrim analizi gerçekleştirilmektedir. Ancak, ÇDH yüksek yanlış pozitif oranlar üreten bir yöntemdir [202, 203]. Bu nedenle araştırmacılar alternatif yaklaşımlar geliştirmeye başlamışlar ve hesaplama-tabanlı yöntemleri bu alana uygulamışlardır. Protein dizilimlerinin hesaplama-tabanlı yaklaşımlarla analiz edilmesi, protein dizilimleri arasındaki etkileşimleri arttırmakta ve yanlış pozitif oranlarını azaltmaktadır. Birlikte evrim analizinde seçilen

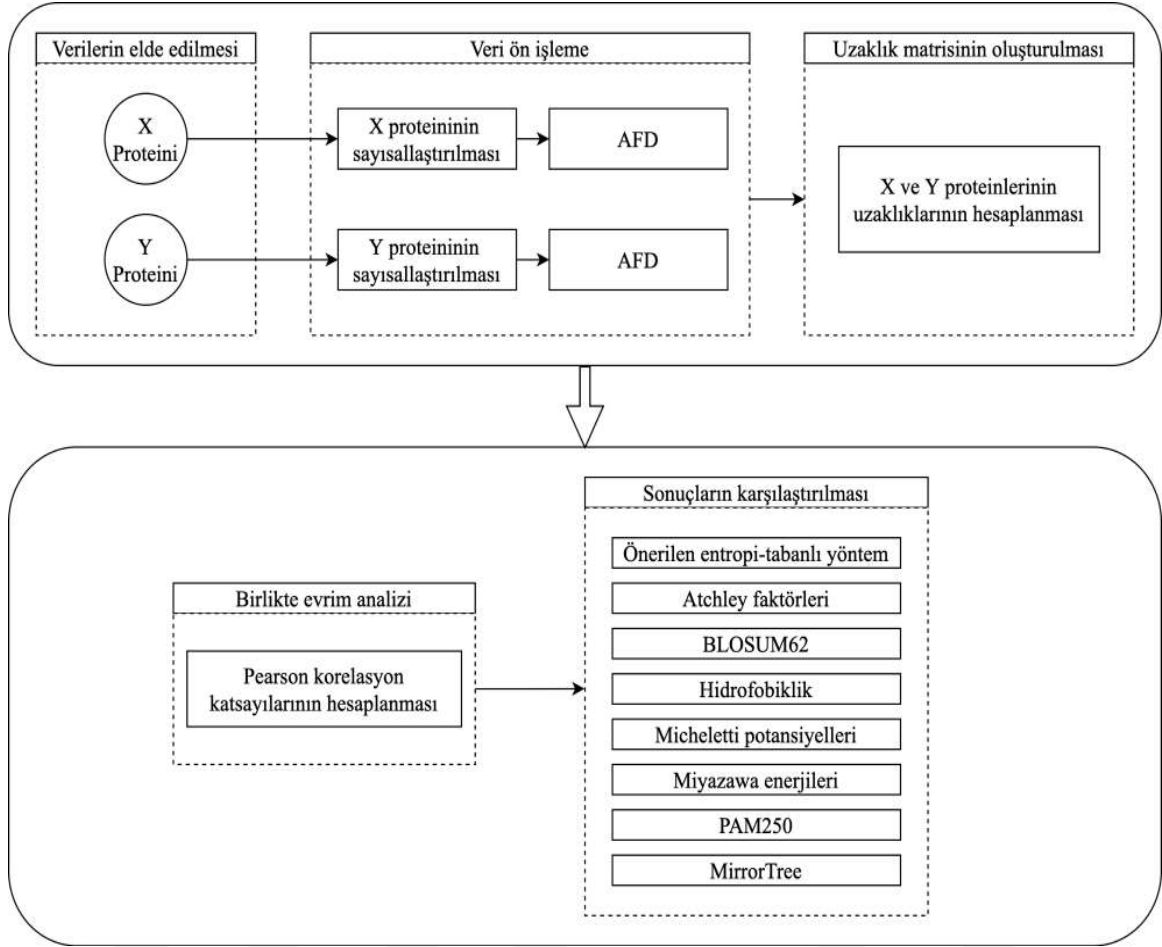
protein sayısallaştırma yöntemlerinin etkileşimleri belirlemede etkisi yüksek olmaktadır. Bundan dolayı seçilecek olan protein sayısallaştırma yönteminin önemi büyük olmaktadır.

Gerçekleştirilen çalışmada, yeni bir protein sayısallaştırma yöntemi önerilmiştir. Önerilen yöntemde ÇDH kullanılmamış ve birlikte evrim analizi gerçekleştirilmiştir. Çalışmanın ilk aşamasında protein dizilimleri hem önerilen entropi tabanlı hem de çeşitli protein sayısallaştırma yöntemleri ile sayısal hale getirilmiştir. Ardından AFD uygulanmış ve proteinler arasındaki benzerlik ve farklılıklar tanımlanmıştır. Sonraki aşamada, sayısal hale getirilmiş her protein diziliminin Öklid mesafesi ölçülerek uzaklık matrisi oluşturulmuş ve Pearson korelasyon katsayıları hesaplanmıştır. Son olarak ise katsayılar kullanılarak birlikte evrim analizi gerçekleştirilmiştir. Önerilen protein sayısallaştırma yönteminin başarımı iki farklı viral genom üzerinde test edilmiş ve performansı çeşitli protein sayısallaştırma yöntemleri ile kıyaslanmıştır. Ayrıca, önerilen yöntemin performansı, birlikte evrim analizinde sıklıkla kullanılan ÇDH-tabanlı MirrorTree yaklaşımı ile de karşılaştırılmıştır. Gerçekleştirilen çalışmanın katkıları şu şekilde özetlenebilir;

- Shannon entropisine dayalı yeni bir protein sayısallaştırma yöntemi önerilmiş ve literatüre kazandırılmıştır.
- Proteinler arasındaki etkileşimler, etkileşim açısından ziyade birlikte evrim analizi ile tahmin edilmiştir.
- Birlikte evrim analizinde kullanılan geleneksel bir yöntem olan MirrorTree ile hesaplanan yanlış pozitif oranlar gösterilmiştir.
- Hesaplama-tabanlı yaklaşımların daha etkili olduğu gözlemlenmiştir.

Gerçekleştirilen çalışma ile önerilen sayısallaştırma yönteminin protein-protein etkileşimleri üzerindeki etkisi gösterilmiştir. Uygulama sonuçları önerilen protein sayısallaştırma yönteminin doğru ve etkili bir sınıflandırma yaptığını belirtmiştir. İki protein arasındaki etkileşimi belirleyebilmek için korelasyonlardan yararlanılmaktadır. Eğer iki protein etkileşim halinde ise, aralarında yüksek korelasyon gözlemlenmektedir. Bunun yanı sıra, korelasyon değerinin düşük olması, iki protein arasındaki etkileşimin az olduğunu ya da etkileşimin hiç olmadığını göstermektedir. Bu bilgiler ışığında, gerçekleştirilen çalışmada, viral genomlara ait protein etkileşimleri birlikte evrim analizi ile tahmin edilmiştir. Birlikte evrim analizi Ebola virüsü ve Influenza A virüsü olmak üzere iki farklı viral genom üzerinde gerçekleştirilmiştir. Bu virüslerin seçilmesinin ana nedeni, her birinin genomlarında farklı proteinlerin kodlanmasıdır. Ebola virüsü, insanlarda ve primatlarda ciddi ateşe neden olan hastalık formu oluşturmaktadır. Yedi farklı proteini kodlamaktadır. Bunlar; VP40, GP (Glikoprotein), NP (Nükleoprotein), VP24, VP30, VP35 ve L (RNA Polimeraz) proteinleridir. VP24, VP30, VP35 ve L proteinleri RNA transkripsiyonundan sorumluyken, geri kalan proteinler virüsün genel yapısının oluşmasında rol almaktadırlar. Grip olarak da bilinen Influenza A virüsü ise boğaz, burun ve akciğerleri enfekte

eden bir virüstür. Bazı durumlarda ölüme bile sebebiyet vermektedir. 11 farklı proteini kodlayan 8 segmentten meydana gelmektedir. Bu proteinler şu şekilde sıralanabilir; PB2 (Polimeraz Bazik 2), PB1 (Polimeraz Bazik 1), PA (Polimeraz Asidik protein), PA-X, NP (Nükleoprotein), NEP (Nükleer İhraç Protein), NS1 (Yapısal Olmayan protein), M1 (Matris protein 1), M2 (Matris protein 2), HA (Hemaglutinin) ve NA (Nöraminidaz) [204, 205]. NA ve HA proteinleri virüsün enfeksiyonundan sorumlu olan proteinlerdir [206]. M1 ve M2 proteinleri Influenza A virüsünde partikül oluşumunda önemli rol oynamaktadır [207]. NP proteini ise RNA replikasyonu ve transkripsiyonundan sorumludur [208]. Influenza A virüsündeki NEP protein ise viral ribonükleoproteinlerin ihracatında etkili bir görev üstlenmektedir [208]. NS1 proteini organizmadaki birçok hücresel bileşenle etkileşime girmektedir. PB1, PB2, PA ve PA-X ise polimeraz komplekslerini temsil etmektedir. Literatürde makul düzeyde protein sayısallaştırma yöntemi bulunmaktadır. Çalışmadaki ana amaçlardan bir tanesi literatüre yeni bir protein sayısallaştırma yöntemi kazandırmaktır. Bu nedenle, önerilen protein sayısallaştırma yönteminin performansı, literatürde sıklıkla kullanılan protein sayısallaştırma yöntemleri ile kıyaslanmıştır. Bu protein sayısallaştırma yöntemleri PAM250, BLOSUM62, Miyazawa enerjileri, Micheletti potansiyelleri, Atchley faktörleri ve hidrofobiklik yöntemleridir. İki farklı virüs genomu üzerinde gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir. Bunlara ek olarak, önerilen yöntemin performansı hem çeşitli protein sayısallaştırma yöntemleri hem de MirrorTree yöntemi ile kıyaslanmıştır. MirrorTree yaklaşımı birlikte-evrim analizinde sıklıkla kullanılan bir yöntemdir. Bu yöntemde, protein dizilimleri arasındaki uzaklıklar hesaplanır ve filogenetik ağaçlar oluşturularak karşılaştırma yapılmaktadır. MirrorTree yöntemi ÇDH kullanmaktadır. ÇDH benzer uzunluktaki üç ya da daha fazla biyolojik dizilimin hizalanması olarak ifade edilmektedir. Bu yöntem sayesinde dizilimler arasındaki evrimsel ilişkiler incelenebilmektedir. Özetlemek gerekirse, biyolojik (DNA, protein) dizilimler arasındaki yapısal, fonksiyonel ve evrimsel ilişkileri saptayabilecek benzer bölgeleri inceleyebilmek için kullanılmaktadır. Bu çalışmada olduğu gibi MirrorTree yönteminde de Pearson katsayıları kullanılmaktadır. MirrorTree yaklaşımına göre, iki protein arasındaki korelasyon değeri 0.80'in üzerinde ise, bu etkileşim iyi bir etkileşimdir. Şekil 6.2'de çalışmanın akış şeması verilmiştir.



Şekil 6.2. Önerilen entropi yönteminin birlikte evrim analiziyle değerlendirme süreci

6.2.1. Ebola Virüsüne Ait Protein Etkileşimlerinin Birlikte Evrim Analizi

Çalışma kapsamında Ebola virüsüne ait 7 adet protein değerlendirilmiş ve birlikte evrim analizi gerçekleştirilmiştir. Tablo 6.1’de önerilen protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.1. Önerilen protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,91	0,90	0,92	0,85	0,88	0,88
NP		1	0,99	0,93	0,75	0,99	0,86
VP24			1	0,94	0,78	0,98	0,87
VP30				1	0,86	0,96	0,91
VP35					1	0,79	0,88
VP40						1	0,89
L							1

Tablo 6.1’deki sonuçlar değerlendirildiğinde, önerilen yöntem ile GP-NP, GP-VP24, GP-VP30, GP-VP35, GP-VP40, GP-L, NP-VP24, NP-VP30, NP-VP40, NP-L, VP24-VP30, VP24-VP40, VP24-L, VP30-VP35, VP30-VP40, VP30-L, VP35-L ve VP40-L protein çiftleri arasındaki korelasyon değeri 0,8 ve üzeri olduğu için, mükemmel derecede bir korelasyon gözlemlenmiştir. Bunlara ek olarak, NP-VP35, VP24-VP35 ve VP35-VP40 protein çiftleri arasında ise yüksek dereceli korelasyon görülmüştür. Tablo 6.2’de PAM250 protein sayısallaştırma yöntemi ile yapılan birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.2. PAM250 protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,78	0,81	0,81	0,71	0,79	0,75
NP		1	0,99	0,95	0,69	0,99	0,85
VP24			1	0,94	0,69	0,98	0,87
VP30				1	0,67	0,94	0,82
VP35					1	0,72	0,80
VP40						1	0,85
L							1

Tablo 6.2’de verilen uygulama sonuçlarına göre, PAM250 protein sayısallaştırma yöntemi ile çoğu protein çifti arasında başarılı etkileşimler elde edilmiştir. GP-VP24, GP-VP30, NP-VP24, NP-VP30, NP-VP40, NP-L, VP24-VP30, VP24-VP40, VP24-L, VP30-VP40, VP30-L, VP35-L ve VP40-L protein çiftleri arasında mükemmel derecede korelasyon gözlemlenmiştir. Geri kalan bütün protein çiftleri arasında ise yüksek derecede korelasyon hesaplanmıştır. Önerilen yöntem dışında, bu yöntemde de hem yüksek hem de mükemmel derecede korelasyon elde edilmiştir. Tablo 6.3’te BLOSUM62 protein sayısallaştırma yöntemi ile yapılan birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.3. BLOSUM62 protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,74	0,74	0,83	0,71	0,81	0,78
NP		1	0,99	0,94	0,63	0,98	0,85
VP24			1	0,92	0,67	0,97	0,87
VP30				1	0,68	0,94	0,84
VP35					1	0,60	0,77
VP40						1	0,82
L							1

Başka bir evrim-tabanlı yöntem olan BLOSUM62 matrisi ile hem yüksek hem de mükemmel derecede korelasyon değerleri elde edilmiştir. GP-VP30, GP-VP40, NP-VP24, NP-VP30, NP-VP40, NP-L, VP24-VP30, VP24-VP40, VP24-L, VP30-VP40, VP30-L ve VP40-L protein çiftleri arasında mükemmel derecede korelasyon gözlemlenmiştir. GP-NP, GP-VP24, GP-VP35, GP-L, NP-VP35, VP24-VP35, VP30-VP35, VP35-VP40 ve VP35-L protein çiftleri arasında ise yüksek derecede korelasyon hesaplanmıştır. Hem önerilen yöntem hem de evrimsel yöntemlerle elde edilen başarılı performans yapı-tabanı protein sayısallaştırma yöntemlerinde gözlemlenmemiştir. Tablo 6.4'te Miyazawa enerjileri protein sayısallaştırma yöntemi ile yapılan birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.4. Miyazawa enerjileri protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,75	0,66	0,74	0,64	0,71	0,70
NP		1	0,96	0,93	0,64	0,96	0,79
VP24			1	0,90	0,54	0,97	0,75
VP30				1	0,61	0,92	0,78
VP35					1	0,61	0,75
VP40						1	0,87
L							1

Tablo 6.4'te elde edilen uygulama sonuçlarına göre, sadece NP-VP24, NP-VP30, NP-VP40, VP24-VP30, VP24-VP40, VP30-VP40 ve VP40-L protein çiftleri arasında mükemmel derecede korelasyon görülmüştür. Bunlara ek olarak, GP proteinini, bütün protein çiftleri ile yüksek derecede korelasyon göstermiştir. NP proteininde ise sadece NP-VP35 protein çifti arasında yüksek derecede korelasyon gözlemlenmiştir. VP24-L protein çifti arasında yüksek korelasyon hesaplanmışken, VP24-VP35 protein çifti arasında ise orta derecede korelasyon gözlemlenmiştir. VP30-VP35 ve

VP30-L protein çiftleri arasında ise orta derecede etkileşim saptanmıştır. Benzer şekilde bir etkileşim VP35-VP40 ve VP35-L protein çiftleri içinde elde edilmiştir. Tablo 6.5'te başka bir yapı-tabanlı protein sayısallaştırma yöntemi olan Micheletti potansiyellerinin etkileşim skorları verilmiştir.

Tablo 6.5. Micheletti potansiyelleri protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,74	0,71	0,63	0,62	0,75	0,69
NP		1	0,97	0,89	0,58	0,98	0,72
VP24			1	0,90	0,66	0,97	0,80
VP30				1	0,53	0,90	0,70
VP35					1	0,62	0,79
VP40						1	0,71
L							1

Tablo 6.5'te verilen sonuçlar doğrultusunda, GP proteini hiçbir protein ile mükemmel derecede etkileşim kuramamıştır. GP proteini ve diğer proteinler arasında yalnızca yüksek derecede korelasyon görülmüş ve en yüksek etkileşim GP-VP40 protein çifti arasında gözlemlenmiştir. NP proteini ise VP24, VP30 ve VP40 protein çiftleri ile mükemmel derecede etkileşim kurmuştur. En yüksek korelasyon değeri ise NP-VP40 protein çifti arasında gerçekleşmiş ve korelasyon değeri 0,98 bulunmuştur. Ancak, NP-VP35 protein çifti arasında ise orta derecede bir korelasyon gözlemlenmiştir. NP-L protein çifti arasında ise yüksek derecede korelasyon hesaplanmıştır. VP24 proteini ise, VP35 proteini hariç diğer bütün proteinler ile mükemmel derecede etkileşim göstermiştir. VP30 proteini ise en düşük etkileşimi VP35 proteini ile kurmuş ve orta derecede korelasyon gözlemlenmiştir. VP30-VP40 protein çifti arasındaki korelasyon 0,90 olurken, VP30-L protein çifti arasındaki korelasyon ise 0,70 olarak hesaplanmıştır. VP35-VP40 ve VP35-L proteinleri arasında ise yüksek derecede etkileşim gözlemlenmiştir. Son olarak VP40-L protein çifti arasında da yüksek derecede etkileşim görülmüş ve korelasyon skoru 0,71 hesaplanmıştır. Tablo 6.6'da Atchley faktörleri ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.6. Atchley faktörleri protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,80	0,72	0,79	0,74	0,78	0,80
NP		1	0,98	0,94	0,70	0,99	0,83
VP24			1	0,93	0,58	0,98	0,77
VP30				1	0,60	0,95	0,79
VP35					1	0,63	0,77
VP40						1	0,79
L							1

Tablo 6.6’da verilen bulgular doğrultusunda GP proteini sadece NP ve L protein çiftleri ile mükemmel derecede etkileşim göstermiştir. Geri kalan bütün protein çiftleri ile yüksek derece korelasyon sergilemiştir. NP-VP24, NP-VP30, NPVP40 ve NP-L protein çiftleri arasında ise mükemmel derecede korelasyon görülmüştür. NP proteini sadece VP35 proteini ile yüksek derecede korelasyon sergilemiştir. VP24 proteini incelendiğinde ise, VP24-VP35 arasında orta derecede etkileşim gözlemlenmiştir. VP24 proteini L proteini ile yüksek etkileşim kurmuşken, VP30 ve VP40 proteinleri ile mükemmel derecede etkileşim sergilemiştir. VP30 proteini ise sadece VP40 proteini ile mükemmel derecede etkileşim göstermişken, VP35 ve L proteinleri ile yüksek derecede korelasyon sergilemiştir. VP35-VP40 ve VP35-L protein çiftleri arasında ise yüksek derecede korelasyon gözlemlenmiştir. Benzer şekilde VP40-L protein çiftleri arasında da yüksek korelasyon değeri hesaplanmıştır. Tablo 6.7’de hidrofobiklik protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.7. Hidrofobiklik protein sayısallaştırma yöntemi ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,78	0,72	0,78	0,72	0,80	0,80
NP		1	0,99	0,94	0,67	0,99	0,81
VP24			1	0,93	0,58	0,98	0,78
VP30				1	0,59	0,95	0,77
VP35					1	0,64	0,76
VP40						1	0,79
L							1

Tablo 6.7’de verilen sonuçlar doğrultusunda, hidrofobiklik protein sayısallaştırma yöntemiyle, GP-VP40, GP-L, NP-VP24, NP-VP30, NP-VP40, NP-L, VP24-VP30, VP24-VP40 ve VP30-VP40 protein çiftleri arasında mükemmel derecede korelasyon gözlemlenmiştir. Buna ek olarak GP

proteini NP, VP24, VP30 ve VP35 proteinleri ile yüksek derecede etkileşim kurmuştur. Bunlara ek olarak, NP-VP35, VP24-L, VP30-L, VP35-VP40, VP35-L ve VP40-L protein çiftleri arasındaki etkileşimlerde yüksek bulunmuştur. Hidrofobiklik protein sayısallaştırma yönteminde sadece VP24-VP35 ve VP30-VP35 protein çiftleri arasında orta derecede etkileşim gözlemlenmiştir.

6.2.2. Influenza A Virüsüne Ait Protein Etkileşimlerinin Birlikte Evrim Analizi

Influenza A virüsüne ait 11 adet protein kullanılarak, önerilen protein sayısallaştırma yöntemi, diğer protein sayısallaştırma yöntemleri ve MirrorTree yöntemi ile birlikte evrim analizi gerçekleştirilmiştir. Tablo 6.8’de önerilen protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.8. Önerilen Shannon entropisine dayalı yöntem ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,21	0,23	0,85	0,63	0,77	0,94	0,70	0,83	0,75	0,89
M1		1	0,98	0,31	0,23	0,18	0,33	0,13	0,09	0,18	0,15
M2			1	0,31	0,08	0,17	0,24	0,19	0,12	0,35	0,82
NA				1	0,35	0,66	0,81	0,63	0,78	0,54	0,77
NP					1	0,33	0,35	0,27	0,65	0,32	0,41
NEP						1	0,93	0,76	0,84	0,88	0,76
NS1							1	0,92	0,60	0,78	0,89
PA								1	0,55	0,68	0,78
PAX									1	0,60	0,79
PB1										1	0,89
PB2											1

Tablo 6.8’deki sonuçlara göre HA proteinin NA, NS1, PAX ve PB2 proteinleri ile mükemmel etkileşim gösterdiği görülmüştür. Bunun yanı sıra NP, NEP, PA ve PB1 proteinleri ile yüksek korelasyon sergilemiştir. Ancak, M1 ve M2 proteinleri ile düşük etkileşim göstermiştir. M1 protein ise sadece M2 proteini ile mükemmel etkileşim göstermiş ve korelasyon sonucu bu iki protein çifti için 0,98 olmuştur. Sonuçlar genel olarak incelendiğinde M1 proteinine ait başarımın yeterli olmadığı gözlemlenmiştir. M1-NA, M1-NP ve M1-NS1 arasında düşük derecede korelasyon görülmüşken, geri kalan bütün protein çiftleri için çok düşük etkileşim elde edilmiştir. M2 proteini ise sadece PB2 proteini ile mükemmel derecede etkileşim kurmuş ve bu iki proteinin korelasyon skoru 0,82 olmuştur. Geri kalan proteinler arasında düşük ve çok düşük olmak üzere iki farklı derecede etkileşim gözlemlenmiştir. NA-NS1 protein çifti arasında mükemmel derecede korelasyon gözlemlenmişken, NA-NEP, NA-PA, NA-PAX ve NA-PB2 protein çiftleri arasında ise

yüksek derecede etkileşim elde edilmiştir. Bunlara ek olarak, NA proteini NP proteini ile düşük derecede etkileşim kurmuşken, PB1 proteini ile orta derecede etkileşim sergilemiştir. Bunun yanı sıra, NP proteini incelendiğinde, bu proteinin sadece PAX proteiniyle yüksek etkileşim gösterdiği gözlemlenmiştir. Geri kalan protein çiftleri ile orta ve düşük olmak üzere iki farklı etkileşim göstermiştir. NEP-NS1, NEP-PAX ve NEP-PB1 protein çiftleri arasında mükemmel derecede korelasyon elde edilmiştir. NEP-PA ve NEP-PB2 protein çiftleri için ise yüksek derecede korelasyon hesaplanmıştır. NS1-PA, NS1-PB2 ve PB1-PB2 protein çiftleri mükemmel korelasyon sergilemişlerdir. PA-PAX protein çifti hariç, NS1-PAX, NS1-PB2, PA-PB1, PA-PB2, PAX-PB1 ve PAX-PB2 protein çiftleri arasında yüksek korelasyon gözlemlenmiştir. Tablo 6.9’da PAM250 protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.9. PAM250 protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,14	0,28	0,86	0,38	0,82	0,73	0,51	0,28	0,62	0,60
M1		1	0,96	0,12	0,23	0,43	0,48	0,20	0,07	0,07	0,17
M2			1	0,23	0,18	0,51	0,57	0,26	0,04	0,18	0,25
NA				1	0,33	0,78	0,72	0,40	0,22	0,52	0,49
NP					1	0,39	0,38	0,23	0,03	0,24	0,26
NEP						1	0,92	0,40	0,20	0,51	0,42
NS1							1	0,45	0,15	0,47	0,49
PA								1	0,19	0,75	0,75
PAX									1	0,50	0,22
PB1										1	0,70
PB2											1

Tablo 6.9’da verilen bulgular doğrultusunda, HA proteini sadece NA ve NEP proteinleri ile mükemmel derecede etkileşim göstermiştir. Bunun yanı sıra, HA-NS1, HA-PB1 ve HA-PB2 protein çiftleri arasında ise yüksek korelasyon gözlemlenmiştir. HA proteini sadece PA proteini ile orta derecede etkileşim sergilemiş ve korelasyon değeri bu iki protein için 0,51 olmuştur. HA-M2, HA-NP ve HA-PAX protein çiftleri arasında düşük etkileşim belirlenmişken, sadece HA-M1 protein çifti arasında çok düşük korelasyon hesaplanmıştır. M1 proteini ise sadece M2 proteini ile mükemmel derecede etkileşim oluşturmuştur. Yalnızca, NEP ve NS1 protein çiftleri ile orta derecede korelasyon gösterirken, geri kalan proteinler için hem çok düşük hem de düşük olmak üzere iki farklı korelasyon sergilemiştir. M2 proteini ise, sadece NEP ve NS1 proteinleri ile orta derecede etkileşim göstermiştir. NA, PA ve PB2 proteinleri ile düşük korelasyon elde etmiştir. Bunlara ek olarak, NP, PAX ve PB1 proteinleri ile çok düşük etkileşim kurmuştur. NP proteini ise

PAX proteini hariç, NEP, NS1, PA, PB1 ve PB2 proteinleri ile düşük etkileşim kurmuştur. Bunun yanı sıra, NEP-NS1 proteini arasındaki mükemmel derecede etkileşim görülmüştür. NEP-PA, NEP-PB1 ve NEP-PB2 protein çiftleri arasında orta derecede korelasyon hesaplanmışken, NEP-PAX arasında düşük derecede korelasyon gözlemlenmiştir. NS1 proteini ise yalnızca PA, PB1 ve PB2 proteinleri ile orta düzeyde etkileşim oluşturmuştur. Diğer proteinlerde de olduğu gibi, PA proteini PAX proteini ile çok düşük etkileşim kurmuşken hem PB1 hem de PB2 proteinleri ile yüksek etkileşim ağı oluşturmuştur. PAX-PB1 protein çifti arasındaki etkileşim orta düzeydeyken, PAX-PB2 protein çifti arasındaki etkileşim düşük olmuştur. Son olarak PB1 proteini PB2 proteini ile yüksek oranda etkileşim oluşturmuştur. Tablo 6.10'da BLOSUM62 protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.10. BLOSUM62 protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,23	0,23	0,85	0,26	0,68	0,72	0,59	0,26	0,47	0,63
M1		1	0,96	0,19	0,15	0,40	0,51	0,09	0,02	0,07	0,05
M2			1	0,18	0,08	0,34	0,48	0,08	0,15	0,11	0,03
NA				1	0,21	0,67	0,69	0,41	0,22	0,36	0,44
NP					1	0,19	0,22	0,14	0,05	0,02	0,15
NEP						1	0,75	0,32	0,01	0,21	0,31
NS1							1	0,32	0,16	0,32	0,39
PA								1	0,21	0,53	0,69
PAX									1	0,64	0,33
PB1										1	0,61
PB2											1

Tablo 6.10'da verilen BLOSUM62 protein sayısallaştırma sonuçları doğrultusunda, HA yalnızca NA proteini ile mükemmel derecede etkileşim kurmuştur. HA-NEP, HA-NS1 ve HA-PB2 protein çiftleri arasında ise yüksek derecede korelasyon elde edilmiştir. HA-PA ve HA-PB1 protein çiftleri arasında ise orta düzeyde etkileşim gözlemlenmiştir. HA proteini PA ve PB1 ile orta derecede etkileşim oluşturmuşken, M1, M2, NP ve PAX proteinleri ile düşük seviyede etkileşim kurmuştur. M1 proteini ise sadece M2 ile mükemmel etkileşim oluşturmuştur. M1-NEP ve M1-NS1 arasında ise orta düzeyde etkileşim gözlemlenmişken, M1 ve diğer bütün proteinler ile çok düşük etkileşim ağı oluşmuştur. M2-NS1 protein çiftleri arasında ise orta düzeyde korelasyon gözlemlenmiştir. M2-NEP protein çiftleri arasındaki korelasyon derecesi düşükken, M2-NA, M2-NP, M2-PA, M2-PAX, M2-PB1 ve M2-PB2 protein çiftleri arasındaki korelasyon çok düşük olmuştur. NA proteini ise, NEP ve NS1 proteinleri ile yüksek oranda etkileşim sergilemiştir. Bunun yanı sıra, PA ve PB2

proteinleri ile de orta seviyede etkileşim ağı meydana getirmiştir. Geri kalan proteinler (NP, PAX ve PB1) ile düşük korelasyon elde etmiştir. NP proteini yalnızca NS1 proteini ile düşük oranda etkileşim kurarken, NEP, PA, PAX, PB1 ve PB2 proteinleri ile çok düşük etkileşim oluşturmuştur. Bunlara ek olarak, NEP-NS1 protein çiftleri arasında yüksek korelasyon elde edilmişken, NEP-PA, NEP-PB1 ve NEP-PB2 protein çiftleri arasında düşük korelasyon gözlemlenmiştir. NEP-PAX protein çifti arasında ise çok düşük etkileşim meydana gelmiştir. NS1 proteini ise, PA, PB1 ve PB2 proteinleri ile düşük oranda etkileşim sergilemişken, PAX proteini ile çok düşük etkileşim kurmuştur. PA proteini incelendiğinde ise, PA-PAX protein çiftleri arasındaki korelasyon düşük, PA-PB1 protein çiftleri arasındaki korelasyon orta ve PA-PB2 protein çiftleri arasındaki korelasyon yüksek olmuştur. PAX-PB1 protein çiftleri yüksek etkileşim sergilemişken, PAX-PB2 arasındaki etkileşim düşük olmuştur. PB1 ve PB2 arasındaki etkileşim ise yüksek olmuştur. Tablo 6.11’de Miyazawa enerjileri protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.11. Miyazawa enerjileri protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,17	0,19	0,84	0,34	0,74	0,78	0,37	0,33	0,58	0,48
M1		1	0,93	0,14	0,24	0,46	0,51	0,05	0,07	0,10	0,09
M2			1	0,15	0,11	0,43	0,48	0,07	0,01	0,13	0,10
NA				1	0,28	0,73	0,75	0,21	0,29	0,49	0,34
NP					1	0,28	0,31	0,12	0,05	0,20	0,17
NEP						1	0,89	0,20	0,17	0,45	0,25
NS1							1	0,16	0,28	0,47	0,28
PA								1	0,01	0,54	0,52
PAX									1	0,42	0,16
PB1										1	0,57
PB2											1

Tablo 6.11’de verilen sonuçlara göre, Miyazawa enerjileri protein sayısallaştırma yöntemi ile, HA proteini yalnızca NA proteini ile mükemmel düzeyde etkileşim göstermiştir. Buna ek olarak, NEP ve NS1 proteinleri ile yüksek derecede etkileşim ağı oluşturmuştur. HA-PB1 ve HA-PB2 protein çiftleri arasında orta derecede korelasyon gözlemlenmişken, HA-NP, HA-PA ve HA-PAX protein çiftleri arasında düşük derecede korelasyon hesaplanmıştır. Son olarak, HA proteini, M1 ve M2 proteinleri ile çok düşük derecede etkileşime girmiştir. M1 proteini ise sadece M2 proteini ile mükemmel etkileşim göstermiştir. M1-NEP ve M1-NS1 arasında elde edilen korelasyon değeri ise orta derecede olmuştur. M1 proteininin NP proteini ile etkileşimi düşükken, NA, PA, PAX, PB1

ve PB2 proteinleri ile etkileşimi ise çok düşük olmuştur. M2 proteinini ise yalnızca NEP ve NS1 proteinleri ile orta düzeyde etkileşim ağı geliştirmiştir. Geri kalan proteinler ile (NA, NP, PA, PAX, PB1 ve PB2) çok düşük korelasyon gözlemlenmiştir. Bunun yanı sıra, NA-NEP ve NA-NS1 protein çiftleri arasında yüksek etkileşim görülmüştür. Orta derece korelasyon NA-PB1 protein çiftleri arasında elde edilmişken, NA-NP, NA-PA, NA-PAX ve NA-PB2 arasında düşük korelasyon değerleri hesaplanmıştır. NP-NEP, NP-NS1 ve NP-PB1 protein çiftleri arasında düşük etkileşim ağı oluşturulmuşken, NP-PA, NP-PAX ve NP-PB2 protein çiftleri arasında ise çok düşük etkileşim ağı gözlemlenmiştir. Bunlara ek olarak, NEP-NS1 arasında mükemmel düzeyde etkileşim elde edilmiştir. NEP-PB1 arasında orta derecede korelasyon hesaplanmışken, NEP-PA ve NEP-PB2 arasında düşük korelasyonlar gözlemlenmiştir. NEP-PAX protein çiftleri arasındaki etkileşim ise çok düşük olarak gözlemlenmiştir. NS1 ve PB1 proteinleri arasında ise orta derecede etkileşim elde edilmiş ve korelasyon sonucu 0,47 olmuştur. NS1-PAX ve NS1-PB2 arasında ise düşük oranda etkileşimler görülmüştür. En düşük etkileşim ise NS1-PA proteinleri arasında elde edilmiştir. PA-PB1 ve PA-PB2 arasında orta derecede etkileşim saptanmışken, PA-PAX arasında çok düşük etkileşim belirlenmiştir. PAX ve PB1 proteinleri arasındaki etkileşim orta derecedeyken, PAX ve PB2 arasındaki etkileşim çok düşük olmuştur. PB1 ve PB2 protein çiftleri arasındaki etkileşim ise orta düzeyde elde edilmiştir. Tablo 6.12’de Micheletti potansiyelleri protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.12. Micheletti potansiyelleri protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,08	0,22	0,85	0,37	0,70	0,69	0,51	0,35	0,52	0,49
M1		1	0,93	0,06	0,22	0,27	0,50	0,13	0,13	0,03	0,09
M2			1	0,18	0,25	0,31	0,58	0,15	0,12	0,31	0,22
NA				1	0,32	0,65	0,66	0,40	0,25	0,44	0,41
NP					1	0,22	0,31	0,18	0,05	0,14	0,19
NEP						1	0,64	0,73	0,06	0,31	0,35
NS1							1	0,19	0,16	0,30	0,26
PA								1	0,03	0,37	0,40
PAX									1	0,46	0,27
PB1										1	0,68
PB2											1

Tablo 6.12’de verilen sonuçlar incelendiğinde, HA-NA protein çifti arasındaki etkileşimin mükemmel olduğu görülmüştür. Bunun yanı sıra, HA proteinini NEP ve NS1 proteinleri ile yüksek derecede etkileşim sergilemiştir. HA-PA, HA-PB1 ve HA-PB2 protein çiftleri arasında orta

düzeyde etkileşim saptanmışken, HA-M2, HA-NP ve HA-PAX proteinleri arasında düşük seviyeli etkileşimler meydana gelmiştir. Son olarak, HA-M1 protein çifti arasında ise çok düşük korelasyon hesaplanmıştır. M1 proteini M2 proteini ile mükemmel düzeyde etkileşim kurmuştur. M1-NS1 proteini ile orta düzeyde etkileşim oluşmuşken, geri kalan proteinler ile (NA, PA, PAX, PB1 ve PB2) çok düşük düzeyde etkileşim kurulmuştur. M2-NS1 protein çifti arasında orta seviyede etkileşim elde edilmiştir. Bunun yanı sıra, M2 proteini NP, NEP, PB1 ve PB2 proteinleri ile düşük derecede etkileşim sergilemiştir. Geri kalan proteinler ile (NA, PA ve PAX) çok düşük derecede etkileşim ağı oluşturulmuştur. NA proteini ise yalnızca NEP ve NS1 proteinleri ile yüksek oranda etkileşim göstermiştir. Bunlara ek olarak, NA proteini ile PA, PB1 ve PB2 arasında orta düzeyde etkileşim kurulmuştur. Son olarak NA-NP ve NA-PAX protein çiftleri arasında çok düşük korelasyon skorları hesaplanmıştır. NP proteini ise yalnızca NEP ve NS1 proteinleri ile düşük derecede korelasyon sergilemiştir. Geri kalan (PA, PAX, PB1 ve PB2) proteinler ile çok düşük etkileşim kurulmuştur. NEP-PA ve NEP-NS1 protein çiftleri arasında yüksek korelasyon hesaplanmışken, NEP-PB1 ve NEP-PB2 proteinleri arasında düşük korelasyon gözlemlenmiştir. Son olarak NEP proteini PAX ile çok düşük derecede etkileşim göstermiş ve korelasyon değeri 0,06 olarak hesaplanmıştır. NS1 proteini PB1 ve PB2 proteinleri ile düşük derece etkileşim kurmuşken, PA ve PAX proteinleri ile çok düşük seviyede etkileşim oluşturmuştur. Diğer yandan ise, PA proteini sadece PB2 proteini ile orta düzeyde etkileşim oluşturmuştur. PB1 ile düşük derecede korelasyon gösteren PB2 proteini PAX proteini ile çok düşük derecede etkileşim sergilemiştir. PAX ve PB1 proteinleri arasındaki korelasyon değeri orta derecede hesaplanırken, PAX ve PB2 proteinleri arasındaki bu değer çok düşük dereceye gerilemiştir. Son olarak PB1 ile PB2 protein çifti arasındaki etkileşim yüksek olmuştur. Tablo 6.13'te Atchley faktörleri protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.13. Atchley faktörleri protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,17	0,34	0,86	0,36	0,62	0,76	0,46	0,32	0,54	0,55
M1		1	0,95	0,14	0,17	0,41	0,48	0,02	0,04	0,09	0,06
M2			1	0,27	0,16	0,50	0,57	0,12	0,02	0,23	0,17
NA				1	0,31	0,62	0,75	0,34	0,30	0,47	0,38
NP					1	0,28	0,35	0,10	0,06	0,20	0,21
NEP						1	0,81	0,29	0,10	0,42	0,21
NS1							1	0,33	0,25	0,46	0,31
PA								1	0,01	0,63	0,63
PAX									1	0,29	0,07
PB1										1	0,62
PB2											1

Tablo 6.13'te verilen birlikte evrim analizi sonuçları doğrultusunda, HA proteini sadece NA proteini ile mükemmel etkileşim sergilemiştir. HA-NEP ve HA-NS1 protein çiftleri arasında yüksek korelasyon gözlemlenmişken, HA-PA, HA-PB1 ve HA-PB2 protein çiftleri arasında ise orta derecede korelasyon elde edilmiştir. Bunların yanı sıra, M2, NP ve PAX proteinleri ile düşük derecede etkileşim oluşturmuştur. HA proteini yalnızca M1 proteini ile çok düşük etkileşim kurmuştur. M1 proteini M2 proteini ile mükemmel derecede etkileşim ağı oluşturmuşken, NEP ve NS1 proteinleri ile orta düzeyde etkileşim oluşturmuştur. Son olarak M1 proteini geri kalan proteinler (NA, NP, PA, PAX, PB1 ve PB2) ile çok düşük derecede korelasyon göstermiştir. M2 proteini ise NEP ve NS1 proteinleri ile orta seviyede etkileşim kurmuşken, NA ve PB1 proteinleri ile düşük seviyede etkileşim meydana getirmiştir. NP, PA, PAX ve PB2 proteinleri ile çok düşük etkileşime girmiştir. NA proteini ise NEP ve NS1 ile yüksek etkileşim göstermiştir. Sadece PB1 proteini ile orta derecede etkileşime giren NA proteini NP, PA, PAX ve PB2 proteinleri ile düşük düzeyde etkileşim sergilemiştir. NP-NEP, NP-NS1 ve NP-PB1 protein çiftleri arasında ise düşük derecede korelasyon gözlemlenmiştir. PA, PAX ve PB2 proteinleri ile ise çok düşük derecede korelasyon hesaplanmıştır. NEP-NS1 protein çifti ise mükemmel etkileşim kurarak etkileşim ağını oluşturmuştur. NEP-PB1 protein çifti arasında orta düzeyde etkileşim görülmüşken, NEP-PA ve NEP-PB2 protein çiftleri arasında düşük düzeyde etkileşim gözlemlenmiştir. NEP ve PAX proteinleri arasında ise çok düşük etkileşim belirlenmiştir. NS1-PB1 protein çifti arasında orta düzeyde korelasyon hesaplanmışken, NS1-PA, NS1-PAX ve NS1-PB2 protein çiftleri arasındaki korelasyon ise düşük olmuştur. PA proteini PB1 ve PB2 proteinleri ile yüksek etkileşim meydana getirmişken, PAX proteini ile çok düşük etkileşime girmiştir. PAX-PB1 protein çiftleri arasındaki korelasyon ise düşük olarak hesaplanmıştır. PAX proteinin PB2 proteini ile etkileşimi ise çok düşük

olarak belirlenmiştir. Son olarak PB1 ise PB2 proteini ile yüksek etkileşim meydana getirmiştir. Tablo 6.14'te hidrofobiklik protein sayısallaştırma yöntemi ile gerçekleştirilen birlikte evrim analizinin sonuçları verilmiştir.

Tablo 6.14. Hidrofobiklik protein sayısallaştırma yöntemi ile gerçekleştirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,15	0,27	0,84	0,43	0,75	0,95	0,77	0,81	0,62	0,87
M1		1	0,95	0,11	0,10	0,10	0,11	0,06	0,04	0,15	0,19
M2			1	0,18	0,08	0,17	0,25	0,18	0,12	0,28	0,77
NA				1	0,38	0,63	0,79	0,61	0,79	0,54	0,77
NP					1	0,32	0,39	0,27	0,64	0,31	0,40
NEP						1	0,76	0,73	0,67	0,83	0,73
NS1							1	0,81	0,63	0,68	0,84
PA								1	0,55	0,67	0,71
PAX									1	0,40	0,56
PB1										1	0,69
PB2											1

Tablo 6.14'te verilen sonuçlar doğrultusunda, HA proteini NA, NS1, PAX ve PB2 proteinleri ile mükemmel etkileşim kurmuştur. Bunun yanı sıra, HA-NEP, HA-PA ve HA-PB1 protein çiftleri arasında ise yüksek etkileşim gözlemlenmiştir. HA-NP protein çifti arasında orta derece korelasyon hesaplanmışken, HA-M2 protein çiftleri arasındaki korelasyon ise düşük olarak belirlenmiştir. HA proteini ile M1 proteini arasındaki korelasyon ise çok düşük olmuştur. M1 proteini sadece M2 proteini ile mükemmel derecede etkileşim kurmuşken, NA, NP, NEP, NS1, PA, PAX, PB1 ve PB2 proteinleri ile çok düşük etkileşim meydana getirmiştir. M2-PB2 protein çifti arasında yüksek korelasyon gözlemlenmişken, M2-NS1 ve M2-PB1 protein çiftleri arasında düşük korelasyon elde edilmiştir. M2 proteini geri kalan proteinler (NA, NP, NEP, PA ve PAX) ile çok düşük düzeyde etkileşim kurmuştur. NA-NEP, NA-NS1, NA-PA, NA-PAX ve NA-PB2 protein çiftleri arasında yüksek korelasyon elde edilmişken, NA-PB1 protein çifti arasında orta derecede korelasyon hesaplanmıştır. Son olarak, NA proteini NP proteini ile çok düşük etkileşim sergilemiştir. NP-PAX protein çifti arasında yüksek derecede etkileşim ağı oluşturulmuştur. NP-PB2 protein çifti arasındaki etkileşim ise orta dereceyken, NP-NEP, NP-NS1, NP-PA ve NP-PB1 protein çiftleri arasındaki etkileşim düşük olmuştur. NEP proteini genel olarak başarılı bir etkileşim göstermiştir. NEP-NS1, NEP-PA, NEP-PAX ve NEP-PB2 protein çiftleri arasındaki korelasyon değeri yüksek bulunmuştur. NEP-PB1 protein çifti arasındaki korelasyon mükemmel olmuştur. NS1 proteini PA ve PB2 proteinleri ile mükemmel seviyede etkileşim kurmuşken, PAX ve PB1 proteinleri ile yüksek

düzeyde etkileşim sergilemiştir. PA-PAX protein çifti arasında ise orta düzeyde korelasyon elde edilmiştir. PA-PB1 ve PA-PB2 protein çiftleri arasında ise korelasyon yüksek olmuştur. PAX proteini hem PB1 hem de PB2 proteinleri ile orta düzeyde etkileşim oluşturmuştur. Son olarak PB1 ve PB2 proteinleri arasında yüksek etkileşim gözlemlenmiştir.

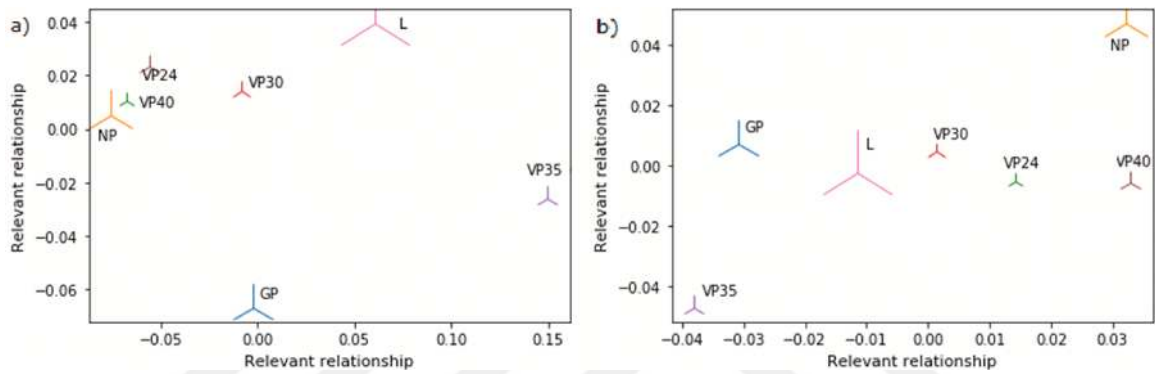
6.3. Uygulama Sonuçlarının Değerlendirilmesi

Birlikte evrim analizi Ebola virüsünde bulunan NP proteinin diğer proteinler ile güçlü etkileşim kurduğunu göstermiştir. [209] numaralı çalışmada araştırmacılar NP, VP24, VP30 ve VP40 proteinlerinin yüksek etkileşim oluşturduğunu kanıtlamıştır. Birlikte evrim analizi bu etkileşimleri doğrulamıştır. Önerilen protein sayısallaştırma yöntemi ile GP proteinin diğer proteinler ile mükemmel etkileşim kurduğu gözlemlenmiştir. Ebola virüsünde GP proteini viral olan tek proteindir ve reseptör bağlanmasından sorumludur. Virüsün, konak organizmanın hücrelerine bağlanmasını sağlamaktadır. Bundan dolayı, GP proteinin diğer proteinler ile etkileşimi güçlü olmalıdır [209]. Birlikte evrim analizi sonuçları bu etkileşimi doğrulamıştır. Bunlara ek olarak, EBOV genomunun transkripsiyonu ve replikasyonu VP24 proteini ile gerçekleştirilmektedir. VP24 bu işlemleri NP, VP30 ve VP40 proteinleri ile etkileşime girerek gerçekleştirmektedir [210]. Önerilen sayısallaştırma yöntemiyle gerçekleştirilen birlikte evrim analizi sonuçları bu ilişkileri doğrulamıştır. Ebola virüsüne ait proteinler arasındaki etkileşimleri birlikte evrim analizi ile belirleyebilmek için bu alanda etkin bir şekilde kullanılan MirrorTree yaklaşımına da başvurulmuştur. MirrorTree yaklaşımının hem önerilen hem de diğer protein sayısallaştırma yöntemlerinden farkı, sayısallaştırma işlemi yerine ÇDH ile birlikte evrim analizi yapmasıdır. Tablo 6.15'te MirrorTree yaklaşımı ile gerçekleştirilen birlikte evrim analizi skorları verilmiştir.

Tablo 6.15. MirrorTree yaklaşımı ile gerçekleştirilen Ebola virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	GP	NP	VP24	VP30	VP35	VP40	L
GP	1	0,97	0,94	0,97	0,98	0,91	0,98
NP		1	0,98	0,99	0,70	0,96	0,99
VP24			1	0,98	0,97	0,98	0,97
VP30				1	0,98	0,97	0,98
VP35					1	0,95	0,99
VP40						1	0,95
L							1

Tablo 6.15’te verilen sonuçlar doğrultusunda MirrorTree yaklaşımıyla NP-VP35 protein çifti hariç bütün protein çiftleri arasında mükemmel derecede etkileşim kurulmuştur. NP-VP35 protein çifti arasında ise yüksek oranda korelasyon görülmüştür. Sonuçlar genel olarak incelendiğinde MirrorTree yaklaşımının çok başarılı olduğu gözlemlenmiştir. Daha önceden de belirtildiği üzere, bu yaklaşım ile yüksek oranda yanlış pozitif oranlar elde edilebilmektedir. Yanlış pozitif oranlarının yüksek olup olmadığını belirlemek için ÇBÖ (Çok Boyutlu Ölçekleme) analizi uygulanmıştır. Şekil 6.3’te hem önerilen yöntem hem de MirrorTree yaklaşımı için uygulanan ÇBÖ analizinin sonuçları verilmiştir.



Şekil 6.3. ÇBÖ analizi. a) Önerilen protein sayısalştırma yöntemi. b) MirrorTree

Şekil 6.3a’da GP ve VP35 dışındaki protein çiftleri arasında yüksek bir ilişki olduğu görülmektedir. Birlikte evrim analizi bu sonuçları doğrulamış ve önerilen yöntem ile GP proteininin diğer proteinler ile iyi etkileşim kurduğu belirlenmiştir. Bununla birlikte, ÇBÖ analizi bu ilişkileri doğrulamamıştır, çünkü ÇBÖ sonucuna göre GP proteini diğer proteinlerle iyi etkileşim kuramamıştır. Bundan dolayı, önerilen protein sayısalştırma yöntemi, GP proteinin etkileşimi için yüksek oranda yanlış pozitif değerler oluşturmuştur. Öte yandan, MirrorTree yaklaşımıyla gerçekleştirilen birlikte evrim analizine göre Ebola virüsüne ait tüm proteinler birbirleriyle mükemmel etkileşim kurmuştur. Ancak, ÇBÖ sonuçlarına göre böyle bir durum söz konusu olmamıştır. Şekil 6.3b’ye göre tüm proteinler farklı bölgelerde yer aldığı için aralarında doğrudan ya da net bir ilişki görülmemiştir. Bu durum, MirrorTree yaklaşımının Ebola virüsü için yüksek oranda yanlış pozitif oranlar ürettiğini göstermektedir.

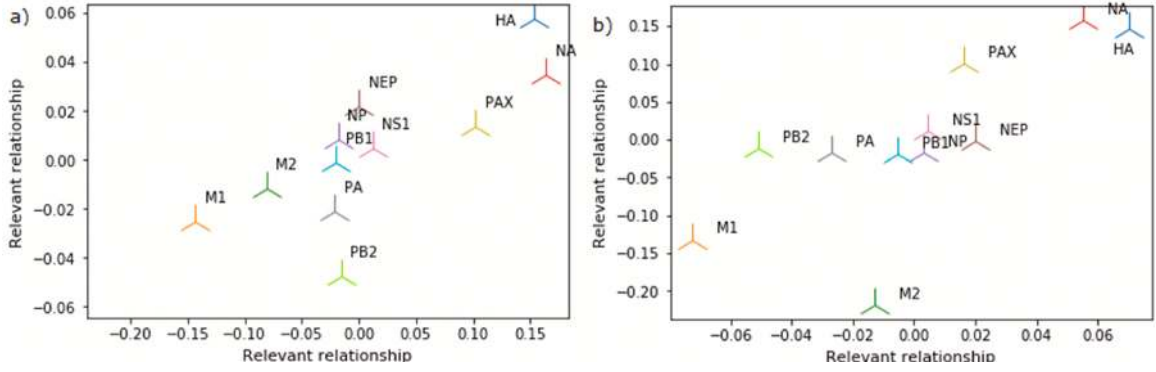
Birlikte evrim analizi ile tahmin edilen Influenza A virüsüne ait protein etkileşimleri deneysel yaklaşımlarla da kanıtlanmıştır. HA proteini sialik aside bağlanarak virüs enfeksiyonundan sorumludur [211]. NA proteini ise bu sialik asidi parçalamakla görevlidir [212]. Bu nedenlerden dolayı bu iki protein birbiriyle etkileşim halindedir. Ayrıca deneysel çalışmalara göre HA-PB2 protein çifti, virüs patojenitesi için etkileşim kurmaktadır [213]. Deneysel

çalıřmalarda mutasyonlara dayalı M1 ve M2 proteinleri arasındaki etkileřimler kanıtlanmıřtır [214]. NS1 ve NEP proteinleri, Y2H yöntemiyle etkileřim kurduđu için mükemmel bir iliřkiye sahip olmaktadır [215]. NEP proteini ise NS1, PB1 ve PB2 proteinleri ile iyi bir etkileřim kurmaktadır [216]. Genel olarak, önerilen protein sayısallařtırma yöntemi, protein çiftleri arasındaki bu etkileřimleri dođrulamıřtır. Ebola virüsünde olduđu gibi önerilen protein sayısallařtırma yönteminin performansı MirrorTree yaklařımıyla kıyaslanmıřtır. Tablo 6.16’da MirrorTree yaklařımının birlikte evrim analizi sonuçları verilmiřtir.

Tablo 6.16. MirrorTree yaklařımı ile gerekleřtirilen Influenza A virüsüne ait proteinlerin birlikte evrim analizi sonuçları

	HA	M1	M2	NA	NP	NEP	NS1	PA	PAX	PB1	PB2
HA	1	0,65	0,65	1,00	0,81	0,84	0,88	0,79	0,99	0,82	0,76
M1		1	0,90	0,63	0,86	0,85	0,87	0,87	0,79	0,90	0,90
M2			1	0,63	0,85	0,85	0,85	0,83	0,29	0,88	0,83
NA				1	0,80	0,83	0,87	0,78	0,99	0,81	0,81
NP					1	0,96	0,97	0,97	0,97	0,99	0,93
NEP						1	0,97	0,95	0,97	0,97	0,92
NS1							1	0,96	0,97	0,98	0,93
PA								1	0,97	0,97	0,97
PAX									1	0,97	0,97
PB1										1	0,96
PB2											1

Tablo 6.16’dan elde edilen sonuçlar ıřıđında, MirrorTree yaklařımıyla elde edilen analiz sonuçlarının mükemmel olduđu gözlemlenmiřtir. MirrorTree yaklařımının önerilen protein sayısallařtırma ve diđer protein sayısallařtırma yöntemlerinden ok daha etkili olduđu görölmüřtür. Elde edilen bu sonuçların yanlış pozitif olup olmadıđını belirlemek için, Ebola virüsünde olduđu gibi BÖ analizi yapılmıřtır. řekil 6.4’te hem önerilen yöntem hem de MirrorTree yaklařımı için uygulanan BÖ analizinin sonuçları verilmiřtir.



Şekil 6.4. ÇBÖ analizi. a) Önerilen protein sayısallaştırma yöntemi. b) MirrorTree

Şekil 6.4'e bakıldığında, bazı proteinlerin birbirleriyle etkileşim kurduğu, bazılarının ise etkileşim kurmadığı görülmüştür. Örnek verilecek olursa, önerilen yöntemle HA-NA protein çiftleri arasında 0,85'lik bir korelasyon değeri gözlemlenmiştir. Başka bir örnekte ise, NA-NP protein çiftleri arasında ise 0,35 korelasyon değeri hesaplanmıştır. Şekil 6.4a incelendiğinde, bu sonuçların ÇBÖ analizi ile uyumlu olduğu görülmektedir. Ancak benzer bir çıkarım MirrorTree yaklaşımı için söz konusu olmamıştır. MirrorTree yaklaşımının birlikte evrim analizi sonuçlarına göre, tüm proteinler birbirleriyle güçlü bir etkileşim göstermiştir. Ancak, Şekil 6.4b'de böyle bir durum gözlemlenmemiştir. ÇBÖ analizine göre, M1 ve NEP proteinleri arasındaki mesafe oldukça büyüktür. Bu durumda bu iki proteininin etkileşim halinde olması mümkün değildir. Tablo 6.16'da verilen sonuçlara göre bu iki protein arasındaki etkileşimin mükemmel olması gerekmektedir. Benzer bir çıkarımı M2-PB2 protein çiftleri için de yapılabilir. Bu nedenle MirrorTree yaklaşımı yüksek oranda yanlış pozitif oranlar üretmiştir.

İki virüs tipi için sonuçlar incelendiğinde önerilen yöntemin etkili olduğu gözlemlenmiştir. Evrim-tabanlı protein sayısallaştırma yöntemlerinin yeterince etkili olamamasının nedeni, bu yöntemlerin genellikle protein hizalama çalışmalarında kullanılması olabilir. Hizalama işlemlerinde genellikle iki amino asit dizilimi arasındaki benzerlikler belirlenir ve puan oluşturulur. Ancak birlikte evrim analizinde benzerlik yapısı bulunmamaktadır. Bunlara ek olarak, proteinler arasındaki etkileşimler belirlenirken bir hizalama işlemi de gerçekleştirilmemiştir. Bu nedenler evrim-tabanlı protein sayısallaştırma yöntemlerinin etkili olamamasına sebebiyet vermiş olabilir. Yapı-tabanlı protein sayısallaştırma yöntemleri de evrim-tabanlı yöntemler gibi etkisiz olmuştur. Bunun en büyük nedenlerinden birisi, çalışmada proteinlerin yapı bilgisine ihtiyaç duyulmaması olabilir. Yapı-tabanlı yöntemler genel olarak protein ailelerini tahmin etme çalışmalarında başarılı sonuçlar vermektedir. Bu tür çalışmalarda proteinlerin birincil ya da ikincil yapı bilgileri gerekmektedir. Bu iki kategori dışında en etkili olan protein sayısallaştırma yöntemleri fizikokimyasal-tabanlı yöntemler olmuştur. Bunun en büyük nedeni, proteinler arasındaki etkileşimler belirlenirken kimyasal yaklaşımlara ihtiyaç duyulmasıdır. Proteinler arasındaki

etkileşimleri hesaplama-tabanlı yöntemler ile tahmin etmek ya da belirlemek için ön bilgi gerekmektedir. Bu ön bilgi kimyasal analizler ile elde edilmektedir. Ancak, önerilen protein sayısallaştırma yöntemi en iyi performansı sergilemiştir. Bu başarımın nedeni şu şekilde sıralanabilir;

- Önerilen yöntemde belirli bir protein bilgisi gerekmemektedir. Sayısallaştırma işlemi entropi hesaplamasına dayandığı için, dizilimde bilgi kaybı az olmaktadır. Bu nedenle, sayısallaştırma işlemi daha güvenilir ve etkili olmaktadır.
- Mevcut sayısallaştırma yöntemlerinden farklı olarak önerilen yöntem dinamik ve adaptif bir yapıya sahiptir. Mevcut yöntemlerin çoğunda dizilimdeki amino asitler sabit değerlere sahipken önerilen protein sayısallaştırma yönteminde amino asitlerin değeri değişmektedir. Önerilen yöntemde amino asitlerin sayısal değerleri dizilimin uzunluğuna ve amino asitlerin tekrarlanma sıklığına bağlıdır. Bu şekilde farklı amino asit dizilimleri aynı sayısal değeri almayabilir. Bu durumda dejenerasyon görülmektedir ve bilgi kaybı olmamaktadır.

Önerilen yöntemin bu avantajları, proteinler arasındaki etkileşimin başarılı bir şekilde belirlenmesinde önemli bir faktör olmuştur.

Sonuç olarak bu çalışmada, birlikte evrim analizinde kullanılmak üzere yeni bir entropi tabanlı protein sayısallaştırma yöntemi önerilmiştir. İlk aşamada, protein dizilimleri hem önerilen hem de çeşitli protein sayısallaştırma yöntemleri ile sayısal hale getirilmiştir. Daha sonra sayısallaştırılan her protein dizilimine Fourier dönüşümü uygulanmış ve dizilimler arasındaki uzaklık değerleri Öklid ile hesaplanarak uzaklık matrisi oluşturulmuştur. Son aşamada ise, uzaklık matrislerinden Pearson korelasyon değerleri hesaplanmış ve birlikte evrim analizi gerçekleştirilmiştir. Çalışmada Ebola ve Influenza A olmak üzere iki virüse ait proteinler değerlendirilmiştir. Önerilen yöntem, en yüksek etkileşimi NP-VP24 ve NP-VP40 protein çiftleri arasında tahmin etmiş ve 0,99 korelasyon değerine ulaşmıştır. Benzer şekilde, diğer protein sayısallaştırma yöntemleri ile de bu protein çiftleri arasında mükemmel derecede etkileşimler gözlemlenmiştir. En kötü etkileşim sonucu 0,75 korelasyon değeri ile NP-VP35 proteinleri arasında görülse bile bu değer yüksek olarak kabul edilmektedir. Sonuçlar genel olarak incelendiğinde, Ebola virüsüne ait proteinlerin etkileşimlerini belirlemede en başarısız yöntemlerin fizikokimyasal-tabanlı ve yapı-tabanlı yöntemler olduğu görülmüştür. Bu yöntemler dışında hiçbir protein sayısallaştırma yönteminde orta derecede etkileşim gözlemlenmemiştir. Ayrıca protein sayısallaştırma yöntemlerinin sonuçları MirrorTree yaklaşımı ile de kıyaslanmıştır. MirrorTree yaklaşımına göre tüm protein çiftlerinin (NP-VP35 hariç) korelasyon değeri 0,90'ın üzerinde olmuştur. Ancak, MirrorTree yaklaşımı ÇDH kullanmaktadır. Bundan dolayı yanlış pozitif değerler elde edilmiştir. Influenza A virüsü için de aynı analiz işlemi gerçekleştirilmiştir. Birlikte evrim analizine göre, önerilen yöntemle en iyi etkileşim M1-M2 protein çiftleri arasındaki gerçekleştirilmiştir. Sonuçlar genel olarak incelendiğinde hem yüksek hem de düşük derecede

etkileşimlerin olduğu görülmüştür. Influenza A virüsündeki etkileşimler de deneysel olarak ispat edilmiştir. Bu, hesaplama-tabanlı yöntemlerin proteinler arasındaki etkileşimleri belirlemede etkili olduğunu kanıtlamıştır. Ancak çalışmadan da anlaşıldığı üzere seçilen protein sayısallaştırma yönteminin başarı üzerinde etkisi büyük olmaktadır. MirrorTree ile gerçekleştirilen birlikte evrim analizine göre bütün protein etkileşimleri güçlü olmuştur. Ancak Ebola virüsünde olduğu gibi, bu sonuçlarda yanlış pozitif oranlardan oluşmuştur. Çalışmanın sonuçları önerilen protein sayısallaştırma yönteminin başarılı olduğunu sergilemiştir.



7. FİBONACCİ HASH TABANLI YENİ BİR SAYISALLAŞTIRMA YÖNTEMİ

Tezin bu bölümünde, mevcut protein sayısallaştırma yöntemlerine ek olarak, Fibonacci hash tabanlı yeni bir protein sayısallaştırma yöntemi önerilmiştir. Önerilen bu yöntem, protein ailelerini sınıflandırmak için değerlendirilmiş ve yöntemin literatürdeki diğer protein sayısallaştırma yöntemleri ile kıyaslanması gerçekleştirilerek başarımı belirlenmiştir. Önerilen protein sayısallaştırma yöntemi, Fibonacci sayılarına ve hash tablosuna dayanmaktadır. Önerilen yöntemin içeriğinde bulunan hash tablosu ve Fibonacci sayıları algoritma analizinde sıklıkla kullanılmaktadır. Hash tablosu verilerin hesaplama karmaşıklığını azaltmak için kullanılan bir veri yapısıdır. Bunun yanı sıra, Fibonacci sayıları ise arama algoritması olarak kullanılabilir. Önerilen bu hibrit yöntem bu algoritmik yapısı nedeniyle yeni bir kategorisel yaklaşım olarak tanımlanan algoritmik-tabanlı protein sayısallaştırma kategorisinde değerlendirilmiştir.

7.1. Önerilen Sayısallaştırma Yöntemi

Tez çalışması kapsamında yeni bir protein sayısallaştırma yöntemi önerilmiş ve önerilen yaklaşım Fibonacci sayıları ve hash tablosu kullanılarak tasarlanmıştır. Fibonacci sayıları, her terimi kendisinden önce gelen iki sayının toplamından oluşan bir seridir. Denklem 7.1’de Fibonacci sayılarının hesaplanması verilmiştir;

$$F(n) = F_{n-1} + F_{n-2}, F_1 = 1 \text{ ve } F_2 = 1, n > 3 \quad (7.1)$$

Denklem 7.1’de n değeri n . sayıyı ifade etmektedir. Doğada Fibonacci serisi net bir şekilde gözlemlenmektedir. Fibonacci sayıları çiçeklerde, insan organlarında, fizikte ve birçok farklı alanda görülmektedir [217]. Fibonacci sayıları o kadar yaygındır ki, bu benzerlik proteinlerde dahi görülmektedir. Altın oranla ilgili bazı çalışmalar, Fibonacci sayılarının, organizmaların yapılarının yeniden şekillenmesinde kusursuz bir role sahip olduğunu göstermiştir. İnsan genomunun Fibonacci serisini ve altın oranı taklit eden davranış sergilediği gözlemlenmiştir [218]. Başka bir çalışmada ise araştırmacılar, DNA dizilimlerinde gen kodlayan bölgelerin Fibonacci sayıları ile yüksek oranda ilişkili olduğunu göstermiştir [219]. [220] numaralı çalışmada araştırmacılar Fibonacci dizilimini kullanarak, amino asit dizilimlerinin genetik kodlarını belirlemişlerdir. Çalışmada ilk olarak, protein dizilimleri karbon, nitrojen, oksijen ve kükürt dahil olmak üzere atom numaralarına göre sıralanmıştır. Daha sonra protein dizilimleri küçük ve büyük olmak üzere iki farklı atoma dayalı olarak Fibonacci zinciri ile sayısallaştırılmıştır. Bu çalışmalardan ve gözlemlerden esinlenerek, bu çalışmada protein dizilimleri Fibonacci sayıları kullanılarak sayısal hale getirilmiştir. Proteinler, yaşamı sürdürmek için çeşitli fonksiyonları yerine getirdikleri için

temel moleküllerdir. Moleküllerin bir hücreden diğer hücreye taşınmasından, DNA replikasyonundan, metabolik reaksiyonların hızlanmasından ve diğer birçok önemli fonksiyonlardan sorumludurlar. Kabul edilen ve çalışmalarda kullanılan amino asit sayısı 20 olduğu için, çalışmada 1'den 20'ye kadar olan değerler hesaplanmıştır. Hesaplama işleminin ardından her bir amino aside Fibonacci sayıları atanmıştır. Amino asitler alfabetik bir şekilde sıralanmıştır. Tablo 7.1'de amino asitlerin Fibonacci değerleri verilmiştir.

Tablo 7.1. Amino asitlerin Fibonacci değerleri

Amino Asit Kodu	Fibonacci Değeri	Amino Asit Kodu	Fibonacci Değeri
A	1	M	89
C	1	N	144
D	2	P	233
E	3	Q	377
F	5	R	610
G	8	S	987
H	13	T	1597
I	21	V	2584
K	34	W	4181
L	55	Y	6765

Fibonacci sayılarının 12. indeksten sonraki ondalık basamaklarının boyutu nedeniyle, bu değerler hash tablosuna yerleştirilmiştir. Fibonacci sayıları hash tablosuna yerleştirilmeden kullanıldığında, hesaplama süreci karmaşık olmakta ve zaman açısından verimli olmamaktadır. Bu nedenle Fibonacci değerlerinin hash tablosuna yerleştirilmesi gerekmiştir. Hash tablosu verileri ilişkili bir şekilde saklayan bir veri yapısıdır. Hash tablosunda her bir değer, kendine özgü bir indekste tutulmaktadır. Verinin boyutuna bakılmaksızın, ekleme ve arama işlemleri hızlı olduğu için verilere çabuk erişilmektedir [221]. Değerler hash tablosuna hash yöntemi ile eklenmektedir. Bir yöntemde mod operatörü kullanılmakta ve değerler ona göre tabloya yerleştirilmektedir. Örneğin [26, 13, 35, 44, 12] şeklinde verilen bir dizi hash tablosuna eklendiğinde Tablo 7.2'deki durum elde edilir.

Tablo 7.2. Hash yöntemi örneği

Değer	Hash	İndeks
26	26%5	1
13	13%5	3
35	35%5	0
44	44%5	4
12	12%5	2

Ancak bazı durumlarda, değerler aynı indeksi paylaşmak zorunda kalmaktadır. Bu duruma çakışma denir. Bu çakışmaları çözmek için doğrusal sondalama, karesel sondalama gibi yöntemler kullanılır [221]. Önerilen yöntemde de çakışma olduğu için, bu çakışma doğrusal sondalama işlemi ile giderilmiştir. Doğrusal sondalama işleminde, çakışmaya neden olan değer, olması gereken indeksten sonraki ilk boş indekse yerleştirilir. 20 adet amino asit olduğu için 20 boyutunda bir hash tablosu geliştirilmiş ve değerler bu hash tablosuna yerleştirilmiştir. Tablo 7.3'te amino asitlerin FIBHASH değerleri verilmiştir.

Tablo 7.3. Amino asitlerin FIBHASH değerleri

Amino asitler	Amino Asit Kodları	Sayısallaştırma	Hash	Çakışma (eğer varsa)	Index
Alanin	A	1	1%20	-	1
Sistein	C	1	1%20	+1	2
Aspartik asit	D	2	2%20	+1	3
Glutamik asit	E	3	3%20	+1	4
Fenilalanin	F	5	5%20	-	5
Glisin	G	8	8%20	-	8
Histidin	H	13	13%20	-	13
İzolösin	I	21	21%20	+5	6
Lizin	K	34	34%20	-	14
Lösin	L	55	55%20	-	15
Metiyonin	M	89	89%20	-	9
Asparajin	N	144	144%20	+3	7
Prolin	P	233	233%20	+3	16
Glutamin	Q	377	377%20	-	17
Arjinin	R	610	610%20	-	10
Serin	S	987	987%20	+4	11
Treonin	T	1597	1597%20	+1	18
Valine	V	2584	2584%20	+8	12
Triptofan	W	4181	4181%20	+18	19
Tirozin	Y	6765	6765%20	+15	20

Tablo 7.4’te verilen deęerler doęrultusunda, $P(N) = [M, W, F, Y, P, C, \dots]$ řeklindeki bir protein dizilimi, FIBHASH yntemi ile $M(N) = [9, 19, 5, 20, 16, 2, \dots]$ biřiminde sayısallařtırılmaktadır.

7.2. Protein Ailelerine Uygulanması

Protein ailelerinin sınıflandırılması, biyoenformatik alıřmaları iin nemli bir arařtırma konusudur. Protein aileleri, ok sayıda proteinlerden meydana gelmektedir. Sınıflandırma iřlemi esnasında, benzer fonksiyonlara sahip olan protein aileleri bir grup olarak ifade edilmektedir. Fakat, bir protein ailesinin yapısal bilgilere dayalı olarak sınıflandırılması, karakterize edilmemiř protein dizilimlerinin elde edilmesinin zor olması nedeniyle, saęlıklı olamamaktadır [77,78]. Bundan dolayı, alıřmalarda eřitli yntemler denenmektedir. Arařtırılan yntemler doęrultusunda kullanılan en popler yaklařım ise protein dizilimleriyle gerekleřtirilen alıřmalardır [79-81]. Bunun yanı sıra, protein fonksiyonlarını belirlemek iin genellikle 3-boyutlu protein yapıları kullanılmaktadır. Bu da srecin maliyetli olmasına ve zaman almasına neden olmaktadır. Bu problemlerden dolayı protein dizilimlerine ynelik yaklařımların nemi ve poplaritesi artmıřtır. Protein dizilimlerini analiz etmek ve protein dizilimlerinden anlamlı, ayrıık ve srekli znitelikler toplamak iin, znitelik mhendislięine bařvurulmaktadır. Gereklı znitelikler elde edildikten sonra ise, derin ęrenme ve makine ęrenmesi algoritmaları kullanılmakta ve protein aileleri sınıflandırılmaktadır. Ancak, sınıflandırma algoritmalarının kullanılabilmesi iin, protein dizilimlerinin sayısallařtırılması gerekmektedir. Bu nedenle bu alıřmalarda protein sayısallařtırma yntemleri sıklıkla kullanılmaktadır. Gerekleřtirilen alıřmada da protein dizilimleri sayısal hale getirilmiř ve yinelemeli sinir aęları ve birincil protein dizilimleri kullanılarak protein aileleri sınıflandırılmıřtır. İlk ařamada amino asit dizilimleri CPNR, EIIP, hidrofobiklik, ikili-kodlama, Atchley faktrleri, PAM250, BLOSUM62, rastgele kodlama ve FIBHASH protein sayısallařtırma yntemleri ile sayısal hale getirilmiřtir. Sayısallařtırma iřleminin ardından veriler yinelemeli sinir aęı ile sınıflandırılmıřtır. Sınıflandırıcıların ve protein sayısallařtırma yntemlerinin bařarımları doęruluk, kesinlik, F1-skor, duyarlılık ve AUC skorları ile belirlenmiřtir. Gerekleřtirilen alıřmanın ana katkıları řu řekilde zetlenebilir;

- Amino asit dizilimlerini sayısal hale getirmek iin yeni protein sayısallařtırma yntemi nerilmiřtir.
- Protein aileleri sadece birincil yapılar kullanılarak tahmin edilmiřtir.

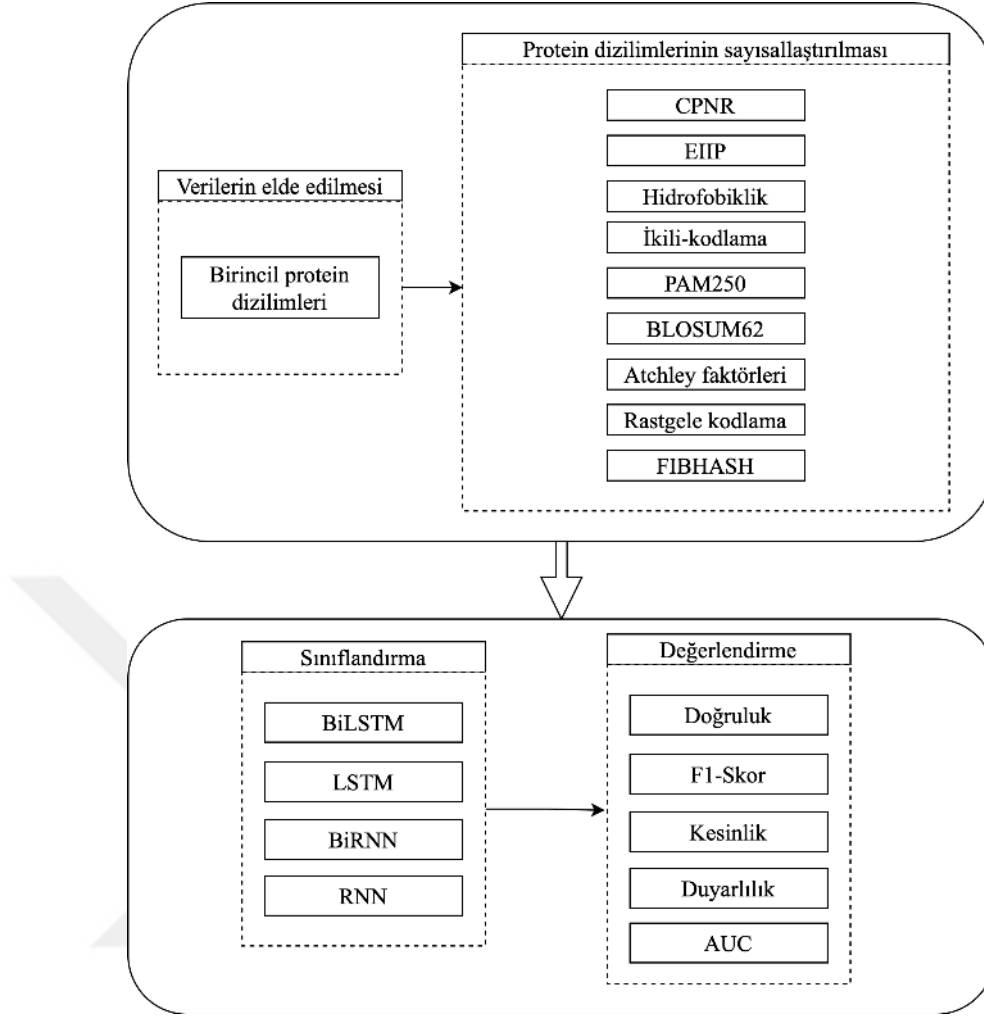
alıřmada 60 farklı protein ailesine ait toplamda 97.576 adet protein dizilimi kullanılmıřtır. Veriler UniProt veri setinden elde edilmiřtir. UniProt veri seti protein dizilimlerini ve fonksiyonel bilgilerini ieren ve serbeste eriřilebilen bir veri setidir. Veri seti, proteinlerin biyolojik fonksiyonları, etkileřimleri ve aileleri hakkında byk miktarda bilgi ieren birok biyoenformatik

projelerinden elde edilen proteinlerden oluşmaktadır. Tablo 7.4'te çalışmada kullanılan protein aileleri ve bu ailelere ait protein dizilimlerinin adeti verilmiştir.

Tablo 7.4. Çalışmada kullanılan protein aileleri

Protein ailesi	Adet	Protein ailesi	Adet
Sınıf-II aminosil-tRNA sentatazları	7,335	Evrensel ribozomal protein uS3	993
ABC taşıyıcı	5,707	Krueppel C2H2-tip çinko-parmak protein	990
Sınıf-I aminosil-tRNA sentatazları	5,656	LDH	981
G proteini kenetli reseptör 1	5245	MDH	981
Sınıf I benzeri SAM-bağlayıcı metiltransferaz	4,572	Mitokondriyal salım faktörü	960
Protein kinaz	4,480	Enolaz	915
TRAFAC sınıfı translayon faktörü GTPaz	4,179	Evrensel ribozomal protein uL1	900
Radikal SAM	2,480	TRAFAC sınıfı OBG-HflX benzeri GTpaz	889
Prokaryor	2,408	FKBP-tip PPIaz	888
TrmE-Era-EngA-EngB-Septin-benzeri GTPaz	2,293	MnmG	888
Sitokrom b	2,242	NAD-bağımlı DNA ligaz	876
Ana kolaylaştırıcı	2,191	Evrensel ribozomal protein uL3	867
Siklohidraz	2,190	Evrensel ribozomal protein uL4	848
Metallo-bağımlı hidrolazlar	1,859	DNA yanlış eşleşme tamiri MutS	843
Transferaz heksapeptit tekrarı	1,772	Tetrahidrofolat dehidrojenaz	842
Sitokrom P450	1,500	SHMT	836
DEAD kutusu sarmalı	1,460	RecA	833
MurCDEF ailesi	1,405	Adenylosukinat sentaz	829
Isı şoku proteini 70	1,378	Tetrahidrofolat dehidrojenaz siklohidrolaz	828
EPSP sintaz	1,218	EF-Ts	826
AB hidrolaz	1,206	SecA	818
Evrensel ribozomal protein uS4	1,150	RNA polimeraz alfa zinciri	802
GHMP kinaz	1,143	IPP transferaz	801
Şaperonin (HSP60)	1,125	Glikosiltransferaz 28	787
Evrensel ribozomal protein uS2	1,094	Poliribonükleotid nükleotidiltransferaz	783
Sınıf III piridoksal-fosfat	1,081	Fmt	765
Kısa zincirli dehidrojenazlar	1,070	ATCase	705
Evrensel ribozomal protein uL2	1,055	Metiltransferaz üst ailesi	519
Metiltiyotransferazlar	1,053	OTCase	406

Tablo 7.4'te verilen her bir aileye ait protein dizilimleri CPNR, EIIP, hidrofobiklik, ikili-kodlama, Atchley faktörleri, PAM250, BLOSUM62, rastgele kodlama ve önerilen FIBHASH yöntemi ile sayısal hale getirilmiş ve sınıflandırılmıştır. Sınıflandırma işlemi ile protein dizilimlerinin hangi aileye ait oldukları tahmin edilmiştir. Çalışmanın akış şeması Şekil 7.1'de verilmiştir.



Şekil 7.1. Önerilen FIBHASH yönteminin derin öğrenme modelleriyle değerlendirilme süreci

7.2.1. Derin Öğrenme Kullanılarak Protein Ailelerinin Sınıflandırılması

Bu kısımda kullanılan her bir sınıflandırıcının uygulama sonuçları verilmiştir. Sınıflandırıcıların başarımları doğruluk, F1-skör, kesinlik, duyarlılık ve AUC skoru ile belirlenmiştir. Sınıflandırma işlemi dört farklı senaryo üzerinden gerçekleştirilmiştir. Birinci senaryoda RNN, ikinci senaryoda LSTM, üçüncü senaryoda BiRNN ve son senaryoda ise BiLSTM sınıflandırıcıları kullanılmıştır. RNN ile gerçekleştirilen sınıflandırma işleminde veri setinin %80'i eğitim için geri kalan %20'si ise test için kullanılmıştır. RNN sınıflandırıcısının parametreleri deneme-yanılma yaklaşımı ile belirlenmiştir. Eğitim işlemi 200 iterasyon ile gerçekleştirilmiştir. RNN sınıflandırıcısının parametreleri şu şekildedir;

- Girdi katmanında 97.576 adet protein dizilimi değerlendirilmiştir.
- İkinci katmanda 256 RNN birimi kullanılmış ve aktivasyon işlemi ReLU ile yapılmıştır.
- Ardından iki adet tam-bağlantılı katman kullanılmış ve bu katmanlarda sırasıyla 1.024 ve 512 adet nöron değerlendirilmiştir.

- Son katmanda ise Softmax aktivasyon fonksiyonu kullanılmış ve sınıflandırma işlemi gerçekleştirilmiştir.
- Modelin kaybı kategorik çapraz entropi ile belirlenmiştir.
- En iyileme algoritması olarak Adam fonksiyonu kullanılmıştır.

RNN ile gerçekleştirilen sınıflandırma işleminin sonuçları Tablo 7.5'te verilmiştir.

Tablo 7.5. RNN sınıflandırıcısının sonuçları

Sayısallaştırma Yöntemi	Doğruluk	F1-Skor	Kesinlik	Duyarlılık	AUC
EIIP	%34,46	%21,08	%25,23	%18,11	0,52
Hidrofobiklik	%48,97	%37,89	%57,79	%28,19	0,58
CPNR	%49,76	%37,62	%61,75	%27,06	0,61
BLOSUM62	%46,02	%19,98	%55,27	%12,20	0,53
Atchley faktörleri	%55,15	%40,20	%69,70	%28,25	0,65
PAM250	%50,55	%44,74	%37,97	%54,46	0,55
İkili kodlama	%55,19	%46,10	%68,09	%31,53	0,64
Rastgele kodlama	%46,00	%41,66	%52,52	%34,53	0,52
FIBHASH	%57,00	%44,38	%65,36	%33,60	0,67

Tablo 7.5'te verilen sonuçlar doğrultusunda hiçbir protein sayısallaştırma yöntemi etkili bir sınıflandırma işlemi gerçekleştirememiştir. En düşük doğruluk skoru EIIP protein sayısallaştırma yöntemi ile elde edilmiş ve bu yöntem ile %34,46 oranında doğruluk skoru gözlemlenmiştir. Bu yöntem ile elde edilen AUC skoru ise 0,52 olmuştur. Ardından en başarılı sınıflandırma işlemi ise Rastgele kodlama yöntemi ile hesaplanmıştır. Rastgele kodlama yöntemi ile %46,00 doğruluk skor elde edilmişken, 0,52 AUC skoru gözlemlenmiştir. Bunun yanı sıra, hidrofobiklik, CPNR ve BLOSUM62 yöntemleri de %50'nin üzerinde doğruluk sergileyememişlerdir CPNR yöntemi ile %46,76 doğruluk hesaplanırken, hidrofobiklik ve BLOSUM ile sırasıyla. %48,97 ve %46,02 doğruluk elde edilmiştir. Öte yandan, hidrofobiklik ve BLOSUM62 yöntemlerinin AUC skoru 0,5'i geçemezken, CPNR yönteminin AUC skoru 0,61 olmuştur. PAM250 protein sayısallaştırma yöntemi ile %50,55 doğruluk hesaplanırken, AUC skoru ise 0,55 olmuştur. Atchley faktörleri ve ikili kodlama yöntemleri birbirine yakın doğruluk skoru sergilemişlerdir. Atchley yönteminin doğruluk skoru %55,15'ken, ikili kodlamanın doğruluk skoru %55,19'dur. Bu yakınlık AUC skoruna da yansımış ve sırasıyla 0,65 ve 0,64 AUC skorları elde edilmiştir. En iyi başarımlar önerilen FIBHASH yöntemi ile elde edilmiştir. FIBHASH yöntemi ile %57,00 doğruluk hesaplanırken, 0,67 AUC skoruna ulaşılmıştır.

İkinci senaryoda ise LSTM derin öğrenme modeli kullanılmıştır. RNN sınıflandırıcısında olduğu gibi bu senaryoda da veri setinin %80'i eğitim için kullanılırken %20'si test için değerlendirilmiştir. Bu sınıflandırıcı için de 200 iterasyon kullanılmış ve parametreler deneme-yanılma yaklaşımı ile belirlenmiştir. LSTM derin öğrenme modelinin parametreleri aşağıda verilmiştir;

- Girdi katmanında 97.576 adet protein dizilimi değerlendirilmiştir.
- İkinci katmanda 256 LSTM birimi kullanılmış ve aktivasyon işlemi ReLU ile yapılmıştır.
- Ardından iki adet tam-bağlantılı katman kullanılmış ve bu katmanlarda sırasıyla 1.024 ve 512 adet nöron değerlendirilmiştir.
- Son katmanda ise Softmax aktivasyon fonksiyonu kullanılmış ve sınıflandırma işlemi gerçekleştirilmiştir.
- Modelin kaybı kategorik çapraz entropi ile belirlenmiştir.
- En iyileme algoritması olarak Adam fonksiyonu kullanılmıştır.

LSTM ile gerçekleştirilen sınıflandırma işleminin sonuçları Tablo 7.6'da verilmiştir.

Tablo 7.6. LSTM sınıflandırıcısının sonuçları

Sayısallaştırma Yöntemi	Doğruluk	F1-Skor	Kesinlik	Duyarlılık	AUC
EIIP	%51,16	%43,92	%59,92	%34,67	0,70
Hidrofobiklik	%42,53	%63,99	%68,74	%59,87	0,50
CPNR	%73,93	%71,65	%78,84	%65,67	0,86
BLOSUM62	%66,92	%60,40	%70,77	%52,69	0,77
Atchley faktörleri	%64,54	%60,33	%65,72	%55,76	0,68
PAM250	%64,54	%59,79	%68,83	%52,85	0,73
İkili kodlama	%60,05	%53,54	%66,85	%44,66	0,74
Rastgele kodlama	%59,63	%68,63	%69,66	%67,64	0,71
FIBHASH	%69,47	%67,85	%74,55	%62,27	0,83

Tablo 7.6'da verilen sonuçlar incelendiğinde, genel olarak LSTM yönteminin RNN yönteminden etkili olduğu görülmektedir. Her ne kadar etkili olsa da sınıflandırma işlemi gene de istenilen seviyeye gelememiştir. RNN sınıflandırıcısının aksine LSTM modelinde en düşük sınıflandırma işlemi hidrofobiklik yöntemi ile gerçekleştirilmiştir. Bu yöntem ile %42,53 doğruluk elde edilirken, 0,50 AUC skoru hesaplanmıştır. Hidrofobiklik yöntemi dışında sadece EIIP ve rastgele kodlama yöntemleri %60'ın üzerinde doğruluk sergileyememişlerdir. EIIP yönteminin doğruluk skoru %51,16 olmuşken, bu değer rastgele kodlama yönteminde %59,63'e çıkmıştır. Bu iki yöntemin AUC skorları ise sırasıyla 0,70 ve 0,74 olmuştur. Ardından BLOSUM62, Atchley faktörleri, PAM250, ikili kodlama ve önerilen FIBHASH yöntemi %60'ın üzerinde doğruluk göstermiştir.

BLOSUM62 ile hesaplanan doğruluk skoru %66,92'yken, bu oran Atchley faktörleri ve PAM250 yöntemlerinde %64,54'e gerilemiştir. İkili kodlama yönteminin doğruluk skoru %60,05 olarak hesaplanmışken önerilen FIBHASH yöntemi ile %69,47 doğruluk elde edilmiştir. Bu yöntemler içerisinde en yüksek AUC skoru ise önerilen FIBHASH yöntemiyle elde edilmiş ve 0,83 AUC skoru hesaplanmıştır. LSTM derin öğrenme modeli ile en yüksek sınıflandırma işlemi ise CPNR protein sayısallaştırma yöntemi ile gerçekleştirilmiştir. Sadece CPNR protein sayısallaştırma yöntemi %70'in üzerinde doğruluk sergilemiştir. Benzer şekilde en yüksek AUC skoru da CPNR yöntemi ile hesaplanmış ve 0,86 olmuştur.

Üçüncü senaryoda ise BiRNN derin öğrenme modeli kullanılmıştır. Birinci ve ikinci senaryoda da olduğu gibi veri seti %80-%20 şeklinde ayrıştırılmıştır. Bu modelin de parametreleri deneme-yanılma yaklaşımı kullanılarak belirlenmiş ve eğitim aşaması 200 iterasyon ile gerçekleştirilmiştir. Tasarlanan BiRNN modelinin parametreleri şu şekilde özetlenebilir;

- Girdi katmanında 97.576 adet protein dizilimi kullanılmıştır.
- İkinci katmanda 256 birim BiRNN kullanılmış ve bu katmanda aktivasyon işlemi ReLU fonksiyonu ile gerçekleştirilmiştir.
- Üçüncü katmanda da 256 birim BiRNN kullanılmış ve aktivasyon fonksiyonu olarak ReLU uygulanmıştır.
- Ardında düzleştirme işlemi yapılmış ve veriler 1-boyutlu hale getirilmiştir.
- Düzleştirme işleminin ardından sırasıyla yığın normalleştirme ve seyreltme adımları gerçekleştirilmiştir. Seyreltme işleminde verilerin %25'i unutulmuştur.
- İki adet tam-bağımlı katman kullanılmış ve sırasıyla bu katmanlarda 1.024 ve 512 adet nöron değerlendirilmiştir.
- Sınıflandırma katmanında ise Softmax aktivasyon fonksiyonu kullanılmıştır.
- Modelin kaybı kategorik çapraz entropi ile belirlenmiştir. En iyileme işlemi için ise Adam fonksiyonu kullanılmıştır.

BiRNN ile gerçekleştirilen sınıflandırma işleminin sonuçları Tablo 7.7'de verilmiştir.

Tablo 7.7. BiRNN sınıflandırıcısının sonuçları

Sayısallaştırma Yöntemi	Doğruluk	F1-Skor	Kesinlik	Duyarlılık	AUC
EIIP	%73,88	%72,98	%79,02	%67,81	0,74
Hidrofobiklik	%79,54	%78,53	%82,82	%74,67	0,88
CPNR	%79,05	%77,53	%83,39	%72,45	0,83
BLOSUM62	%71,97	%70,65	%78,92	%63,95	0,81
Atchley faktörleri	%79,59	%80,31	%82,29	%78,44	0,85
PAM250	%74,03	%72,74	%77,45	%68,58	0,79
İkili kodlama	%64,19	%63,77	%69,45	%58,95	0,73
Rastgele kodlama	%69,23	%70,35	%78,95	%63,45	0,72
FIBHASH	%78,47	%79,17	%81,93	%76,60	0,90

Tablo 7.7’de verilen BiRNN sonuçlarına göre, ikili kodlama ve rastgele kodlama dışında bütün protein sayısallaştırma yöntemleri %70’in üzerinde doğruluk skoru üretmiştir. İkili kodlama yöntemiyle %64,19 doğruluk elde edilirken, rastgele kodlama ile %69,23 oranında doğruluk hesaplanmıştır. AUC skorları ise birbirine yakın değerlerden meydana gelmiştir. Rastgele kodlama yönteminin AUC skoru 0,72’yken, ikili kodlamanın AUC skoru 0,73 olarak elde edilmiştir. BLOSUM62, EIIP ve PAM250 yöntemleriyle birbirine yakın doğruluk skorları elde edilmiştir. EIIP protein sayısallaştırma yöntemi ile %73,88, BLOSUM62 protein sayısallaştırma yöntemi ile %71,97 ve PAM250 protein sayısallaştırma yöntemi ile %74,03 doğruluk skorları hesaplanmıştır. Bunun yanı sıra, en yüksek skorlar FIBHASH, CPNR ve hidrofobiklik protein sayısallaştırma yöntemleri ile elde edilmiştir. Önerilen FIBHASH yöntemi ile %78,47 doğruluk ve 0,90 AUC skoru elde edilmiştir. CPNR protein sayısallaştırma yöntemi ile %79,05 doğruluk skoru hesaplanırken, hidrofobiklik protein sayısallaştırma yöntemiyle bu oran %79,54 olmuştur. AUC skorlarında ise hidrofobiklik değeri 0,88’e ulaşmışken, CPNR yönteminde AUC skoru 0,83 olmuştur.

Dördüncü ve son senaryoda ise BiLSTM derin öğrenme modeli kullanılmıştır. Diğer bütün senaryolarda olduğu gibi bu senaryoda da veri seti eğitim ve test veri seti olmak üzere ikiye ayrılmıştır. Eğitim işleminde orijinal veri setinin %80’i kullanılırken, test veri setinde %20’si kullanılmıştır. Eğitim aşamasında model 200 kere eğitilmiş ve bütün parametreleri deneme-yanılma yaklaşımına göre belirlenmiştir. Tasarlanan BiLSTM derin öğrenme modelinin parametreleri aşağıda verilmiştir;

- Girdi katmanında 97.576 adet protein dizilimi kullanılmıştır.
- İkinci katmanda 256 birim BiLSTM kullanılmış ve bu katmanda aktivasyon işlemi ReLU fonksiyonu ile gerçekleştirilmiştir.
- Üçüncü katmanda da 256 birim BiLSTM kullanılmış ve aktivasyon fonksiyonu olarak ReLU uygulanmıştır.
- Ardında düzleştirme işlemi yapılmış ve veriler 1-boyutlu hale getirilmiştir.

- Düzleştirme işleminin ardından sırasıyla yığın normalleştirme ve seyreltme adımları gerçekleştirilmiştir. Seyreltme işleminde verilerin %25'i unutulmuştur.
- İki adet tam-bağılantılı katman kullanılmış ve sırasıyla bu katmanlarda 1.024 ve 512 adet nöron değerlendirilmiştir.
- Sınıflandırma katmanında ise Softmax aktivasyon fonksiyonu kullanılmıştır.
- Modelin kaybı kategorik çapraz entropi ile belirlenmiştir. En iyileme işlemi için ise Adam fonksiyonu kullanılmıştır.

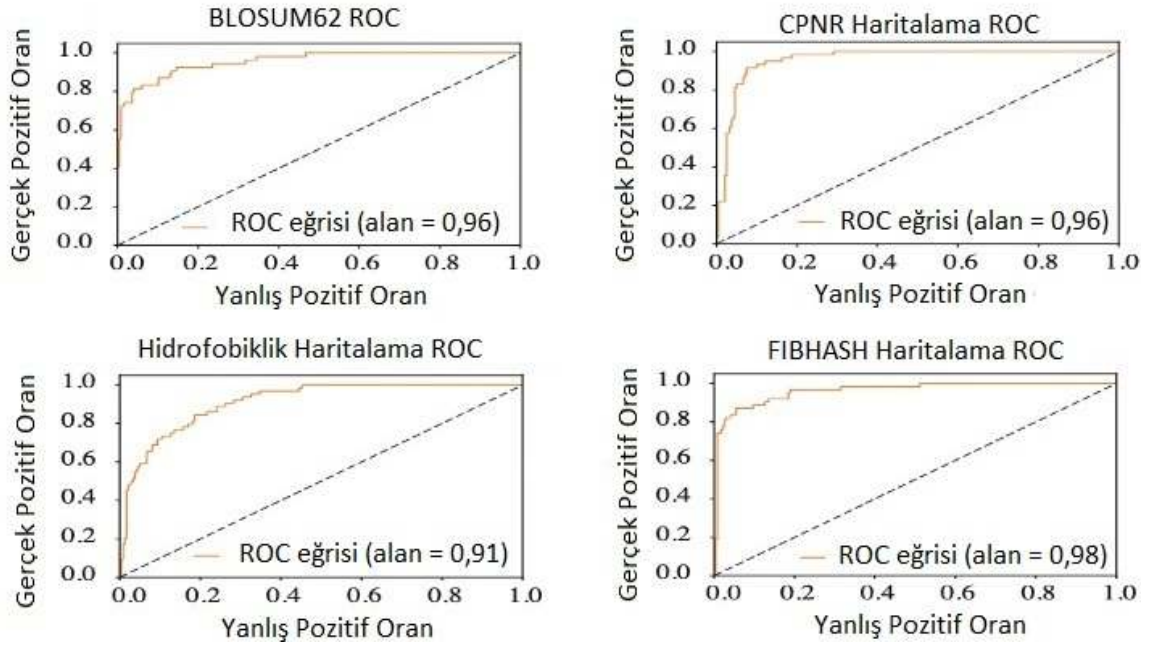
BiLSTM ile gerçekleştirilen sınıflandırma işleminin sonuçları Tablo 7.8'de verilmiştir.

Tablo 7.8. BiLSTM sınıflandırıcısının sonuçları

Sayısallaştırma Yöntemi	Doğruluk	F1-Skor	Kesinlik	Duyarlılık	AUC
EIIP	%78,76	%75,55	%87,51	%66,48	0,86
Hidrofobiklik	%82,32	%74,84	%96,22	%61,24	0,91
CPNR	%88,50	%85,06	%94,23	%77,53	0,96
BLOSUM62	%85,63	%82,33	%92,62	%74,10	0,96
Atchley faktörleri	%78,23	%73,22	%88,79	%62,31	0,84
PAM250	%84,46	%83,31	%96,76	%73,15	0,90
İkili kodlama	%82,44	%77,82	%90,41	%68,32	0,85
Rastgele kodlama	%85,42	%79,89	%89,25	%72,32	0,90
FIBHASH	%92,77	%88,31	%97,88	%80,46	0,98

Tablo 7.8'de verilen BiLSTM sonuçları incelendiğinde genel olarak en etkili senaryonun dördüncü senaryo olduğu gözlemlenmiştir. En düşük doğruluk skoru EIIP ve Atchley faktörleri ile elde edilmiş ve sonuçlar sırasıyla %78,76 ve %78,23 olmuştur. Bu yöntemlerin AUC skorları ise 0,86 ve 0,84 olarak elde edilmiştir. Diğer yöntemlere bakıldığında ise, önerilen FIBHASH yöntemi dışında hiçbir yöntemin %90'ın üzerinde doğruluk sergileyemediği görülmüştür. Hidrofobiklik ve ikili kodlama yöntemleri birbirine yakın doğruluk skorları üretmiş ve bu değerler sırasıyla %82,32 ve %82,44 olmuştur. AUC skorlarına göre incelendiğinde ise hidrofobiklik yönteminin ikili kodlamadan daha iyi sonuca ulaştığı belirlenmiştir. Hidrofobiklik yöntemi ile 0,91 AUC skoru elde edilirken, ikili kodlama ile 0,85 AUC skoru hesaplanmıştır. Geri kalan protein sayısallaştırma yöntemleri içerisinde ise, BLOSUM62, PAM250 ve rastgele kodlama yöntemleri benzer doğruluk sergilemişlerdir. BLOSUM62 yönteminin doğruluk skoru %85,63 olarak elde edilirken, bu oran PAM250 protein sayısallaştırma yöntemi ile %84,46'ya düşmüştür. Rastgele kodlama yönteminde ise %85,42 doğruluk hesaplanmıştır. AUC skorlarına göre kıyaslama yapıldığında ise bahsi geçen protein sayısallaştırma yöntemleri 0,90'ın üzerinde başarımleri sergilemiştir. En başarılı sınıflandırma işlemi ise CPNR ve önerilen FIBHASH yöntemleri ile elde edilmiştir. CPNR protein

sayısallaştırma yöntemiyle %88,50 doğruluk hesaplanırken, FIBHASH yöntemi ile %92,77 doğruluk elde edilmiştir. AUC skorlarında ise birbirine benzer değerler gözlemlenmiştir. CPNR ve FIBHASH ile sırasıyla 0,96 ve 0,98 AUC skorları hesaplanmıştır. Şekil 7.2’de en etkili dört protein sayısallaştırma yönteminin ROC grafikleri verilmiştir.



Şekil 7.2. En etkili protein sayısallaştırma yöntemlerinin ROC eğrileri

7.3. Uygulama Sonuçlarının Değerlendirilmesi

Sonuçlar genel olarak incelendiğinde, bütün protein sayısallaştırma tekniklerinin etkili olduğu söylenebilir. Ancak, tüm sonuçlar dikkate alındığında, EIIP ve hidrofobiklik yöntemlerinin yeteri kadar başarılı olmadığı gözlemlenmiştir. Bunun nedeni bu yöntemlerin dejenerasyona uğraması olabilir. Farklı amino asit değerleri bu yöntemlerde benzer sayısal değerler almaktadır. Bundan dolayı dejenerasyon görülmekte ve bilgi kaybı olmaktadır. Bunlara ek olarak, rastgele kodlama yönteminin başarımı üretilen sayıya göre değişiklik göstermektedir. Amino asitlere atanan sayıların farklı olması durumunda rastgele kodlama yönteminin başarımı azalabilmekte ya da artabilmektedir. Evrim-tabanlı yöntemlerden olan BLOSUM62 ve PAM250 matrisleri genellikle hizalama çalışmalarında kullanılmaktadır. Ancak bu çalışmada hizalama işlemi yapılmamasına rağmen iki yöntemde başarılı sonuçlar sergilemiştir. Yapı-tabanlı bir yöntem olan Atchley faktörleri ise her ne kadar en iyi sonuca ulaşmasa da başarılı bir sınıflandırma işlemi gerçekleştirmiştir. İkili kodlama yönteminin performansı ise makul olmuştur. Bu durumun nedeni bu yöntemin boyutundan kaynaklanmaktadır. Boyut sayısı arttıkça performans düşük olmaktadır.

İkili kodlama yerine altılı kodlama ya da beşli kodlama gibi yöntemlerin kullanılması performansı arttırabilir. En iyi sınıflandırma işlemleri ise CPNR ve önerilen FIBHASH yöntemleri ile gerçekleştirilmiştir. İki yöntemde belirli bir protein bilgisine ihtiyaç duymamaktadır. Bundan dolayı bilgi kaybı olmamakta ve dizilimler daha etkili analiz edilebilmektedir. Önerilen yöntemin CPNR yönteminden en büyük farkı, yöntemin Fibonacci sayısı gibi doğada gözlemlenebilen bir yapıya dayandırılmasıdır. CPNR yönteminde böyle bir yapı bulunmamakta, sayısallaştırma işlemi sadece proteinlerin asal sayılarla ilişkisine göre gerçekleştirilmektedir. Diğer yöntemler kadar olmasa da bu yöntemde de bundan dolayı bilgi kaybı gözlemlenebilmektedir.

Sonuç olarak, çalışmada yeni bir protein sayısallaştırma yöntemi önerilmiş ve protein aileleri yinelemeli sinir ağları ile sınıflandırmıştır. Bu doğrultuda RNN, LSTM, BiRNN ve BiLSTM yöntemleri kullanılmıştır. Önerilen yöntem geliştirilirken Fibonacci sayıları ve hash tablosu kullanılmıştır. Çalışmada protein dizilimleri, FIBHASH, Atchley faktörleri, ikili-kodlama, CPNR, EIIP, BLOSUM62, PAM250, hidrofobiklik ve rastgele kodlama yöntemleri ile sayısallaştırılmıştır. Sayısallaştırma adımının ardından girdiler yinelemeli sinir ağlarında kullanılmıştır. Protein sayısallaştırma yöntemlerinin başarımları doğruluk, AUC skor, F1-skor, kesinlik ve duyarlılık değerlendirme ölçütleri ile belirlenmiştir. RNN ile gerçekleştirilen sınıflandırma işleminde FIBHASH yöntemi ile %57 doğruluk ve 0,67 AUC skoru elde edilmiştir. AUC skoru 0,70'ten küçük olduğu için başarılı bir sınıflandırma gerçekleşmemiştir. İkinci senaryoda ise LSTM derin öğrenme modeli kullanılmıştır. Bu sınıflandırıcı ile performansta artış gözlemlenmiştir. Önerilen yöntem ile doğruluk skoru %69,47'ye yükselirken, AUC skoru ise 0,83'e yükselmiştir. AUC skoru 0,70'ten büyük olduğu için bu sonuçlar makul olarak değerlendirilmiştir. Önerilen yöntem her ne kadar en başarılı sınıflandırmayı yapmasa da etkili bir sınıflandırma işlemi gerçekleştirmiştir. RNN ve LSTM modellerinin dezavantajlarından dolayı çalışmada ayrıca BiRNN ve BiLSTM derin öğrenme modelleri de kullanılmıştır. Üçüncü senaryoda BiRNN kullanılmış ve en yüksek AUC skoru FIBHASH yöntemiyle elde edilmiş ve 0,90 AUC skoru gözlemlenmiştir. Bu senaryoda bütün protein sayısallaştırma yöntemleri 0,70'ten yüksek sonuçlar ürettiği için BiRNN derin öğrenme modeli başarılı olmuştur. Bahsi geçen sınıflandırıcılar içinde ise en yüksek sınıflandırma işlemi BiLSTM derin öğrenme modeli ile gerçekleştirilmiştir. Bu senaryoda, protein sayısallaştırma yöntemlerinin tamamı neredeyse mükemmel AUC skorlarına ulaşmıştır. Ancak en iyi performans %92,77 doğruluk ve 0,98 AUC skoru ile önerilen FIBHASH yönteminden elde edilmiştir. Çalışmanın sonunda önerilen yöntemin etkili olduğu gözlemlenmiştir.

8. ÖNERİLEN PROTEİN SAYISALLAŞTIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Tezin bu bölümünde, geliştirilen protein sayısallaştırma yöntemlerini karşılaştıran bir çalışma gerçekleştirilmiştir. Gerçekleştirilen çalışmada COVID-19'a neden olan SARS-CoV-2 virüsü ve konak hücre arasındaki etkileşimler tahmin edilmiş ve bu doğrultuda hem önerilen hem de literatürde sıklıkla kullanılan protein sayısallaştırma yöntemleri kullanılmıştır. Tahmin işlemi için ise derin öğrenme algoritmaları kullanılmış ve protein sayısallaştırma yöntemlerinin başarımları karşılaştırılmıştır. Ayrıca, sınıflandırma adımı protein sayısallaştırma yöntemlerinin başarımları doğruluk, F1-skor ve AUC skor metrikleri ile belirlenirken, bu yöntemlerin donanımsal başarımları ise algoritmik süreçleri değerlendirmede kullanılan bellek boyutu, analiz süresi ve zaman karmaşıklığı metrikleri açısından değerlendirilmiştir.

8.1. Protein Sayısallaştırma Yöntemlerinin Viral-Konak Etkileşimlerine Uygulanması

COVID-19 hastalığı ile mücadele edebilmek için SARS-CoV-2 hastalığının genetik karakteristiklerinin iyi bir şekilde bilinmesi gerekmektedir. Virüsün genetik karakteristikleri sayesinde COVID-19 hastalığı ile ilgili çeşitli ilaç ve aşı tedavileri ortaya konabilmektedir [222, 223]. SARS-CoV-2 virüsü spike glikoprotein (S), membrane glikoprotein (M), small envelope glikoprotein (E) ve nükleokapsit protein (N) olmak üzere dört ana yapısal proteinlerden oluşmaktadır. Bu virüs proteinlerinin hayatta kalma ve üreme için konakçı hücreler ile nasıl etkileşime girdiğinin anlaşılması, ilaçların geliştirilebilmesi ve kullanılabilmesi için büyük bir önem arz etmektedir [224]. Virüsler, konak hücreler ile protein-protein etkileşimleri yoluyla iletişim kurmaktadır. Bundan dolayı, virüs ve konak proteinler arasındaki protein etkileşimlerinin belirlenmesi, virüs proteinlerinin nasıl çalıştıkları, nasıl çoğaldıkları ve nasıl hastalığa neden oldukları hakkında bilgi vermektedir. Proteinler ve virüs-konak arasındaki etkileşimleri belirleyebilmek için deneysel yaklaşımlar sıklıkla kullanılmaktadır. Deneysel yaklaşımlar çok sayıda veri üretmekte ve hem zaman hem de laboratuvar ekipmanları açısından maliyetli olmaktadır [225]. Bunlara ek olarak protein etkileşimleri kullanılan deneysel yöntemlere göre değişiklik göstermektedir. Deneysel yöntemler hem çevresel hem de operasyonel işlemlere dayalı oldukları için, yanlış pozitif ve yanlış negatif sonuçlar ortaya çıkabilmektedir [226]. Bu problemlerden dolayı, hesaplama-tabanlı yaklaşımlar daha çok tercih edilmekte ve bu yöntemlerin popülaritesi sürekli artış göstermektedir [227, 228].

Gerçekleştirilen bu çalışmada da hesaplama tabanlı yaklaşımlar kullanılmış ve SARS-CoV-2 ve insana ait proteinler arasındaki viral-konak etkileşimleri tahmin edilmiştir. Çalışma dört aşamadan meydana gelmektedir. Birinci aşamada SARS-CoV-2 ve insana ait viral-konak etkileşim

verileri elde edilmiştir. İkinci aşamada protein dizilimleri sayısallaştırma yöntemleri ile sayısal hale getirilmiştir. Bu amaç doğrultusunda 14 adet protein sayısallaştırma yöntemi kullanılmıştır. Bu yöntemler; AESNN1, Atchley faktörleri, AVL, ikili kodlama, BLOSUM62, CPNR, EIIP, FIBHASH, hidrofobiklik, entropi, Meiler parametreleri, Micheletti potansiyelleri, Miyazawa enerjileri ve PAM250 matrisidir. Üçüncü aşamada ise derin öğrenme modeli tasarlanmış ve SARS-CoV-2 virüsü ve insan proteinleri arasındaki viral-konak etkileşimleri sınıflandırılmıştır. Son aşamada ise sayısallaştırma yöntemlerinin performansları doğruluk, AUC skoru ve F1-skor değerlendirme ölçütleriyle belirlenmiştir. Çalışmanın öne çıkan noktaları şu şekilde özetlenebilir;

- Bu çalışma ile SARS-CoV-2 ve insana ait proteinlerin arasındaki viral-konak etkileşimler hesaplama tabanlı yaklaşımlar ile tahmin edilmiştir.
- Bildiğimiz kadarıyla ilk kez bu çalışmada AVL ağacı, Shannon entropisine dayanan yöntem ve FIBHASH yöntemleri kullanılarak proteinler analiz edilmiştir.
- Bu çalışma ile SARS-CoV-2 ve insana ait proteinler arasındaki viral-konak etkileşimler derin öğrenme ile etkili bir şekilde sınıflandırılmıştır.

Literatür incelendiğinde hesaplama-tabanlı yaklaşımların viral-konak etkileşimlerini tahmin etmek için sıklıkla kullanıldığı gözlemlenmiştir. [224] numaralı çalışmada araştırmacılar SARS-CoV-2 ve insan proteinleri arasındaki viral-konak etkileşimlerini makine öğrenmesi teknikleri ile tahmin etmişlerdir. Çalışmada sınıflandırma işlemi için DVM, KNN, NB ve RO algoritmaları kullanılmıştır. Protein dizilimleri AAB, BT ve SAAB (Sözde AAB) yöntemleri ile sayısallaştırılmış ve özellikler elde edilmiştir. Ardından ÖVK (Öğrenmeli Vektör Kuantalama) yöntemiyle özellik seçilimi gerçekleştirilmiştir. Sınıflandırma algoritmalarının başarımları doğruluk, kesinlik, duyarlılık ve F1-skor ile ölçülmüştür. En iyi sonuçlar DVM ve RO modelleri ile elde edilmiştir. Bu modeller ile ortalama %72,33 oranında doğruluk elde edilmiştir. [229] numaralı çalışmada araştırmacılar SARS-CoV-2 enfeksiyonuna ait konak-hücre interaktomunu analiz etmişlerdir. Çalışmada analiz işlemi için ağ tabanlı bir yaklaşım kullanılmıştır. Çalışmada insana ait koronovirüs molekülleri (SARS-CoV, MERS-COV ve HCoV-229E) kullanılarak SARS-CoV-2 virüsünün patojeni tahmin edilmiş ve protein-protein etkileşimlerinden yararlanılmıştır. Çalışmanın sonunda geliştirilen yöntemin başarılı olduğu ve potansiyel yeni biyolojik hedeflerin seçilmesinde kullanılabileceği belirtilmiştir. [230] numaralı çalışmada araştırmacılar SARS-CoV-1, SARS-CoV-2 virüslerine ve insan proteinlerine ait viral-konak etkileşimlerini tamsayı doğrusal programlama kullanarak hizalamışlardır. Çalışmada gen ontolojilerinden yararlanılmış ve etkileşimler MF, CC ve BP seviyeleri kullanılarak incelenmiştir. Çalışmanın ilk aşamasında SARS-CoV-1 ve SARS-CoV-2 virüslerine ait proteinler hizalanmıştır. Hizalama işlemi her bir gen ontolojisi için ayrı ayrı gerçekleştirilmiştir. Hizalama işleminin ardından insana ait proteinler kullanılmış ve bu virüslerin etkileşim halinde olduğu proteinler belirlenmiştir. Spike protein türleri için en iyi BP ve CC sonucu Q7Z5G4 proteini ile elde edilmiştir. MF için ise Q7Z5G4 ve Q9C0B5

proteinlerinden elde edilmiştir. Envelope protein türleri için ise en iyi hizalama sonucu BB kategorisinde O00203 protein ile elde edilmiştir. CC kategorisinde Q8IWA5 proteini en iyi hizalama sonucu verirken, MF kategorisinde ise en iyi hizalama etkileşimi O00203 proteininde hesaplanmıştır. Membran proteinin de ise en iyi hizalama skoru CC kategorisinde Q96D53 proteininden elde edilmiştir. BP kategorisinde ise en iyi skor P27105 proteininden elde edilirken, MF kategorisinde ise tekrardan Q96D53 proteininden elde edilmiştir. Nükleokapsit protein türünde ise en iyi hizalama skoru BP için P67870 proteininden elde edilmiştir. CC gen ontoloji türünde etkili hizalama işlemi P11940 protein ile hesaplanmıştır. MF kategorisinden ise birden çok proteinin hizalama işleminde etkili olduğu gözlemlenmiştir. Çalışmanın sonunda hizalama işleminin SARS-CoV-2 ve insana ait proteinlerin virüs-konak etkileşimlerinin belirlenmesinde etkili olduğu belirtilmiştir. [231] numaralı çalışmada araştırmacılar yeni bir derin öğrenme modeli geliştirerek protein dizilimlerinden ve bulaşıcı hastalık fenotiplerinden virüs-konak etkileşimlerini tahmin etmişlerdir. Çalışmada insan proteinleri ve virüs taksonomileri ikili-kodlama yöntemi ile sayısallaştırılmıştır. Sayısallaştırma işleminin ardından ESA tabanlı derin öğrenme modeli geliştirilmiş ve veriler sınıflandırılmıştır. Sınıflandırma işlemi aile-tabanlı, takson-tabanlı ve protein-tabanlı olmak üzere üç farklı kategoride gerçekleştirilmiş ve önerilen derin öğrenme modelinin performansı ROC eğrisi ile hesaplanmıştır. Aile-tabanlı sınıflandırmada %81,30 oranında doğruluk başarımı elde edilirken bu oran takson-tabanlı sınıflandırmada %82,90 olarak hesaplanmıştır. Protein-tabanlı sınıflandırmada ise bu oran %80 olmuştur. Çalışmanın sonunda önerilen derin öğrenme modelinin SARS-CoV-2 ve insana ait proteinler arasındaki viral-konak etkileşimlerini etkili bir şekilde belirlediği gözlemlenmiştir. [232] numaralı çalışmada ise araştırmacılar SARS-CoV-2 virüsü ve insana ait proteinler arasındaki viral-konak etkileşimlerini ağ-tabanlı yöntem ile tahmin etmişlerdir. Virüs ve insana ait proteinlerin yapısal ve içerik bilgileri kullanılmış ve bu bilgiler birleştirilmiştir. Ardından YSA ile sınıflandırma işlemi gerçekleştirilmiştir. Geliştirilen model literatürde bulunan çeşitli modellerle de karşılaştırılmış ve sonuçlar irdelenmiştir. Sınıflandırma işleminin başarımı doğruluk, duyarlılık, kesinlik, F1-skor ve AUC değerleri ile hesaplanmıştır. Önerilen derin öğrenme modeli ile %97,1 oranında doğruluk skoru elde edilirken, 0,99 oranında AUC skoru hesaplanmıştır.

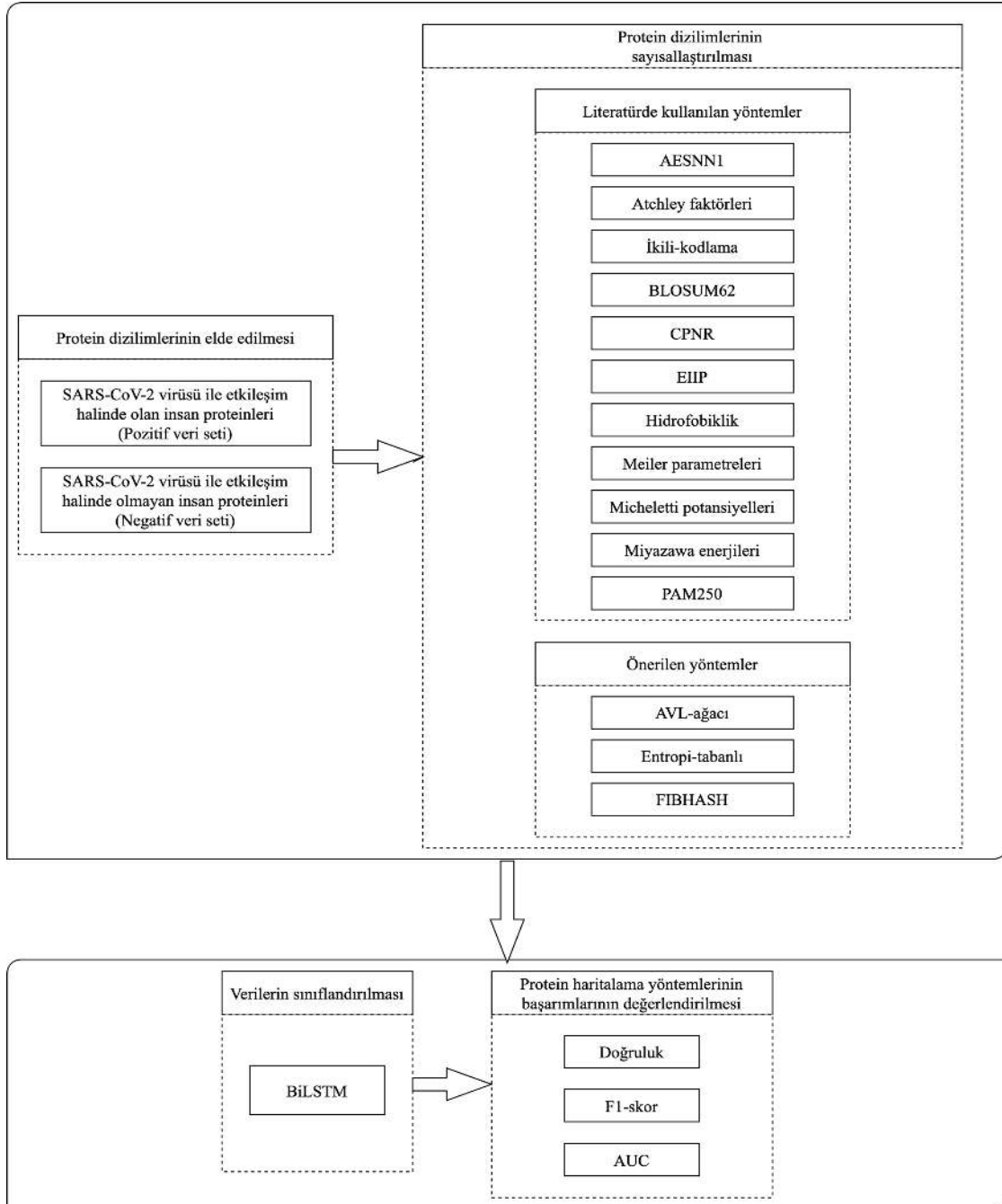
Gerçekleştirilen çalışmada, [164] numaralı kaynakta verilen veri seti kullanılmıştır. Çalışmada araştırmacılar SARS-CoV-2 virüsü ve insan proteinleri arasındaki viral-konak etkileşimlerini içeren veri seti paylaşmışlardır. Bu veri seti oluşturulurken afinite saftlaştırma kütle spektrometrisi kullanılmıştır. Veri setinde 332 adet insana ait protein ile 4 adet yapısal ve 20 adet yardımcı koronavirüs proteini arasındaki etkileşimler bulunmaktadır. Bu veri seti çalışmada pozitif veri setini oluşturmak için kullanılmıştır. Protein-protein ve viral-konak etkileşimlerinde negatif veri setini oluşturmak için belirli bir altın standart bulunmamaktadır. Negatif veri seti oluşturulurken genellikle rastgele eşleştirme ve hücre altı lokalizasyonu olmak üzere iki yöntem kullanılmaktadır

[224]. Çoğu çalışmada etkileşim halinde olmayan proteinler rastgele oluşturulur ve pozitif veri setinde kullanılan protein çiftleri elenir [225, 233]. Bunun yanı sıra bazı çalışmalarda ise negatif veri seti hücre altı lokalizasyonu ile oluşturulmaktadır [76]. Ancak, bu iki yöntemle elde edilen negatif veri setleri çok güvenilir olmamaktadır. Rastgele eşleştirme işleminde, pozitif etkileşimler yanlışlıkla negatif etkileşim olarak değerlendirilebilir ve farklı rastgele eşleştirme işlemi için farklı doğruluk skorları hesaplanabilir. Bunun yanı sıra, hücre altı lokalizasyonu ile de önyargılı doğruluk skorları elde edilebilmektedir [233]. Bu gibi nedenlerden dolayı bu çalışmada negatif veri seti oluştururken bu iki yöntemde kullanılmamıştır. Negatif veri seti [224] numaralı çalışmadan elde edilmiştir. Çalışmada araştırmacılar negatif veri setini oluştururken derece-tabanlı bir yaklaşım kullanmışlardır. Bu yaklaşım ile oluşturulan negatif veri setinin hem rastgele eşleştirme hem de hücre altı lokalizasyon yöntemlerine göre gerek eğitim aşamasında gerek test aşamasında daha etkili olduğu belirlenmiştir. Pozitif veri setinde toplamda 1.327 adet protein dizilimi kullanılırken, negatif veri seti için ise 4.554 adet protein dizilimi kullanılmıştır. Kullanılacak veri seti hazırlandıktan sonra protein dizilimleri 14 adet protein sayısallaştırma yöntemi ile sayısal hale getirilmiştir. Sayısallaştırma işleminin ardından etkileşimler sınıflandırılmış ve viral-konak hücreler arasındaki protein-protein etkileşimleri tahmin edilmiştir. Sınıflandırma adımı BiLSTM derin öğrenme modeli kullanılmış ve protein sayısallaştırma yöntemlerinin başarımları doğruluk, F1-skor ve AUC değerlendirme ölçütleri ile belirlenmiş ve kıyaslanmıştır. Bu çalışmada tasarlanmış olan BiLSTM ağının parametreleri şu şekilde özetlenebilir;

- Girdi katmanında her bir protein sayısallaştırma tekniği ile sayısallaştırılmış olan protein dizilimleri kullanılmıştır.
- İkinci katmanda BiLSTM kullanılmış ve bu amaç için 64 adet LSTM birimi değerlendirilmiştir. Aktivasyon fonksiyonu olarak ReLU kullanılmıştır.
- Üçüncü katmanda tekrar BiLSTM kullanılmış ve bu katmanda 32 adet LSTM birimi değerlendirilmiştir. Aktivasyon fonksiyonu olarak ReLU kullanılmıştır.
- Ardından çıktılar düzleştirme işlemine tabi tutulmuş ve veriler 1-boyutlu hale getirilmiştir.
- Daha sonra tüm verilerin aynı aralıkta olmasını sağlayabilmek için normalleştirme işlemi yapılmış ve yığın normalleştirilmesi kullanılmıştır.
- Son olarak üç adet tam-bağılantılı katman kullanılmış ve bu kısım için sırasıyla 240, 120 ve 60 adet nöron ele alınmıştır.
- Son katmanda ise sınıflandırma yapılmış ve etkileşimin olup olmadığını belirlemek için bu katmanda 2 adet nöron kullanılmıştır. Aktivasyon fonksiyonu olarak ise Softmax hesaplamasından yararlanılmıştır.
- Modelin kaybını belirleyebilmek için kategorik çapraz-entropi kullanılmıştır.
- En iyileme işlemi için SGD kullanılmıştır.
- Eğitim işlemi 50 iterasyon ile yapılmıştır.

- Geliştirilen modelin ve protein sayısallaştırma yöntemlerinin performansları 10 katlı çapraz-doğrulama ile belirlenmiştir.
- Çalışmada kullanılan bütün parametreler deneme-yanılma yaklaşımı ile belirlenmiş ve en iyi sonuçları veren parametreler değerlendirilmiştir.

Şekil 8.1’de çalışmanın akış şeması verilmiştir.



Şekil 8.1. Protein sayısallaştırma yöntemlerinin başarımlarının değerlendirme süreci

8.2. Derin Öğrenme Kullanılarak Viral-Konak Etkileşimlerin Tahmin Edilmesi

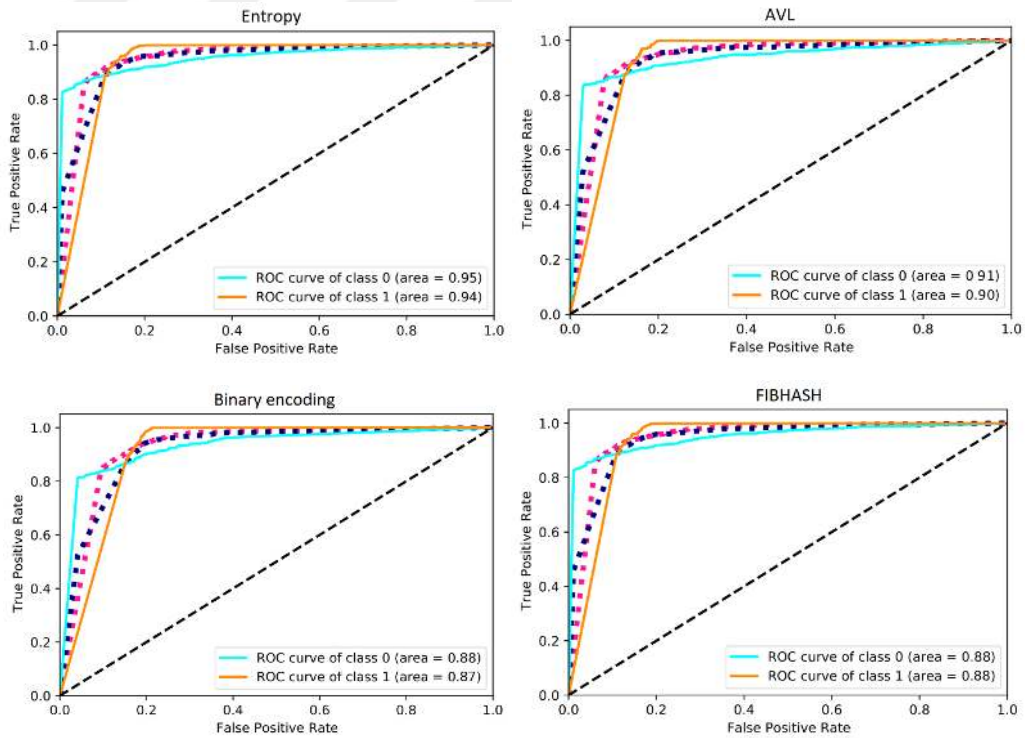
Çalışmada sınıflandırma işleminin ardından protein sayısallaştırma yöntemlerinin başarımları değerlendirilmiş ve karşılaştırılmıştır. Protein sayısallaştırma yöntemlerinin performanslarını kıyaslayabilmek için doğruluk, F1-skor ve AUC değerleri hesaplanmıştır. Tablo 8.1’de sınıflandırma sonuçları verilmiştir.

Tablo 8.1. Protein sayısallaştırma yöntemlerinin başarımlarının karşılaştırılması

Sayısallaştırma Yöntemi	Doğruluk	F1-Skor	AUC
Entropi-tabanlı	%94,74	%93,41	0,95
AVL-ağacı	%90,44	%90,39	0,91
FIBHASH	%89,33	%89,09	0,88
AESNN1	%88,94	%89,41	0,81
Atchley faktörleri	%82,58	%81,52	0,83
İkili-kodlama	%86,20	%85,54	0,88
BLOSUM62	%85,54	%84,71	0,83
CPNR	%88,52	%88,66	0,87
EIIP	%78,89	%79,00	0,82
Hidrofobiklik	%79,65	%76,90	0,77
Meiler parametreleri	%81,10	%83,10	0,82
Micheletti potansiyelleri	%75,39	%77,38	0,79
Miyazawa enerjileri	%77,31	%75,32	0,81
PAM250	%87,63	%87,50	0,87

Tablo 8.1’deki sonuçlar incelendiğinde en etkili doğruluk skorunun entropi tabanlı yöntem ile elde edildiği görülmektedir. Bu yöntem ile %94,74 oranında bir doğruluk skoru hesaplanmıştır. Ardından en etkili yöntemin AVL yöntemi olduğu görülmektedir. Bu yöntem ile de %90,44 oranında bir doğruluk skoru elde edilmiştir. Sadece bu iki yöntem %90 üzerinde bir başarımlar sergilemiştir. Bu yöntemlerin dışında en başarılı yöntemlerin FIBHASH, AESNN1, CPNR, PAM250, ikili kodlama, LUCPROT ve BLOSUM62 olduğu görülmektedir. Bu yöntemler ile sırasıyla %89,33, %88,94, %88,52, %87,63, %86,20, %86,14 ve %85,54 oranlarında doğruluk skoru elde edilmiştir. Bu protein sayısallaştırma yöntemlerine en yakın performansı Atchley faktörleri ve Meiler parametreleri göstermiş ve sırasıyla %82,58 ve %81,10 oranında doğruluk skorları hesaplanmıştır. Bahsi geçen tüm bu yöntemlerin dışındaki diğer protein sayısallaştırma yöntemleri %80’in altında bir performans sergilemiştir. Hidrofobiklik ile %79,65 oranında doğruluk değeri elde edilirken, EIIP ile %78,89 oranında doğruluk skoru hesaplanmıştır. En etkisiz yöntemler ise Micheletti potansiyelleri ve Miyazawa enerjileri olmuştur. Bu yöntemler ile sırasıyla %75,39 ve %77,31 oranında doğruluk skorları hesaplanmıştır. AUC skorlarına göre kıyaslama

gerçekleştirildiğinde ise en yüksek sonucun gene Shannon entropisine dayanan protein sayısalılaştırma yöntemi ile elde edildiği gözlemlenmiştir. Bu yöntem ile hesaplanan AUC skoru 0,95 olmuştur. Ardından hesaplanan en yüksek AUC skoru ise tekrar önerilen yöntemlerden birisi olan AVL-ağacına dayalı protein sayısalılaştırma yöntemi ile hesaplanmıştır. Bu yöntem ile elde edilen AUC skoru ise 0,91'dir. FIBHASH ve ikili-kodlama yöntemleri ise benzer AUC skoru hesaplamışlardır. Bu iki yöntem ile elde edilen AUC skoru 0,88 olarak belirtilmiştir. Bunun yanı sıra CPNR ve PAM250 protein sayısalılaştırma yöntemleri ise en az önerilen yöntemler kadar etkili AUC skoru sergilemiştir. Bu yöntemler ile hesaplanan AUC skorları 0,87 olmuştur. Bu AUC skorları dışında en yüksek AUC skoru ise 0,83 ile hem Atchley faktörleri hem de BLOSUM62 yöntemleri ile elde edilmiştir. Meiler parametreleri ve EIIP protein sayısalılaştırma yöntemleri ile 0,82 oranında AUC skorları gözlemlenirken, AESNN1 ve Miyazawa enerjileri ile 0,81 oranında AUC skorları elde edilmiştir. Sadece Micheletti potansiyelleri ve hidrofobiklik protein sayısalılaştırma yöntemlerinin AUC skorları 0,80'in altında kalmıştır. Micheletti potansiyelleri ile 0,79 oranında AUC skoru elde edilirken, bu skor hidrofobiklik protein sayısalılaştırma yöntemi ile 0,77 olmuştur. Şekil 8.2'de en etkili sınıflandırma işlemi yapan protein sayısalılaştırma yöntemlerinin dört tanesine ait AUC grafikleri verilmiştir.



Şekil 8.2. En etkili protein sayısalılaştırma yöntemlerinin AUC grafikleri

Bunun yanı sıra, Tablo 8.2’de protein sayısallaştırma yöntemlerinin bir protein dizilimi için göstermiş oldukları donanımsal performansları verilmiş ve kıyaslanmıştır.

Tablo 8.2. Protein sayısallaştırma yöntemlerinin donanım performansları

Sayısallaştırma Yöntemi	Zaman Karmaşıklığı	Bellek Boyutu	Analiz Süresi
Entropi-tabanlı	-	234.87 MB	16.567 s
AVL-ağacı	$O(\log N)$	140.34 MB	1.223 s
FIBHASH	-	176.65 MB	3.678 s
AESNN1	-	173.45 MB	3.657 s
Atchley faktörleri	-	204.66 MB	13.458 s
İkili-kodlama	-	199.56 MB	10.432 s
BLOSUM62	-	178.23 MB	4.125 s
CPNR	-	172.34 MB	3.542 s
EIIP	-	185.64 MB	5.643 s
Hidrofobiklik	-	185.89 MB	5.765 s
Meiler parametreleri	-	232.88 MB	15.342 s
Micheletti potansiyelleri	-	190.56 MB	8.965 s
Miyazawa enerjileri	-	182.54 MB	5.324 s
PAM250	-	180.54 MB	4.532 s

Tablo 8.2’de verilen donanım performansları doğrultusunda en yavaş çalışan yöntemin Shannon entropisine dayanan yöntem olduğu gözlemlenmiştir. Bu yöntem ile verilen bir proteinin analiz süresi 16.567 saniye olarak belirlenmiştir. Bu yöntemin uzun bir analiz sürecine sahip olması sürpriz değildir. Bu durumun temel nedeni sayısallaştırma işlemi yapılırken amino asitlerin tekrarlama frekanslarını belirlemesi ve logaritma hesabı yapmasıdır. En düşük ikinci analiz süresi ise Meiler parametrelerinden elde edilmiş ve yöntemin süresi 15.342 saniye olarak hesaplanmıştır. Meiler parametreleri 7-parametrelidir bir protein sayısallaştırma yöntemidir. Boyutun fazla olması yöntemin yavaş çalışmasına neden olmuştur. Benzer bir durum Atchley faktörleri ve ikili-kodlama yöntemlerinde de gözlemlenmiştir. Atchley faktörlerinin analiz süresi 13.458 saniye olarak hesaplanırken, bu süre ikili kodlamada 10.342 saniye olarak belirlenmiştir. Atchley faktörleri de tıpkı, Meiler parametreleri gibi birden çok parametreye sahiptir. Bundan dolayı analiz süreci uzun sürmüştür. İkili-kodlama yöntemi ise 1 parametreden oluşmasına rağmen 20-boyutlu bir yöntemdir. Boyut sayısının fazla olması yöntemin analiz sürecinin uzun olmasına neden olmuştur. Bu yöntemler dışında bütün protein sayısallaştırma yöntemlerinin analiz süreleri 10 saniyenin altında kalmıştır. Micheletti potansiyelleri 9 saniye süren bir analiz işlemi gerçekleştirirken, Miyazawa enerjileri, EIIP ve hidrofobiklik yöntemlerinde bu süre yaklaşık 6 saniyeye kadar gerilemiştir. Bu durumun temel nedeni bu yöntemlerin boyutunun düşük olmasıdır. Bahsi geçen bu yöntemlerin

boyut ve parametre sayıları 1'dir. Tıpkı Miyazawa enerjileri, EIIP, hidrofobiklik ve Micheletti potansiyelleri yöntemlerinde olduğu gibi, FIBHASH, AESNN1 ve CPRN yöntemleri de 1 parametrelili ve 1-boyutlu yöntemlerdir. Ancak, bu yöntemlerde amino asit kodları, EIIP, hidrofobiklik, Micheletti potansiyelleri ve Miyazawa enerjileri yöntemlerinin aksine, tam sayı değerlerden oluşmaktadır. Bundan dolayı bu yöntemlerin analiz süresi daha hızlı olmuştur. Bu yöntemler yaklaşık 4 saniye çalışma süresine erişmişlerdir. En etkili çalışma süresi ise önerilen AVL ağacına dayalı protein sayısallaştırma yöntemiyle elde edilmiştir. Bu yöntemin ağaç yapısına sahip olması, tam sayılardan oluşması ve her zaman çalışma zamanı olarak $O(\log N)$ değerini garantilemesi, yöntemin hızlı olmasına neden olmuştur. Analiz süresi ile bellek kullanımı arasında doğru orantı bulunmaktadır. Analiz süresi arttıkça bellek kullanımı da artmaktadır. Bu doğrultuda incelendiğinde, en fazla bellek kullanımı önerilen entropiye dayalı protein sayısallaştırma yöntemi ile gerçekleştirilmiş ve 234.87 MB bellek kullanımına erişilmiştir. Önerilen yöntem dışında sadece Atchley faktörleri ve Meiler parametreleri 200 MB'nin üstünde bellek kullanmışlardır. Geri kalan bütün protein sayısallaştırma yöntemleri 200 MB'nin altında bellek kullanmışlardır. En az bellek kullanan yöntem ise önerilen yöntemlerden birisi olan AVL ağacına dayalı olan yöntemdir. Bu yöntemin bellek kullanımı 140.34 MB olmuştur. AVL yönteminin en az bellek kullanmasının nedeni analiz sürecinin en hızlı olmasıdır. Çalışma zamanları kıyaslandığında ise sadece önerilen AVL ağacına dayalı gerçekleştirilen yöntemin zaman karmaşıklığı hesaplanabilmiştir. AVL ağacı kullanılarak gerçekleştirilen yöntemin çalışma zamanının $O(\log N)$ olması, yöntemin en hızlı olmasına ve en az bellek kullanmasına sebebiyet vermiştir. Yöntemlerin bellek kullanım boyutu ve analiz süreleri kullanılan donanıma ve donanımda bulunan uygulamalara göre değişiklik göstermektedir. Elde edilen sonuçlar bu açıdan yaklaşık değerler olup, farklı bir donanım platformu üzerinde farklı sonuçlar sergileyebilir.

8.3. Uygulama Sonuçlarının Değerlendirilmesi

Tablo 8.1'deki sonuçlar incelendiğinde, en etkisiz yöntemlerin Micheletti potansiyelleri ve Miyazawa enerjileri olduğu görülmektedir. Micheletti potansiyelleri ve Miyazawa enerjileri yapı-tabanlı yöntemlerdir. Bu iki yöntemin en etkisiz yöntem olmasının nedeni çalışmada protein dizilimlerinin yapı bilgisine ihtiyaç duyulmaması olabilir. Bu nedenle bilgi kaybı yaşanmış olabilir. Fizikokimyasal-tabanlı yöntemlerin de genel olarak etkisiz olduğu görülmektedir. Bu kategoride yer alan Atchley faktörleri, hidrofobiklik, Meiler parametreleri ve EIIP, yapı-tabanlılara göre daha iyi performans gösterebilir de genel olarak etkisiz olmuşlardır. Bu durumun temel nedeni tıpkı yapı-tabanlı yöntemlerde olduğu gibi sayısallaştırma işleminin belirli bir bilgiye ihtiyaç duyulmasıdır. Fizikokimyasal-tabanlı yöntemlerde ise proteinlerin kimyasal bilgilerine ihtiyaç duyulmaktadır. Bu çalışmada da herhangi bir kimyasal bilgi incelenmediği için bilgi kaybı söz konusu olabilir. Evrim-tabanlı yöntemlerin performansları incelendiğinde, bu kategorideki protein sayısallaştırma

yöntemlerinin yapı-tabanlı ve fizikokimyasal-tabanlı yöntemlerden daha başarılı olduğu görülmektedir. Evrim-tabanlı yöntemler genellikle hizalama çalışmalarında kullanılmaktadır [234, 235]. Çalışmada hizalama işlemi gerçekleştirildikten sonra sınıflandırma işleminin yapılması, bu yöntemlerin başarımlarını ciddi oranda geliştirebilir. Karakter-tabanlı yöntemlerin ise daha başarılı olduğu görülmektedir. CPNR, ikili kodlama ve FIBHASH yöntemlerinin hepsi %85'in üzerinde bir doğruluk skoruna erişmiştir. Bu durumun nedeni bu yöntemler ile sayısallaştırma işlemi yapılırken belirli bir bilgiye ihtiyaç duyulmamasıdır. Bu nedenle bir bilgi kaybı söz konusu olmamaktadır. Bu da bu yöntemlerin etkili olmasını sağlamaktadır. Makine öğrenmesi-tabanlı bir yöntem olan AESNN1 protein sayısallaştırma yöntemi en etkili üçüncü yöntem olmuştur. Bunların yanı sıra etkili yöntemin entropi tabanlı yöntem olduğu görülmektedir. Bu durumun temel nedeni, bu yöntem ile sayısallaştırma işleminin statik olmamasıdır. Diğer tüm yöntemlerde protein dizilimleri hep aynı değerleri almaktadır. Protein diziliminin uzunluğu ya da kodon sayısı sayısallaştırmayı etkilememektedir. Ancak entropi yönteminde adaptif bir durum söz konusu olduğu için, dizilimin uzunluğu ve kodon sayısı, sayısallaştırma işleminin her bir protein dizilimi için farklı olmasına sebebiyet vermektedir. Bu da yöntemin daha etkili olmasına vesile olmaktadır. AVL-ağacı tabanlı protein sayısallaştırma yöntemi ise hem karakter-tabanlı hem de algoritmik-tabanlı bir yöntemdir. Bu yöntemde proteinlerin karakter bilgileri algoritmik bir yapıya dayandığı için, sadece AVL ve entropi-tabanlı protein sayısallaştırma yöntemleri harika bir sınıflandırma yapmıştır. Çalışmanın avantajları ve dezavantajları şu şekilde ifade edilebilir;

- Bu çalışma ile protein sayısallaştırma tekniklerinin SARS-CoV-2 virüsü ve insan proteinleri arasındaki viral-konak etkileşimlerini belirlemede başarılı olduğu gözlemlenmiştir. Bu başarının nedenlerinden birisi de kullanılan derin öğrenme modelidir.
- Uygulama sonuçlarına göre hesaplama-tabanlı yaklaşımların bu alanda etkili bir şekilde kullanılabileceği belirlenmiştir. Literatürde bu alanda çok sayıda çalışma bulunmamaktadır. O yüzden çalışmanın çeşitli diğer yaklaşımlarla da (başka sayısallaştırma yöntemleri, hizalama algoritmaları, vb.) desteklenmesi gerekmektedir.
- SARS-CoV-2 ve insan proteinleri arasındaki yeni viral-konak etkileşimlerinin bulunması veri sayısını arttıracaktır. Bu nedenle bu çalışmada elde edilen sonuçlar daha da geliştirilebilir. Veri sayısının çok olması derin öğrenme modelinin daha iyi öğrenmesini sağlayacak ve sonuçlar daha güvenilir olacaktır.
- Farklı bir derin öğrenme algoritması ile sonuçlar farklı bir şekilde değerlendirilebilir. Çalışmanın çeşitliliğini ve başarımını koruması için, diğer derin öğrenme algoritmaları ile de değerlendirilmesi gerekmektedir.

Bu çalışmada SARS-CoV-2 ve insan proteinleri arasındaki viral-konak etkileşimleri hesaplama-tabanlı yaklaşımlarla tahmin edilmiştir. Çalışma dört farklı aşamadan meydana gelmiştir. Birinci aşamada SARS-CoV-2 virüsü ve insan proteinleri arasındaki viral-konak

etkileşimleri elde edilmiş ve pozitif veri seti oluşturulmuştur. Bunun yanı sıra etkileşim halinde olmayan protein dizilimleri de elde edilerek negatif veri seti oluşturulmuştur. İkinci aşamada hem negatif ve hem de pozitif veri setinde bulunan protein dizilimleri çeşitli protein sayısallaştırma yöntemleri ile sayısal hale getirilmiş ve derin öğrenme algoritması ile sınıflandırılmak üzere hazırlanmıştır. Üçüncü aşamada derin öğrenme modeli tasarlanmış ve bu bağlamda BiLSTM algoritması kullanılmıştır. Son aşamada ise protein sayısallaştırma yöntemlerinin başarımları doğruluk, F1-skor ve AUC skorları ile belirlenmiş ve performansları kıyaslanmıştır. En etkisiz protein sayısallaştırma yöntemi Micheletti potansiyelleri olmuş ve %75,39 oranında bir doğruluk elde edilmiştir. En etkili sınıflandırma işlemi ise Shannon entropi hesabına dayanan protein sayısallaştırma yöntemi ile gerçekleştirilmiş ve bu yöntem ile %94,74 oranında bir doğruluk elde edilmiştir. Sonuçlar genel olarak incelendiğinde en etkisiz yöntemlerin yapı-tabanlı yöntemler olduğu belirlenmiştir. Bunun yanı sıra, fizikokimyasal-tabanlı yöntemler için en etkili yöntemin Atchley faktörlerinin olduğu gözlemlenmiştir. Bu yöntem ile %82,58 oranında doğruluk elde edilirken, 0,83 oranında AUC skoru hesaplanmıştır. Ardından en etkili yöntemler evrim-tabanlı yöntemler olmuştur. Bu yöntemler içerisinde ise PAM250 matrisi başarılı bir sınıflandırma işlemi gerçekleştirerek %87,63 doğruluk skoruna erişirken, 0,87 oranında AUC skoru elde etmiştir. En etkili dördüncü yöntem ise makine öğrenmesi-tabanlı yöntemlerden birisi olan AESNN1 protein sayısallaştırma yöntemi ile elde edilmiştir. Bu yöntem ile %88,94 oranında doğruluk skoru hesaplanırken, 0,81 oranında AUC skoru gözlemlenmiştir. En etkili üçüncü sayısallaştırma yöntemi ise karakter-tabanlı protein sayısallaştırma yöntemlerinden birisi olan FIBHASH ile elde edilmiştir. Bu yöntem ile %89,33 doğruluk skoru elde edilirken, 0,88 oranında AUC hesaplanmıştır. Algoritmik-tabanlı bir yöntem olan AVL ise en etkili ikinci sınıflandırma işlemini gerçekleştirmiştir. Bu yöntemin doğruluk skoru %90,44 olarak hesaplanmıştır. AUC skoru ise 0,91 olarak belirlenmiştir. Sonuçlar genel olarak irdelendiğinde, bütün protein sayısallaştırma yöntemlerinin, tasarlanan derin öğrenme modeli ile başarılı bir sınıflandırma işlemi yaptığı gözlemlenmiştir. Bu sayede bu alanda hesaplama-tabanlı yaklaşımların kullanılması, hem virüsünün yapısının daha iyi anlaşılabilmesine hem de insan proteinlerle etkileşimlerinin sonucunun daha hızlı belirlenmesine yardımcı olacaktır.

9. SONUÇ

Bu tez çalışmasında, protein-protein etkileşimlerinin belirlenmesi, viral genomlardaki proteinler arasındaki etkileşimlerin saptanması ve protein ailelerinin sınıflandırılmasına yönelik yeni yaklaşımlar önerilmiştir. Önerilen bu yaklaşımlar yeni bir kategori olarak tanımlanan algoritmik-tabanlı şeklinde isimlendirilmiştir. Bu çerçevede algoritmik tabanlı 3 farklı protein sayısallaştırma yöntemi önerilmiştir. Bu yöntemler AVL ağacına dayalı protein sayısallaştırma yöntemi, Shannon entropisine dayalı protein sayısallaştırma yöntemi ve Fibonacci sayılarını ile hash tablosunu içeren hibrit bir model olan FIBHASH yöntemidir. Algoritmik-tabanlı olarak önerilen bu yöntemler güncel biyoenformatik çalışmaları üzerinde denenmiş ve literatürde bulunan mevcut yöntemlerle kıyaslanması gerçekleştirilmiştir.

Tezin amaçları doğrultusunda geliştirilen yaklaşımlar ve kullanım alanları şu şekilde ifade edilmektedir;

- AVL ağacı kullanılarak yeni bir protein sayısallaştırma yöntemi geliştirilmiş ve bu yöntem COVID-19 proteinleri ile insan proteinleri arasındaki etkileşimleri saptamak için kullanılmıştır. Önerilen yöntemin dışında literatürde kullanılan diğer protein sayısallaştırma yöntemleri de değerlendirilmiş ve önerilen yaklaşımın başarımı kıyaslanmıştır. Derin öğrenme modeli kullanılarak gerçekleştirilen çalışmada, önerilen protein sayısallaştırma yönteminin proteinlerin etkileşimlerini saptamada etkili olduğu gözlemlenmiştir. Protein-protein etkileşimlerine yönelik gerçekleştirilen çalışmalar incelendiğinde, elde edilen doğruluk başarımları genellikle %80 ve %99 arasında değişiklik göstermektedir [51, 69, 71-75]. Önerilen AVL yönteminin COVID-19 protein etkileşimlerinde kullanılmasıyla ortalama %85,33 oranında bir doğruluk skoru elde edilmiştir. Çalışma kapsamında kullanılan diğer protein sayısallaştırma yöntemleri ise ortalama %71,10 doğruluk skoruna erişmiştir. Bu da literatürdeki çalışmalar ile kıyaslandığında önerilen AVL yönteminin başarılı bir yöntem olduğunu göstermektedir. Önerilmiş olan yöntemde ikili arama ağacı olan AVL ağacı kullanılmıştır. Bu ikili arama ağacı algoritma analizinde ve bilgisayar bilimlerinde sıklıkla kullanılmaktadır. Dengeli bir yapıya sahip olması ve zaman karmaşıklığının $O(\log N)$ olması bu yöntemin önerilmesine vesile olmuştur. Ayrıca, önerilen yöntemin ağaç yapısına ve algoritmik süreçlere sahip olması, yöntemin algoritmik-tabanlı kategoride değerlendirilmesine yol açmıştır.
- Fibonacci sayılarına ve hash tablosuna dayalı yeni bir protein sayısallaştırma yöntemi geliştirilmiş ve yöntemin başarımı protein ailelerini sınıflandırmada değerlendirilmiştir. Çalışma kapsamında başka protein sayısallaştırma yöntemleri de kullanılmış ve önerilen protein sayısallaştırma yönteminin performansı kıyaslanmıştır. Çalışma kapsamında protein aileleri dört farklı senaryo kullanılarak sınıflandırılmış ve bu süreç doğrultusunda

çeşitli derin öğrenme modeli kullanılmıştır. Önerilen yöntemin etkili olduğu saptanmış ve sınıflandırma işleminde başarılı olduğu belirlenmiştir. Literatürde protein ailelerinin sınıflandırılmasına yönelik gerçekleştirilen çalışmalarda doğruluk skorları genellikle %74,33 ve %94,84 arasında değişiklik göstermektedir [82-87]. Önerilen FIBHASH yöntemi ile gerçekleştirilen çalışmada ise elde edilen doğruluk skoru %92,77 olmuştur. Bunlara ek olarak, çalışma doğrultusunda kullanılan diğer protein sayısallaştırma yöntemleri ise ortalama %83,22 doğruluk değerine ulaşmışlardır. Bu da literatürdeki çalışmalar ile kıyaslandığında önerilen FIBHASH yönteminin başarılı bir yöntem olduğunu göstermektedir. Önerilmiş olan bu hibrit yöntemin içeriğinde ise hash tablosu ve Fibonacci sayıları bulunmaktadır. Hash tablosu verilerin hesaplama karmaşıklığını azaltmak için kullanılan bir veri yapısıdır. Bunun yanı sıra, Fibonacci sayıları ise arama algoritması olarak kullanılabilir. Önerilen bu hibrit yöntem bu algoritmik yapısı nedeniyle yeni bir kategorisel yaklaşım olan algoritmik-tabanlı protein sayısallaştırma kategorisinde değerlendirilmiştir.

- Viral genomlarda bulunan proteinler arasındaki etkileşimleri tahmin etmek için entropi tabanlı yeni bir protein sayısallaştırma yöntemi önerilmiştir. Önerilen yöntemin performansını kıyaslamak için farklı protein sayısallaştırma yöntemleri değerlendirilmiştir. Proteinler arasındaki etkileşimi saptamak için birlikte evrim analizi kullanılmış ve Pearson korelasyon katsayıları hesaplanmıştır. Çalışmada bu alanda en çok kullanılan yöntemlerden birisi olan MirrorTree yaklaşımının üretmiş olduğu yanlış pozitif sonuçlar gösterilmiş ve önerilen protein sayısallaştırma yönteminin başarımı ön plana çıkarılmıştır. Önerilen protein sayısallaştırma yöntemi Shannon entropisine dayanmaktadır. Shannon entropisi günümüzde verilerin sıkıştırılması ve yorumlanması için bilgisayar bilimlerinde sıklıkla kullanılmaktadır. Önerilen yöntem bu algoritmik yapısı nedeniyle yeni bir kategorisel yaklaşım olan algoritmik-tabanlı protein sayısallaştırma kategorisinde değerlendirilmiştir.
- Çalışmada toplamda üç adet yeni protein sayısallaştırma yöntemi önerilmiştir. Bu yöntemler AVL, FIBHASH ve entropi-tabanlı yöntemlerdir. Önerilen ve literatürde kullanılan yöntemlerin başarımları ayrıca viral-konak etkileşimlerini tahmin etmek için kullanılmış ve bu yöntemlerin başarımları kıyaslanmıştır. Önerilen yöntemler içerisinde en etkili sınıflandırma işlemi entropi-tabanlı yöntem ile elde edilmiş ve %94,74 oranında doğruluk skoru hesaplanmıştır. Bunun yanı sıra, önerilen diğer bir yöntem olan AVL ile de %90,44 oranında doğruluk skoru elde edilmiştir. Önerilen başka bir yöntem olan FIBHASH ile de %89,33 oranında doğruluk skoru gözlemlenmiştir. Önerilen yöntemlerin ortalama doğruluk değerleri %91,5 olmuştur. Literatürde en çok kullanılan yöntemlerin (AESNN1, Atchley faktörleri, ikili-kodlama, BLOSUM62, CPNR, EIIP, hidrofobiklik,

Meiler parametreleri, Micheletti potansiyelleri, Miyazawa enerjileri ve PAM250) ortalama doğruluk değeri ise %82,88 olmuştur. Bu durum önerilen yöntemlerin başarımlarını açıkça ortaya koymaktadır. Bunlara ek olarak, literatürdeki benzer çalışmalar incelendiğinde, doğruluk skorlarının genellikle %72 ve %97 arasında değişiklik gösterdiği gözlemlenmiştir [224, 231, 232]. Önerilen yöntemlerin ortalama başarısı, yöntemlerin etkili olduğunu göstermiştir.

9.1. Sonuçlar ve Tartışma

- Literatürde protein dizilimlerini sayısal ifadelere dönüştürmek için değerlendirilen çok sayıda protein sayısallaştırma yöntemi bulunmaktadır. Bu yöntemler hesaplama-tabanlı protein analizi çalışmalarında sıklıkla kullanılmaktadır. Proteinler arasındaki etkileşimlerin saptanması, protein fonksiyonlarının sınıflandırılması, protein ailelerinin belirlenmesi, ilaç-hedef etkileşimlerinin tahmin edilmesi, viral-konak etkileşimlerinin saptanması, vb. uygulamalarda sayısallaştırma yöntemlerinin önemi büyük olmaktadır. Çalışmada literatürde bulunan yöntemlere ek olarak protein sayısallaştırma yöntemleri önerilmiş ve önerilen yöntemler çeşitli çalışmalarda değerlendirilmiştir.
- Tez kapsamı boyunca gerçekleştirilen birinci yaklaşımda protein dizilimlerini sayısal ifadeleri dönüştürmek için AVL ağacına dayalı yeni bir yöntem geliştirilmiştir. Yöntem yapısı gereği literatürde bulunan diğer yöntemlerden farklı kategoride değerlendirilmiştir. Mevcut protein sayısallaştırma yöntemlerinin çoğu yapı-tabanlı, fizikokimyasal-tabanlı ve evrim-tabanlı yöntemlerdir. Ancak önerilen yöntem algoritmaya dayalı olduğu için algoritmik-tabanlı yöntem olarak kategorileştirilmiştir. AVL ağacına dayalı geliştirilen yöntem COVID-19 proteinleri ve insan proteinleri arasındaki etkileşimleri tahmin etmek için kullanılmıştır. Yapay zeka yöntemi ile gerçekleştirilen çalışmada önerilen yöntemin yüksek doğruluğa ulaştığı gözlemlenmiştir. Elde edilen sonuçlar doğrultusunda, önerilen yöntem literatüre kazandırılmıştır.
- Tez çalışması süresince gerçekleştirilen ikinci çalışmada ise Shannon entropisine dayalı protein sayısallaştırma yöntemi önerilmiş ve önerilen bu yöntem viral genomlar arasındaki protein etkileşimlerini tahmin etmek için kullanılmıştır. Geliştirilen yöntem, entropi hesabına dayandığında için, mevcut protein sayısallaştırma yöntemlerinden farklı olarak algoritmik-tabanlı kategoride değerlendirilmiştir. Literatürde bu kategoride başka protein sayısallaştırma yöntemi olmadığı için bu anlamda bir yenilik elde edilmiştir. Çalışmada birlikte evrim analizi kullanılmış ve proteinler arasındaki etkileşimler belirlenmiştir. Çeşitli protein sayısallaştırma yöntemleri kullanılarak, önerilen protein sayısallaştırma

yönteminin performansı kıyaslanmıştır. Çalışmanın sonunda önerilen yöntemin çok başarılı olduğu gözlemlenmiş ve bu yöntemin literatüre kazandırılması sağlanmıştır.

- Çalışmada önerilen son yöntemde ise Fibonacci sayılarına ve hash tablosuna dayalı yeni protein sayısallaştırma yaklaşımı geliştirilmiştir. Geliştirilen bu yöntemi ile protein aileleri sınıflandırılmış ve ailesi bilinmeyen proteinlerin kategorileştirilmesi sağlanmıştır. Literatür incelendiğinde hibrit bir yöntemin geliştirilmediği gözlemlenmiş ve önerilen yöntem bu anlamda ilk olmuştur. Bunlara ek olarak, geliştirilen yöntem Fibonacci sayılarını ve hash tablosunu kullandığı için, algoritmik-tabanlı protein sayısallaştırma yöntemi kategorisinde değerlendirilmiştir. Literatürde bu kategoriye ait başka bir protein sayısallaştırma yöntemi bulunmamaktadır. Diğer çalışmalarda olduğu gibi, bu çalışmada da önerilen protein sayısallaştırma yöntemi çeşitli yaklaşımlarla kıyaslanmış ve performansı değerlendirilmiştir. Çalışmanın sonunda önerilen yöntemin protein ailelerini sınıflandırmada etkili olduğu gözlemlenmiş ve literatüre yeni bir yöntem kazandırılmıştır.
- Çalışmada gerçekleştirilen son uygulamada ise SARS-CoV-2 virüsü ve insan proteinleri arasındaki viral-konak etkileşimleri tahmin edilmiştir. Tahmin işleminde önerilen üç yöntemin hem kendi içerisinde hem de literatürdeki diğer protein sayısallaştırma yöntemleri ile karşılaştırılması amaçlanmıştır. Çalışmanın sonunda önerilen yöntemlerin, literatürdeki yöntemlerden daha etkili ve başarılı olduğu gözlemlenmiştir. Bunun yanı sıra, önerilen yöntemler içerisinde en etkili yöntemin ise entropi-tabanlı protein sayısallaştırma yöntemi olduğu gözlemlenmiştir. Bu durumun en büyük nedeni, bu yöntemin adaptif bir yöntem olması yani dinamik bir yapıya sahip olmasıdır.

9.2. Öneriler

- Önerilen protein sayısallaştırma yöntemleri, proteinler arasındaki etkileşimleri tahmin etmek ve protein ailelerini sınıflandırmak için kullanılmıştır. Önerilen bu yöntemler; protein fonksiyonlarının sınıflandırılması, ilaç-hedef etkileşimlerinin belirlenmesi, viral-konak etkileşimlerinin tahmin edilmesi, kanserli protein dizilimlerinin saptanması, protein yapılarının tahmin edilmesi, kodlayan ve kodlamayan RNA bölgelerinin belirlenmesi, vb. çalışmalarda değerlendirilebilir.
- Gerçekleştirilen çalışmalarda derin öğrenme algoritmaları kullanılmıştır. Bu algoritmalarından farklı olarak diğer derin öğrenme algoritmaları kullanılabilir ve bu sayede gerçekleştirilen çalışmalar ve önerilen yöntemlerin başarımları geliştirilebilir.
- Benzer şekilde derin öğrenme algoritmalarından farklı olarak makine öğrenmesi algoritmaları kullanılabilir ve algoritmaların performansları kıyaslanabilir. Özellik çıkarma adımıyla kullanılacak olan yöntemler araştırılıp yöntemin başarımları artırılabilir.

- Geliştirilen protein sayısallaştırma yöntemleri ortak bir çalışmada kullanılabilir ve performansları kıyaslanabilir.
- Önerilen yöntemlerin performansı, günümüzde pandemiye neden olan SARS-CoV-2 virüsüne ait proteinlerin analizi üzerinde denenebilir ve bu sayede klinik çalışmalar için zemin hazırlanabilir.



TERİMLER SÖZLÜĞÜ

AAB (Amino Asit Bileşimi): Amino asitlerin oluşmasında rol oynayan bileşimlerdir. Bunlar nitrojen, karbon, hidrojen ve oksijendir. Bunlara ek olarak, değişken bir yan zincir grubu da amino asitin oluşmasında etkili olmaktadır.

AVL (Adelson, Velski ve Landi) Ağacı: Adelson, Velsky ve Landis tarafından geliştirilmiş sürekli dengeli olan bir ikili arama ağacıdır. Ağacın ismi geliştiricilerin isimlerinin baş harflerinden gelmektedir.

BLOSUM (Blok İkame Matrisi): Proteinlerin kıyaslanmasıyla elde edilen blokların, değişim frekanslarının gözlemlenmesi sonucu oluşan matrislerdir. Genel bir ifade ile matris hizalama prosedürüdür.

BS (Biyolojik Süreç): Karmaşık yapıda bulunan organik maddelerin basit veya daha kararlı maddelere dönüşümünün mikroorganizmalar tarafından gerçekleştirildiği metabolik süreçlerin tamamıdır.

ÇDH (Çoklu Dizilim Hizalama): Genomik dizilimlerdeki semboller arasına boşlukların eklenmesiyle iki ya da daha çok diziliminin hizalanmasıdır. Bu işlemin amacı, dizilimler arasında eşleşen sembollerin sayısını minimum düzeyde boşuk ekleyerek arttırmaktır.

DNA (Deoksiribo Nükleik Asit): Deoksiribo Nükleik Asit ya da kısaca DNA, tüm organizmaların ve bazı virüslerin canlılık işlevleri ve biyolojik gelişmeleri için gerekli olan genetik talimatları içeren bir nükleik asittir.

GO (Gen Ontolojisi): Gen Ontolojisi ya da GO, herhangi bir organizmadaki genleri ve gen ürünlerini tanımlayıp, bu verileri işlevsel düzeyde hesaplamalı olarak analiz edebilmek için geliştirilmiş olan bir biyoenformatik kaynağıdır.

HB (Hücresel Bileşen): Hücresel bileşen, hücrelerin ve buna bağlı olarak canlı organizmaların oluşturduğu karmaşık düzeydeki biyomoleküller ve yapılarıdır.

MF (Moleküler Fonksiyon): Moleküler fonksiyon, bir gen ürününün moleküler seviyedeki bağlanma veya enzim katalizi gibi ögesel faaliyetleri içiren işlevdir.

mRNA (Haberci RNA): Proteinlerin kendine özgü dizilimlerinin kodlandığı ve proteinlerin sentezlenmesini sağlayan RNA molekülleridir.

Nükleik Asit: Nükleik asitler, bütün canlı hücrelerde ve virüslerde bulunan, nükleotit birimlerden oluşmuş polimerlerdir. En yaygın nükleik asitler DNA ve RNA'dır.

Nükleobaz: Nükleobazlar, DNA ve RNA'daki şekerlere bağlı olan azot yapılı kimyasal moleküllerdir. Bu bazlar adenin, guanin, sitozin (sadece DNA'da), timin ve urasil (sadece RNA'da) bazlarıdır.

Nükleotit: Nükleotit, bir fosfat, beş karbonlu bir şeker ve bir azotlu organik bazdan oluşan bir kimyasal bileşiktir.

PAM (Nokta Kabul Edilen Mutasyon): PAM olarak da bilinen nokta kabul edilen mutasyon, bir proteinin birincil yapısındaki tek bir amino asidin, doğal seçim süreçleri tarafından kabul edilen başka bir tek amino asitle değiştirilmesidir.

PPE (Protein-Protein Etkileşimi): Proteinler arasındaki etkileşimler hücreler arası iletişim, hücrelerin metabolik gelişimi gibi süreçleri temsil etmektedir. Protein-protein etkileşimleri, zincirler arasında bulunan ve kovalent olmayan temaslar ile meydana gelmektedir.

PSSM (Konuma Özgü Puanlama Matrisi): Protein dizilimlerinde bulunan kısa motifleri saptayabilmek için kullanılan yöntemlerden birisidir. Bu yöntemde, belirli pozisyonlarda bulunan amino asitler ya da nükleotitler puanlanmaktadır.

RNA (Ribonükleik Asit): Ribonükleik Asit ya da kısaca RNA, nükleotitlerden oluşan bir polimer, bir nükleik asittir. Her nükleotit bir azotlu baz, bir riboz şeker ve bir fosfattan meydana gelmektedir.

rRNA (Ribozomal RNA): Ribozomal RNA ya da kısaca rRNA, ribozomlarda bulunan bir RNA türüdür. Ribozomal RNA'nın görevi, mRNA'da bulunan bilginin translasyon süreci sırasında amino asit dizilimine çevrilmesi için taşıyıcı RNA ile etkileşime geçmek ve uzayan peptit zincirine amino asit takmaktır.

tRNA (Taşıyıcı RNA): Taşıyıcı RNA ya da kısaca tRNA, hücrelerde protein sentezi sırasında büyüyen polipeptit zincirine belirli bir amino asit eklemekle görevli küçük bir RNA molekülüdür.

Y2H (Maya 2 Hibrit): Fiziksel etkileşimleri test ederek, proteinler arasındaki etkileşimleri ve protein-DNA arasındaki etkileşimleri saptamak için kullanılan bir moleküler biyoloji yöntemidir.

KAYNAKLAR

- [1] Özaltay, A. (2020). *Antibakteriyel Etkili Ofloksasin'in Multispektroskopik Olarak DNA ile Etkileşimi ve Moleküler Kenetlenme Çalışmaları*, Yüksek Lisans Tezi, Kahramanmaraş Üniversitesi, Fen Bilimleri Enstitüsü
- [2] Crick, F. (1970). Central Dogma of Molecular Biology, *Nature*, cilt 227(8), ss. 561-563
- [3] Sinden, R.R. (1994). *DNA Structure and Function*, Academic Press, New York
- [4] Griffiths, D., Uchiyama, M., Nurse, P. ve Wang, T.S. (2000). A Novel Mutant Allele of The Chromatin-Bound Fission Yeast Checkpoint Protein Rad17 Separates the DNA Structure Checkpoints, *Journal of Cell Science*, cilt 113(6), ss. 1075-1088
- [5] Nelson, D.L. ve Cox, M.M. (2017). *Lehninger Principles of Biochemistry*, W.H. Freeman and Company, USA
- [6] Eisenhaber, F.E., Persson, B. ve Argos, P. (2008). Protein Structure Prediction: Recognition of Primary, Secondary, and Tertiary Structural Features from Amino Acid Sequence, *Critical Reviews in Biochemistry and Molecular Biology*, cilt 30(1), ss. 1-94
- [7] Janin, J., Bahadur, R.P. ve Chakrabarti, P. (2008). Protein-Protein Interaction And Quaternary Structure, *Quarterly Reviews of Biophysics*, cilt 41(2), ss. 133-180
- [8] Chen, D., Wang, J., Yan, M. ve Bao, F.S. (2016). A Complex Prime Numerical Representation of Amino Acids for Protein Function Comparison, *Journal of Computational Biology*, cilt 23(8), ss. 669-677
- [9] Yoo, P.D., Zhou, B.B. ve Zomaya, A.Y. (2008). Machine Learning Techniques for Protein Secondary Structure Prediction: An Overview and Evaluation, *Current Bioinformatics*, cilt 3(2), ss. 74-86
- [10] Kyte, J. ve Doolittle, R.F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein, *Journal of Molecular Biology*, cilt 157, ss. 105-132
- [11] Meiler, J., Müller, M., Zeidler, A. ve Schmaschke, F. (2001). Generation And Evaluation Of Dimension-Reduced Amino Acid Parameter Representations By Artificial Neural Networks, *Molecular Modelling Annual*, cilt 7, ss. 360-369
- [12] Atchley, W.R., Zhao, J., Fernandes, A.D. ve Drüke, T. (2005). Solving The Protein Sequence Metric Problem, *PNAS*, cilt 102(18), ss. 6395-6400
- [13] Veljkovic, N., Glisic, S., Prljic, J., Perovic, V., Botta, M. ve Veljkovic, V. (2008). Discovery Of New Therapeutic Targets By The Informational Spectrum Method, *Current Protein and Peptide Science*, cilt 9(5), ss. 493-506
- [14] Xu, W. ve Miranker, D.P. (2004). A Metric Model Of Amino Acid Substitution, *Bioinformatics*, cilt 20(8), ss. 1214-1221
- [15] Pearson, W.R. (1990). Rapid And Sensitive Sequence Comparison With FASTP And FASTA, *Methods in Enzymology*, cilt 183, ss. 63-98
- [16] Henikoff, S. ve Henikoff, J.G. (1992). Amino Acid Substitution Matrices From Protein Blocks, *PNAS*, cilt 89, ss. 10915-10919
- [17] Hattori, T., Totsuka, M., Hobo, T., Kagaya, Y. ve Yamamoto-Toyoda, A. (2002). Experimentally Determined Sequence Requirement of ACGT-Containing Abscissic Acid Response Element, *Plant and Cell Physiology*, cilt 43(1), ss. 136-140

- [18] Miyazawa, S. ve Jernigan, R.L. (1985). Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation, *Macromolecules*, cilt 18, ss. 534-552
- [19] Micheletti, C., Seno, F., Banavar, J.R. ve Maritan, A. (2001). Learning Effective Amino Acid Interactions Through Iterative Stochastic Techniques, *Proteins*, cilt 42(3), ss. 422-431
- [20] Lin, K., May, A.C.W. ve Taylor, W.R. (2002). Amino Acid Encoding Schemes from Protein Structure Alignments: Multi-dimensional Vectors to Describe Residue Types, *Journal of Theoretical Biology*, cilt 216(3), ss. 361-365
- [21] Dong, Q., Jing, X., Zhou, A. ve Hu, X. (2020). *Biological Language Model: Theory And Application*, World Scientific, Singapore
- [22] Asgari, E. ve Mofrad, M.R.K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics, *PLOS ONE*, ss. 1-15
- [23] Klein, S. (2010). The Use of Biorelevant Dissolution Media to Forecast the In Vivo Performance of a Drug, *The AAPS Journal*, cilt 12(3), ss. 397-406
- [24] Kuramochi, J. ve Sakakibara, Y. (2005). Intensive in vitro experiments of implementing and executing finite automata in test tube, *XI International Conference on DNA Computing*, Londra, Kanada
- [25] Atanasov, A.G., Waltenberger, B., Pferschy-Wenzling, EM., Linder, T., Wawrosch, C., Uhrin, P., Temml, V., Wang, L., Schwaiger, S., Heiss, E.H., Rollinger, J.M., Schuster, D., Breuss, J.M., Bochkov, V., Mihovilovic, M.D., Kopp, B., Bauer, R., Dirsch, V.M. ve Stuppner, H. (2015). Discovery And Resupply Of Pharmacologically Active Plant-Derived Natural Products: A Review, *Biotechnology Advances*, cilt 33(8), ss. 1582-1614
- [26] Lipinski, C. ve Hopkins, A. (2004). Navigating Chemical Space for Biology and Medicine, *Nature*, cilt 432, ss. 855-861
- [27] Smith, H. (2001). Discovery of the Anthrax Toxin: The Beginning of Studies of Virulence Determinants Regulated In Vivo, *International Journal of Medical Microbiology*, cilt 291(6-7), ss. 411-417
- [28] Relman, D.A. (1998). Detection and Identification of Previously Unrecognized Microbial Pathogens, *Emerging Infectious Diseases*, cilt 4(3), ss. 382-389
- [29] Altuntaş V. ve Gök, M. (2020). Protein – Protein Etkileşimi Tespit Yöntemleri, Veri Tabanları ve Veri Güvenilirliği, *Avrupa Bilim ve Teknoloji Dergisi*, ss. 722-733
- [30] MacBeath, G. ve Schreiber, S.L. (2000). Printing Proteins as Microarrays for High-Throughput Function Determination, *Science*, cilt 289(5485), ss. 1760-1763
- [31] Pinkel, D., Seagraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, WL., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, BM., Gray, J.W. ve Albertson, D.G. (1998). High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays, *Nature Genetics*, cilt 20, ss. 207-211
- [32] Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., Andrews, B., Tyers, M. ve Boone, C. (2001). Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants, *Science*, cilt 294(5550), ss. 2364-2368
- [33] Kobe, B., Guncar, G., Buchholz, R., Huber, T., Maco, B., Cowieson, N., Martin, J.L., Marfori, M. ve Forwood, J.K. (2008). Crystallography and Protein-Protein Interactions: Biological Interfaces and Crystal Contacts, *Biochemical Society Transactions*, cilt 36(6), ss. 1438-1441

- [34] Urakubo, Y., Ikura, T. ve Ito, N. (2008). Crystal Structural Analysis of Protein–Protein Interactions Drastically Destabilized by A Single Mutation, *Biochemical Society Transactions*, cilt 36(6), ss. 1438-1441
- [35] Erten, S., Li, X., Bebek, G., Li, J. ve Koyutürk, M. (2009). Phylogenetic Analysis of Modularity in Protein Interaction Networks, *BMC Bioinformatics*, cilt 10, ss. 1-14
- [36] Pazos, F., Juan, D., Izarzugaza, J.M.G., Leon, E. ve Valencia, A. (2008). Prediction of Protein Interaction Based on Similarity of Phylogenetic Trees, *Functional Proteomics*, cilt 484, ss. 523-535
- [37] Grigoriev, A. (2001). A Relationship Between Gene Expression and Protein Interactions on the Proteome Scale: Analysis of the Bacteriophage T7 and the Yeast *Saccharomyces Cerevisiae*, *Nucleic Acid Research*, cilt 29(17), ss. 3513-3519
- [38] Sidhu, S.S., Lowman, H.B., Cunningham, C. ve Wells, J.A. (2000). Phage Display for Selection of Novel Binding Peptides, *Methods in Enzymology*, cilt 328, ss. 333-363
- [39] Michnick, S.W., Ear, P.H., Landry, C., Malleshaiah, M.K. ve Messier, V. (2011). Protein-Fragment Complementation Assays for Large-Scale Analysis, Functional Dissection and Dynamic Studies of Protein–Protein Interactions in Living Cells, *Signal Transduction Protocols*, cilt 756, ss. 395-425
- [40] Moresco, J.J., Carvalho, P.C. ve Yates, J.R. (2010). Identifying Components of Protein Complexes In *C. Elegans* Using Co-Immunoprecipitation and Mass Spectrometry, *Journal of Proteomics*, cilt 73(11), ss. 2198-2204
- [41] Morell, M., Espargaro, A., Aviles, F.X. ve Ventura, S. (2007). Detection of Transient Protein–Protein Interactions by Bimolecular Fluorescence Complementation: The Abl-SH3 Case, *Proteomics*, cilt 7(7), ss. 1023-1036
- [42] Denison, C., Rudner, A.D., Gerber, S.A., Bakalarski, C.E., Moazed, D. ve Gygi, S.P. (2005). A Proteomic Strategy for Gaining Insights Into Protein Sumoylation in Yeast, *Molecular and Cellular Proteomics*, cilt 4(3), ss. 246-254
- [43] Golemis, E. (2002). Protein-Protein Interactions: A Molecular Cloning Manual, *The Quarterly Review of Biology*, cilt 77(3), ss. 307
- [44] Arnau, J., Lauritzen, C., Petersen, G.E ve Pedersen, J. (2006). Current Strategies for the Use of Affinity Tags and Tag Removal for the Purification of Recombinant Proteins, *Protein Expression and Purification*, cilt 48(1), ss. 1-13
- [45] Waugh, D.S. (2005). Making the Most of Affinity Tags, *Trends in Biotechnology*, cilt 23(6), ss. 316-320
- [46] Rappsilber, J. (2011). The Beginning of A Beautiful Friendship: Cross-Linking/Mass Spectrometry and Modelling Of Proteins and Multi-Protein Complexes, *Journal of Structural Biology*, cilt 173(3), ss. 530-540
- [47] Bruce, J.E. (2012). In Vivo Protein Complex Topologies: Sights Through A Cross-Linking Lens, *Proteomics*, cilt 12(10), ss. 1565-1575
- [48] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R. Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Quareshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. ve Rothberg, J.M. (2000). A Comprehensive Analysis of Protein-Protein Interactions in *Saccharomyces Cerevisiae*, *Nature*, cilt 403, ss. 623-627
- [49] Rutherford, S.L. (2000). From Genotype to Phenotype: Buffering Mechanisms and the Storage of Genetic Information, *BioEssays*, cilt 22(12), ss. 1095-1105

- [50] Alakuş, T.B. ve Türkoğlu, İ. (2019). Prediction of Protein-Protein Interactions with LSTM Deep Learning Model, *III International Symposium on Multidisciplinary Studies and Innovative Technologies*, Ankara, Türkiye
- [51] Wang, L., Wang, HF., Liu, SR., Yan, X. ve Song, KJ. (2019). Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest, *Scientific Reports*, cilt 9, ss. 1-12
- [52] Chen, KH., Wang, TF. ve Hu, YJ. (2019). Protein-Protein Interaction Prediction Using A Hybrid Feature Representation and A Stacked Generalization Scheme, *BMC Bioinformatics*, cilt 20, ss. 1-17
- [53] Sarkar, D. ve Saha, S. (2019). Machine-Learning Techniques for the Prediction of Protein-Protein Interactions, *Journal of Biosciences*, cilt 44(4), ss. 104-116
- [54] Vivona, S., Gardy, J.L., Ramachandran, S., Brinkman, F.S.L., Raghava, G.P.S., Flower, D.R. ve Filippini, F. (2008). Computer-Aided Biotechnology: From Immuno-Informatics to Reverse Vaccinology, *Trends in Biotechnology*, cilt 26(4), ss. 190-200
- [55] Lescot, M., Dehais, P., Thiis, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P. ve Rombauts, S. (2002). PlantCARE, A Database of Plant Cis-Acting Regulatory Elements and A Portal to Tools for in Silico Analysis of Promoter Sequences, *Nucleic Acid Research*, cilt 30(1), ss. 325-327
- [56] Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. ve Woolsey, J. (2006). DrugBank: A Comprehensive Resource for In Silico Drug Discovery and Exploration, *Nucleic Acid Research*, cilt 34, ss. 668-672
- [57] Chen, XW. ve Jeong, C.H. (2009). Sequence-Based Prediction of Protein Interaction Sites with An Integrative Method, *Bioinformatics*, cilt 25(5), ss. 585-591.
- [58] Huang, YA., You, ZH., Chen, X., Chan, K. ve Luo, X. (2016). Sequence-Based Prediction of Protein-Protein Interactions Using Weighted Sparse Representation Model Combined with Global Encoding, *BMC Bioinformatics*, cilt 17(1), ss. 1-11
- [59] Shin, WH., Christoffer, C.W. ve Kihara, D. (2017). In Silico Structure-Based Approaches to Discover Protein-Protein Interaction-Targeting Drugs, *Methods*, cilt 131, ss. 22-32
- [60] Shortridge, M.D. ve Varani, G. (2015). Structure Based Approaches for Targeting Non-Coding Rnas with Small Molecules, *Current Opinion in Structural Biology*, cilt 30, ss. 79-88
- [61] Wu, C., Zhu, J. ve Zhang, X. (2012). Integrating Gene Expression and Protein-Protein Interaction Network to Prioritize Cancer-Associated Genes, *BMC Bioinformatics*, cilt 13, ss. 1-10
- [62] Chen, R., Zhang, Z., Xue, Z., Wang, L., Fu, M., Lu, Y., Bai, L., Zhang, P. ve Fan, Z. (2015). Protein-Protein Interaction Network of Gene Expression in The Hydrocortisone-Treated Keloid, *International Journal of Dermatology*, cilt 54(5), ss. 549-554
- [63] Thieme, S. ve Groth, P. (2013). Genome Fusion Detection: A Novel Method to Detect Fusion Genes from SNP-Array Data, *Bioinformatics*, cilt 29(6), ss. 671-677
- [64] Latysheva, N.S., Oates, M.E., Maddox, L., Flock, T., Gough, J., Buljan, M., Weatheritt, R.J. ve Babu, M.M. (2016). Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer, *Molecular Cell*, cilt 63(4), ss. 579-592
- [65] De, S. ve Babu, M.M. (2010). Genomic Neighbourhood and the Regulation of Gene Expression, *Current Opinion in Cell Biology*, cilt 22(3), ss. 326-333
- [66] Oliver, B., Parisi, M. ve Clark, D. (2002). Gene Expression Neighborhoods, *Journal of Biology*, cilt 1(1), ss. 1-3

- [67] Palidwor, G., Reynaud, E.G. ve Andrade-Navarro, M.A. (2006). Taxonomic Colouring of Phylogenetic Trees of Protein Sequences, *BMC Bioinformatics*, cilt 7, ss. 1-4
- [68] Pazos, F. ve Valencia, A. (2001). Similarity of Phylogenetic Trees As Indicator of Protein-Protein Interaction, *Protein Engineering Design and Selection*, cilt 14(9), ss. 609-614
- [69] Chen, Y., Xu, J., Yang, B., Zhao, Y. ve He, W. (2012). A Novel Method for Prediction of Protein Interaction Sites Based On Integrated RBF Neural Networks, *Computers in Biology and Medicine*, cilt 42(4), ss. 402-407
- [70] Martin, S., Roe, D. ve Faulon, J.L. (2005). Predicting Protein-Protein Interactions Using Signature Products, *Bioinformatics*, cilt 21(2), ss. 218-226
- [71] Li, H., Gong, X.J., Yu, H. Ve Zhou, C. (2018). Deep Neural Network Based Predictions of Protein Interactions Using Primary Sequences, *Molecules*, cilt 23(8), ss. 1-16
- [72] Ahmed, M.S. (2017). *SIGNET: A Neural Network Architecture for Predicting Protein-Protein Interactions*, Yüksek Lisans Tezi, Batı Ontario Üniversitesi, Bilgisayar Bilimleri
- [73] Niu, B., Liang, C., Lu, Y., Zhao, M., Chen, Q., Zhang, Y., Zheng, L. ve Chou, K.C. (2020). Glioma Stages Prediction Based On Machine Learning Algorithm Combined with Protein-Protein Interaction Network, *Genomics*, cilt 112(1), ss. 837-847
- [74] Chen, K.H., Wang, T.F. ve Hu, Y.J. (2019). Protein-Protein Interaction Prediction Using A Hybrid Feature Representation and A Stacked Generalization Scheme, *BMC Bioinformatics*, cilt 20, ss. 1-17
- [75] Li, Y., Li, L.P., Wang, L., Yu, C.Q., Wang, Z. ve You, Z.H. (2019). An Ensemble Classifier to Predict Protein-Protein Interactions by Combining PSSM-based Evolutionary Information with Local Binary Pattern Model, *International Journal of Molecular Sciences*, cilt 20(14), ss. 1-12
- [76] Sun, T., Zhou, B., Lai, L. ve Pei, J. (2017). Sequence-Based Prediction of Protein Protein Interaction Using A Deep-Learning Algorithm, *BMC Bioinformatics*, cilt 18(1), ss. 1-8
- [77] Nguyen, N.P., Nute, M., Mirarab, S. ve Warnow, T. (2016). HIPPI: Highly Accurate Protein Family Classification with Ensembles Of HMMs, *BMC Genomics*, cilt 17, ss. 1-12
- [78] Dawson, N., Sillitoe, I., Marsden, R.L.ve Orenge, C.A. (2017). The Classification of Protein Domains, *Methods in Molecular Biology*, cilt 1525, ss. 137-164
- [79] Enright, A.J., Dongen, S.V. ve Ouzounis, C.A. (2002). An Efficient Algorithm for Large-Scale Detection of Protein Families, *Nucleic Acid Research*, cilt 30(7), ss. 1575-1584
- [80] Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. ve Yuan, Y. (1998). Predicting Function: From Genes to Genomes and Back, *Journal of Molecular Biology*, cilt 283(4), ss. 707-725
- [81] Remmert, M., Biegert, A., Hauser, A. ve Söding, J. (2011). Hhblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment, *Nature Methods*, cilt 9, ss. 173-175
- [82] Vazhayil, A., Vinayakumar, R. ve Soman, K.P. (2018). DeepProteomics: Protein Family Classification using Shallow and Deep Networks, *arXiv*, ss. 1-17
- [83] Naveenkumar, K.S., Harun, B.R.M., Vinayakumar, R. ve Soman, K.P. (2018). Protein Family Classification using Deep Learning, *bioRxiv*, ss. 1-7
- [84] Lee, T.K. ve Nguyen, T. (2019). Protein family classification with neural networks, Teknik rapor, İngiltere
- [85] Zhang, D. ve Kabuka, M.R. (2018). Protein family classification with multi-layer graph convolutional networks, *IEEE International Conference on Bioinformatics and Biomedicine*, Madrid, İspanya

- [86] Seo, S., Oh, M., Park, Y. ve Kim, S. (2018). Deepfam: Deep Learning Based Alignment-Free Method for Protein Family Modeling and Prediction, *Bioinformatics*, cilt 34(13), ss. 254-262
- [87] Zhang, D. ve Kabuka, M.R. (2021). Protein Family Classification from Scratch: A CNN Based Deep Learning Approach, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, cilt 18(5), ss. 1996-2007
- [88] Yusuf, S.M., Zhang, F., Zeng, M. ve Li, M. (2021). DeepPPF: A Deep Learning Framework for Predicting Protein Family, *Neurocomputing*, cilt 428, ss. 19-29
- [89] Zhang, D. ve Kabuka, M. (2019). Multimodal Deep Representation Learning for Protein Interaction Identification and Protein Family Classification, *Bioinformatics*, cilt 20, ss. 1-14
- [90] Sandaruwan, P.D. ve Wannige, C.T. (2021). An Improved Deep Learning Model for Hierarchical Classification of Protein Families, *Plos ONE*, cilt 16 (10), ss. 1-15
- [91] Rifaioğlu, A.S., Doğan, T., Martin, M.J., Atalay, R.C. ve Atalay, V. (2019). DEEPred: Automated Protein Function Prediction with Multi-Task Feed-Forward Deep Neural Networks, *Scientific Reports*, cilt 9(1), ss. 1-16
- [92] Liolis, K., Chen, M.A., Mayromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V.M. ve Kyrpides, N.C. (2010). The Genomes On Line Database (GOLD) In 2009: Status of Genomic and Metagenomic Projects and Their Associated Metadata, *Nucleic Acid Research*, cilt 38, ss. 1-19
- [93] Cao, R. ve Cheng, J. (2016). Integrated Protein Function Prediction by Mining Function Associations, Sequences, and Protein–Protein and Gene–Gene Interaction Networks, *Methods*, cilt 93, ss. 84-91
- [94] Bonetta, R. ve Velantino, G. (2020). Machine learning Techniques for Protein Function Prediction, *Proteins*, cilt 88(3), ss. 397-413
- [95] Lee, D., Redfern, O. ve Orengo, C. (2007). Predicting Protein Function from Sequence and Structure, *Nature Reviews: Molecular Cell Biology*, cilt 8(12), ss. 995-1005
- [96] Bernardes, J.S. ve Pedreira, C.E. (2013). A Review of Protein Function Prediction Under Machine Learning Perspective, *Recent Patents on Biotechnology*, cilt 7(2), ss. 122-141
- [97] Fa, R., Cozzetto, D., Wan, C. ve Jones, D.T. (2018). Predicting Human Protein Function with Multi-Task Deep Neural Networks, *Plos One*, cilt 13(6), ss. 1-16
- [98] Lobley, A.E., Nugent, T., Orengo, C.A. ve Jones, D.T. (2008). Ffpred: An Integrated Feature-Based Function Prediction Server for Vertebrate Proteomes, *Nucleic Acid Research*, cilt 36, ss. 297-302
- [99] Suthaharan, S. (2014). Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning, *ACM SIGMETRICS Performance Evaluation Review*, cilt 41(4), ss. 70-73
- [100] Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliva, N., Wald, R. ve Muharemagic, E. (2015). Deep Learning Applications and Challenges in Big Data Analytics, *Journal of Big Data*, cilt 2, ss. 1-21
- [101] Cai, Y., Wang, J. ve Deng, L. (2020). SDN2GO: An Integrated Deep Learning Model for Protein Function Prediction, *Frontiers in Bioengineering and Biotechnology*, cilt 8(391), ss. 1-11
- [102] Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H. ve Chen, Z. (2017). ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network, *Molecules*, cilt 22(10), ss. 1-14
- [103] Kulmanov, M., Khan, M.A., Hoehndorf, R. ve Wren, J. (2018). Deepgo: Predicting Protein Functions from Sequence and Interactions Using a Deep Ontology-Aware Classifier, *Bioinformatics*, cilt 34(4), ss. 660-668

- [104] You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H. ve Zhu, S. (2018). GOLabeler: Improving Sequence-Based Large-Scale Protein Function Prediction by Learning to Rank, *Bioinformatics*, cilt 34(14), ss. 2465-2473
- [105] Hakala, K., Kaewphan, S., Biorne, J., Mehryary, F., Moen, H., Tolvanen, M., Salakoski, T. ve Ginter, F. (2020). Neural Network and Random Forest Models in Protein Function Prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, ss. 1-10
- [106] Vascon, S., Frasca, M., Tripodi, R., Valentini, G. ve Pelillo, M. (2020). Protein Function Prediction as A Graph-Transduction Game, *Pattern Recognition Letters*, cilt 134, ss. 96-105
- [107] Makrodimitris, S., van Ham, R.C.H.J. ve Reinders, M.J.T. (2019). Improving Protein Function Prediction Using Protein Sequence And GO-Term Similarities, *Bioinformatics*, cilt 35(7), ss. 1116-1124
- [108] Gligorić, V., Barot, M. ve Bonneau, R. (2018). deepNF: Deep Network Fusion for Protein Function Prediction, *Bioinformatics*, cilt 34(22), ss. 3873-3881
- [109] Kulmanov, M. ve Hoehndorf, R. (2020). DeepGOPlus: Improved Protein Function Prediction from Sequence, *Bioinformatics*, cilt 36(2), ss. 422-429
- [110] Borgwardt, K.M., Ong, C.S., Schönauer, S., Vishwanathan, S.V.N., Smola, A.J. ve Kriegel, H.P. (2005). Protein Function Prediction Via Graph Kernels, *Bioinformatics*, cilt 21, ss. 47-56
- [111] Knowles, J. ve Gromo, G. (2003). Target Selection in Drug Discovery, *Nature Reviews Drug Discovery*, cilt 2, ss. 63-69
- [112] Wang, Y.B., You, Z.H., Yang, S., Yi, H.C., Chen, Z.H. ve Zheng, K. (2020). A Deep Learning-Based Method for Drug-Target Interaction Prediction Based on Long Short-Term Memory Neural Network, *BMC Medical Informatics and Decision Making*, cilt 20, ss. 1-9
- [113] Kapetanovic, I.M. (2008). Computer-Aided Drug Discovery and Development (CADD): In Silico-Chemico-Biological Approach, *Chemico-Biological Interactions*, cilt 171(2), ss. 165-176
- [114] Marcucci, F., Stassi, G. ve De Maria, R. (2016). Epithelial-Mesenchymal Transition: A New Target in Anticancer Drug Discovery, *Nature Reviews Drug Discovery*, cilt 15(5), ss. 311-325
- [115] Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijter, M.B., Matos, R.C., Tran, T.B., Whaley, R., Glennon, R.A., Hert, J., Thomas, K.L.H., Edwards, D.D., Shoichet, B.K. ve Roth, B.L. (2009). Predicting New Molecular Targets for Known Drugs, *Nature*, cilt 462(7270), ss. 175-181
- [116] Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y. ve Lu, H. (2017). Deep-Learning-Based Drug-Target Interaction Prediction, *Journal of Proteome Research*, cilt 16(4), ss. 1401-1409
- [117] Cheng, F., Zhou, Y., Li, J., Li, W., Liu, G. ve Tang, Y. (2012). Prediction Of Chemical-Protein Interactions: Multitarget-QSAR Versus Computational Chemogenomic Methods, *Molecular bioSystems*, cilt 8(9), ss. 2373-2384
- [118] He, Z., Zhang, J., Shi, X.H., Hu, L.L., Kong, X., Cai, Y.D. ve Chou, K.C. (2010). Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features, *Plos One*, cilt 5(3), ss. 1-8
- [119] Lee, I., Keum, J. ve Nam, H. (2019). DeepConv-DTI: Prediction of Drug-Target Interactions Via Deep Learning with Convolution on Protein Sequences, *Plos Computational Biology*, cilt 15(6), ss. 1-21
- [120] Pliakos, K. ve Vens, C. (2020). Drug-Target Interaction Prediction with Tree-Ensemble Learning and Output Space Reconstruction, *BMC Bioinformatics*, cilt 21(1), ss. 1-11

- [121] You, J., McLeod, R.D. ve Hu, P. (2019). Predicting drug-target interaction network using deep learning model, *Computational Biology and Chemistry*, cilt 80, ss. 90-101
- [122] Faulon, J.L., Misra, M., Martin, S., Sale, K. ve Sapra, R. (2008). Genome Scale Enzyme–Metabolite and Drug–Target Interaction Predictions Using the Signature Molecular Descriptor, *Bioinformatics*, cilt 24 (2), ss. 225-233
- [123] Xie, L., He, S., Song, X., Bo, X. ve Zhang, Z. (2018). Deep Learning-Based Transcriptome Data Classification for Drug-Target Interaction Prediction, *BMC Genomics*, cilt 19, ss. 1-10
- [124] Ding, Y., Tang, J. ve Guo, F. (2017). Identification Of Drug-Target Interactions Via Multiple Information Integration, *Information Sciences*, cilt 418-419, ss. 546-560
- [125] Cao, D.S., Zhang, L.X., Tan, G.S., Xiang, Z., Zeng, W.B., Xu, Q.S. ve Chen, A.F. (2014). Computational Prediction of Drug Target Interactions Using Chemical, Biological, and Network Features, *Molecular Informatics*, cilt 33(10), ss. 669-681
- [126] Cech, T.R. ve Steitz, J.A. (2014). The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones, *Cell*, cilt 157(1), ss. 77-94
- [127] Mattick, J.S. (2007). A New Paradigm for Developmental Biology, *The Journal of Experimental Biology*, cilt 210, ss. 1526-1547
- [128] Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammanna, H. ve Gingeras, T.R. (2007). RNA Maps Reveal New RNA Classes And a Possible Function for Pervasive Transcription, *Science*, cilt 316 (5380), ss. 1484-1488
- [129] Hung, T. ve Chang, H.Y. (2010). Long Noncoding RNA In Genome Regulation, *RNA Biology*, cilt 7(5), ss. 582-585
- [130] Li, A., Zhang, J. ve Zhou, Z. (2014). PLEK: A Tool for Predicting Long Non-Coding RNAs And Messenger RNAs Based on An Improved K-Mer Scheme, *BMC Bioinformatics*, cilt 15(1), ss. 1-10
- [131] Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S. ve Kageyama, Y. (2007). Small Peptide Regulators of Actin-Based Cell Morphogenesis Encoded by A Polycistronic mRNA, *Nature*, cilt 9, ss. 660-665
- [132] Galindo, M.I., Puevo, J.I., Fouix, S., Bishop, S.A. ve Couso, J.P. (2007). Peptides Encoded by Short Orfs Control Development and Define A New Eukaryotic Gene Family, *Plos Biology*, cilt 5(5), ss. 1052-1062
- [133] Hüttenhofer, A. ve Vogel, J. (2006). Experimental Approaches to Identify Non-coding RNAs, *Nucleic Acid Research*, cilt 34(2), ss. 635-646
- [134] Zhang, Y., Huang, H., Zhang, D., Oiu, J., Yang, J., Wang, K., Zhu, L., Fan, J. ve Yang, J. (2017). A Review on Recent Computational Methods for Predicting Noncoding RNAs, *BioMed Research International*, ss. 1-15
- [135] Tripathi, R., Patel, S., Kumari, V., Chakraborty, P. ve Varadwaj, P.K. (2016). DeepLNC, A Long Non-Coding RNA Prediction Tool Using Deep Neural Network, *Network Modeling Analysis in Health Informatics and Bioinformatics*, cilt 5(21), ss. 1-14
- [136] Liu, J., Gough, J. ve Rost, B. (2006). Distinguishing Protein-Coding from Non-Coding RNAs through Support Vector Machines, *Plos Genetics*, cilt 2(4), ss. 0529-0536
- [137] Tong, X. ve Liu, S. (2019). CPPred: Coding Potential Prediction Based on The Global Description of RNA Sequence, *Nucleic Acid Research*, cilt 47(8), ss. 1-10

- [138] Panwar, B., Arora, A. ve Raghava, G.P.S. (2014). Prediction And Classification Of ncRNAs Using Structural Information, *BMC Genomics*, cilt 15, ss. 1-13
- [139] Xu, Y., Zhao, X., Liu, S., Liu, S., Niu, Y., Zhang, W. ve Wei, L. (2019). LncPred-IEL: a long non-coding RNA prediction method using iterative ensemble learning, *International Conference on Bioinformatics and Biomedicine*, San Diego, ABD
- [140] Wucher, V., Legeai, F., Hedan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., Cirera, S., Fredholm, M., Botherel, N., Leegwater, P.A.J., Le Beguec, C., Fieten, H., Johnson, J., Alföldi, J., Andre, C., Lindblad-Toh, K., Hitte, C. ve Derrien, T. (2017). FEELnc: A Tool for Long Non-Coding RNA Annotation and Its Application to The Dog Transcriptome, *Nucleic Acid Research*, cilt 45(8), ss. 1-12
- [141] Hu, L., Xu, Z., Hu, B. ve Lu, Z.J. (2017). COME: A Robust Coding Potential Calculation Tool for lncRNA Identification and Characterization Based on Multiple Features, *Nucleic Acid Research*, cilt 45(1), ss. 1-13
- [142] Schneider, H.W., Raiol, T., Brigido, M.M., Walter, M.E.M.T. ve Stadler, P.F. (2017). A Support Vector Machine Based Method to Distinguish Long Non-Coding RNAs From Protein Coding Transcripts, *BMC Genomics*, cilt 18(1), ss. 1-14
- [143] Pian, C., Zhang, G., Chen, Z., Chen, Y., Zhang, J., Yang, T. ve Zhanhg, L. (2016). LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature, *Plos One*, cilt 11(5), ss. 1-20
- [144] Saetorm, P., Sneve, R., Kristiansen, K.I., Snove, O., Grünfeld, T., Rognes, T. ve Seeberg, E. (2005). Predicting Non-Coding RNA Genes in Escherichia Coli with Boosted Genetic Programming, *Nucleic Acid Research*, cilt 33(10). Ss. 3263-3270
- [145] Ode, H., Matsuyama, S., Hata, M., Neva, S., Kakizawa, J., Sugiura, W. ve Hoshino, T. (2007). Computational Characterization of Structural Role of The Non-Active Site Mutation M36I Of Human Immunodeficiency Virus Type 1 Protease, *Journal of Molecular Biology*, cilt 370(3), ss. 598-607
- [146] Lorch, M., Mason, J.M., Sessions, R.B. ve Clarke, A.R. (2000). Effects of Mutations on The Thermodynamics of A Protein Folding Reaction: Implications For The Mechanism Of Formation Of The Intermediate And Transition States, *Biochemistry*, cilt 39(12), ss. 3480-3485
- [147] Yamada, Y., Banno, Y., Yoshida, H., Kikuchi, R., Akao, Y., Murate, T. ve Nozawa, Y. (2006). Catalytic Inactivation of Human Phospholipase D2 by a Naturally Occurring Gly901Asp Mutation, *Archives of Medical Research*, cilt 37(6), ss. 696-699
- [148] Jones, R., Ruas, M., Gregory, F., Moulin, S., Delia, D., Manoukian, S., Rowe, J., Brookes, S. ve Peters, G. (2007). A CDKN2A Mutation in Familial Melanoma that Abrogates Binding of p16^{INK4a} to CDK4 but not CDK6, *Cancer Research*, cilt 67(19), ss. 9134-9141
- [149] Tiede, S., Cantz, M., Spranger, J. ve Bräulke, T. (2006). Missense Mutation in the N-acetylglucosamine-1-phosphotransferase Gene (GNPTA) in a Patient with Mucopolidosis II Induces Changes in the Size and Cellular Distribution of GNPTG, *Mutation in Brief*, ss. 1-8
- [150] Hanahan, D. ve Weinber, R.A. (2000). The Hallmarks of Cancer, *Cell*, cilt 100(1), ss. 57-70
- [151] Reva, B., Antipin, Y. ve Sander, C. (2011). Predicting The Functional Impact of Protein Mutations: Application to Cancer Genomics, *Nucleic Acid Research*, cilt 39(17), ss. 1-14
- [152] Sun, Y., Zhu, S., Ma, K., Liu, W., Yue, Y., Hu, G., Lu, H. ve Chen, W. (2019). Identification of 12 Cancer Types Through Genome Deep Learning, *Scientific Reports*, cilt 9(1), ss. 1-9

- [153] Guda, P., Chittur, S.V. ve Guda, C. (2009). Comparative Analysis of Protein-Protein Interactions in Cancer-Associated Genes, *Genomics, Proteomics and Bioinformatics*, cilt 7(1-2), ss. 25-36
- [154] Huang, CH., Peng, HS. ve Ng, KL. (2015). Prediction of Cancer Proteins by Integrating Protein Interaction, Domain Frequency, and Domain Interaction Data Using Machine Learning Algorithms, *BioMed Research International*, cilt 2015, ss. 1-15
- [155] Matsubara, T., Nacher, J.C., Ochiai, T., Hayashida, M. ve Akutsu, T. (2018). Convolutional Neural Network Approach to Lung Cancer Classification Integrating Protein Interaction Network and Gene Expression Profiles, *XVIII International Conference on Bioinformatics and Bioengineering*, Taichung, Tayvan
- [156] Hosseinzadeh, F., Kayvanjoo, A.H., Ebrahimi, M. ve Goliaei, B. (2013). Prediction Of Lung Tumor Types Based on Protein Attributes by Machine Learning Algorithms, *SpringerPlus*, cilt 2(1), ss. 1-14
- [157] de Guia, J.M., Winnipeg, M.B., Devaraj, M. ve Leung, C.K. (2019). DeepGx: deep learning using gene expression for cancer classification, *International Conference on Advances in Social Network Analysis and Mining*, Vancouver, Kanada
- [158] Hammad, A., Elshaer, M. ve Tang, X. (2021). Identification Of Potential Biomarkers with Colorectal Cancer Based on Bioinformatics Analysis and Machine Learning, *Mathematical Biosciences and Engineering*, cilt 18(6), ss. 8997-9015
- [159] He, B., Bergenstrahle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, A., Maaskola, J., Lundberg, J. ve Zou, J. (2020). Integrating Spatial Gene Expression and Breast Tumour Morphology Via Deep Learning, *Nature Biomedical Engineering*, cilt 4, ss. 827-834
- [160] Weiss, M. (2013). *Data Structures & Algorithm Analysis in C++*, Pearson, İngiltere
- [161] Nagaraj, N., Balasubramanian, K. ve Dey, S. (2013). A New Complexity Measure for Time Series Analysis and Classification, *The European Physical Journal Special Topics*, cilt 222, ss. 847-860
- [162] Tenenbaum, A.M. (1990). *Data Structures Using C*, Prentice Hall, ABD
- [163] Koffman, E.B. (2010). *Data Structures: Abstraction and Design Using Java*, Wiley, ABD
- [164] Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J. ve diğ. (2020). A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing, *Nature*, cilt 583 (7816), ss. 459-468
- [165] Öztürk, T., Talo, M., Yıldırım, E.A., Baloğlu, U.B., Yıldırım, Ö. ve Acharya, U.R. (2020). Automated Detection Of COVID-19 Cases Using Deep Neural Networks with X-Ray Images, *Computers in Biology and Medicine*, cilt 121, ss. 1-11
- [166] de Wit, E., van Doremalen, N., Falzarano, D. ve Munster, V.J. (2016). SARS and MERS: Recent Insights into Emerging Coronaviruses, *Nature Reviews: Microbiology*, cilt 14(8), ss. 524-534
- [167] Gates, B. (2020). Responding to Covid-19 — A Once-in-a-Century Pandemic?, *The New England Journal of Medicine*, cilt 382(18), ss. 1677-1679
- [168] Anderson, R.M., Hesterbeek, H., Klinkenberg, D. ve Hollingsworth, T.D. (2020). How Will Country-Based Mitigation Measures Influence the Course Of The COVID-19 Epidemic?, *Lancet*, cilt 395(10228), ss. 931-934
- [169] Sheahan, T.P., Sims, A.C., Leist, S.R., Schafer, A., Won, J., Brown, A.J., Montgomery, S.A., Hogg, A., Babusis, D., Clarke, M.O., Spahn, J.E., Bauer, L., Sellers, S., Porter, D., Feng, J.Y., Cihlar, T., Jordan, R., Denison, M.R. ve Bari, R.S. (2020). Comparative Therapeutic Efficacy of Remdesivir And Combination Lopinavir, Ritonavir, And Interferon Beta Against MERS-CoV, *Nature Communications*, cilt 11(1), ss. 1-14

- [170] Zarin, D.A., Tse, T., Williams, R.J., Califf, R.M. ve Ide, N.C. (2011). The ClinicalTrials.gov Results Database--Update and Key Issues, *The New England Journal of Medicine*, cilt 364(9), ss. 852-860
- [171] Goncarenco, A., Li, M., Simonetti, F.L., Shoemaker, B.A. ve Panchenko, A.R. (2017). Exploring Protein-Protein Interactions as Drug Targets for Anti-Cancer Therapy with In Silico Workflows, *Methods in Molecular Biology*, cilt 1647, ss. 221-236
- [172] Chene, P. (2006). Drugs Targeting Protein-Protein Interactions, *ChemMedChem*, cilt 1(4), ss. 400-411
- [173] Rao, V.S., Srinivas, K., Sujini, G.N. ve Kumar, G.N.S. (2014). Protein-Protein Interaction Detection: Methods and Analysis, *International Journal of Proteomics*, cilt 2014, ss. 1-13
- [174] Ruffalo, M. ve Bar-Joseph, Z. (2019). Protein Interaction Disruption in Cancer, *BMC Cancer*, cilt 19, ss. 1-10
- [175] Jothi, R., Kann, M.G. ve Przytcka, T.M. (2005). Predicting Protein-Protein Interaction by Searching Evolutionary Tree Automorphism Space, *Bioinformatics*, cilt 21, ss. 241-250
- [176] Khailany, R.A., Safdar, M. ve Özaslan, M. (2020). Genomic Characterization of a Novel SARS-CoV-2, *Gene Reports*, cilt 19, ss. 1-6
- [177] Dimitrova, M., Imbert, I., Kieny, M.P. ve Schuster, C. (2003). Protein-Protein Interactions Between Hepatitis C Virus Nonstructural Proteins, *Journal of Virology*, cilt 77(9), ss. 5401-5414
- [178] Song, J., Liu, Y., Gao, P., Hu, Y., Chai, Y., Zhou, S., Kong, C., Zhou, L., Ge, X., Guo, X., Han, J. ve Yanga, H. (2018). Mapping the Nonstructural Protein Interaction Network of Porcine Reproductive and Respiratory Syndrome Virus, *Journal of Virology*, cilt 92(24), ss. 1-25
- [179] Yau, S.S.T., Wang, J., Niknejad, A., Lu, C., Jin, N. ve Ho, YK. (2003). DNA Sequence Representation Without Degeneracy, *Nucleic Acid Research*, cilt 31(12), ss. 1-3
- [180] Jing, X., Dong, O., Hong, D. ve Lu, R. (2020). Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, cilt 17(6), ss. 1918-1931
- [181] Hagemans, D., van Belzen, I.A.E.M., Luengo, T.M. ve Rüdiger, S.G.D. (2015). A Script to Highlight Hydrophobicity and Charge on Protein Surfaces, *Frontiers in Molecular Biosciences*, cilt 2, ss. 1-11
- [182] Lehmann, J. ve Libchaber, A. (2008). Degeneracy Of the Genetic Code and Stability of The Base Pair at The Second Position of The Anticodon, *RNA*, cilt 14(7), ss. 1264-1269
- [183] Kozarzewski, B. (2012). A Method for Nucleotide Sequence Analysis, *Computational Methods in Science and Technology*, cilt 18(1), ss. 5-10
- [184] Machado, J.A.T. (2012). Shannon Entropy Analysis of the Genome Code, *Mathematical Problems in Engineering*, cilt 2012, ss. 1-13
- [185] Koslicki, D. (2011). Topological Entropy of DNA Sequences, *Bioinformatics*, cilt 27(8), ss. 1061-1067
- [186] Thomas, D., Finan, C., Newport, M.J. ve Jones, S. (2015). DNA Entropy Reveals a Significant Difference In Complexity Between Housekeeping And Tissue Specific Gene Promoters, *Computational Biology and Chemistry*, cilt 58, ss. 19-24
- [187] Phung, D., Tran D., Ma, W., Nguyen, P. ve Pham, T. (2014). Using Shannon entropy as EEG signal feature for fast person identification, *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Brugge, Belçika

- [188] Li, J., Zhang, L., Li, H., Ping, Y., Xu, O., Wang, R., Tan, R., Wang, Z., Liu, B. ve Wang, Y. (2019). Integrated Entropy-Based Approach for Analyzing Exons and Introns in DNA Sequences, *BMC Bioinformatics*, cilt 20, ss. 1-7
- [189] Zhang, J., Zhang, W. ve Yang, H. (2016). In Search of Coding and Non-Coding Regions of DNA Sequences Based on Balanced Estimation of Diffusion Entropy, *Journal of Biological Physics*, cilt 42, ss. 99-106
- [190] Karci, A. (2016). Fractional Order Entropy: New Perspectives, *Optik*, cilt 127(20), ss. 9172-9177
- [191] Karci, A. (2016). New Approach for Fractional Order Derivatives: Fundamentals and Analytic Properties, *Mathematics*, cilt 4(2), ss. 1-15
- [192] Hoang, T., Yin, C., Zheng, H., Yu, C., He, R.L. ve Yau, S.S.T. (2015). A New Method to Cluster DNA Sequences Using Fourier Power Spectrum Author Links Open Overlay Panel, *Journal of Theoretical Biology*, cilt 372, ss. 135-145
- [193] Yin, C. ve Yau, S.S.T. (2005). A Fourier Characteristic of Coding Sequences: Origins and A Non-Fourier Approximation, *Journal of Computational Biology*, cilt 12(9), ss. 1153-1165
- [194] Fuentes, A.R., Ginori, J.V.L. ve Abalo, R.G. (2006). Detection of coding regions in large DNA sequences using the short time Fourier transform with reduced computational load, *XI Iberoamerican Congress in Pattern Recognition*, Cancun, Meksika
- [195] Hou, W., Pan, Q., Peng, Q. ve He, M. (2017). A New Method to Analyze Protein Sequence Similarity Using Dynamic Time Warping, *Genomics*, cilt 109(2), ss. 123-130
- [196] Eisenberg, D., Weiss, R.M. ve Terwilliger, T.C. (1984). The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity, *Proceedings of the National Academy of Sciences of the United States of America*, cilt 81(1), ss. 140-144
- [197] Yin, C. ve Yau, S.S.T. (2015). An Improved Model for Whole Genome Phylogenetic Analysis by Fourier Transform, *Journal of Theoretical Biology*, cilt 382, ss. 99-110
- [198] Sedgwick, P. (2012). Pearson's Correlation Coefficient, *BMJ*, ss. 1-2
- [199] Pagano, R.R. (2012). *Understanding Statistics in the Behavioral Sciences*, Cengage Learning, Boston, ABD
- [200] Clark, G.W., Dar, VUN., Bezginov, A., Yang, J.M., Charlebois, R.L. ve Tillier, E.R.M. (2011). Using Coevolution to Predict Protein-Protein Interactions, *Methods in Molecular Biology*, cilt 781, ss. 237-256
- [201] Oliveria, L., Paiva, A.C.M. ve Vriend, G. (2002). Correlated Mutation Analyses on Very Large Sequence Families, *ChemBioChem*, cilt 3(10), ss. 1010-1017
- [202] Hakes, L., Lovell, S.C., Oliver, S.G. ve Robertson, D.L. (2007). Specificity In Protein Interactions and Its Relationship with Sequence Diversity and Coevolution, *Proceedings of the National Academy of Sciences of the United States of the America*, cilt 104(19), ss. 7999-8004
- [203] Gaboriaud, C., Bissery, V., Benchetrit, T. ve Mormon, J.P. (1987). Hydrophobic Cluster Analysis: An Efficient New Way to Compare and Analyse Amino Acid Sequences, *FEBS Letters*, cilt 224(1), ss. 149-155
- [204] Shapira, S.D., Gat-Viks, I., Shum, B.O.V., Dricot, A., Degrace, M.M., Liguio, W., Gupta, P.B., Hao, T., Silver, S.J., Root, D.E., Hill, D.E., Regeve, A. ve Hacohen, N. (2010). A Physical and Regulatory Map of Host-Influenza Interactions Reveals Pathways in H1N1 Infection, *Cell*, cilt 139(7), ss. 1255-1267

- [205] Ghedin, E., Sengamalay, N.A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D.J., Sitz, J., Koo, H., Bolotov, P., Denovoy, D., Tatusova, T., Bao, Y., George, K., Taylor, J., Lipman, D.J., Fraser, C.M., Taubenberger, J.K. ve Salzberg, S.L. (2005). Large-Scale Sequencing of Human Influenza Reveals the Dynamic Nature of Viral Genome Evolution, *Nature*, cilt 437(7062), ss. 1162-1166
- [206] Gamblin, S.J. ve Skehel, J.J. (2010). Influenza Hemagglutinin and Neuraminidase Membrane Glycoproteins, *The Journal of Biological Chemistry*, cilt 285(37), ss. 28403-28409
- [207] Roberts, P.C., Lamb, R.A. ve Compans, R.W. (1998). The M1 and M2 Proteins of Influenza A Virus Are Important Determinants in Filamentous Particle Formation, *Virology*, cilt 240(1), ss. 127-137
- [208] Portela, A. ve Digard, P. (2002). The Influenza Virus Nucleoprotein: A Multifunctional RNA-Binding Protein Pivotal to Virus Replication, *Journal of General Virology*, cilt 83, ss. 723-734
- [209] Licata, J.M., Johnson, R.F., Han, Z. ve Harty, R.N. (2004). Contribution of Ebola Virus Glycoprotein, Nucleoprotein, and VP24 to Budding of VP40 Virus-Like Particles, *Journal of Virology*, cilt 78(14), ss. 7344-7351
- [210] Watanabe, S., Noda, T., Halfmann, P., Jasenosky, L. ve Kawaoka, Y. (2007). Ebola Virus (EBOV) VP24 Inhibits Transcription and Replication of The EBOV Genome, *The Journal of Infectious Diseases*, cilt 196, ss. 284-290
- [211] Els, M.C., Laver, W.G. ve Air, G.M. (1989). Sialic Acid Is Cleaved from Glycoconjugates at The Cell Surface When Influenza Virus Neuraminidases Are Expressed from Recombinant Vaccinia Viruses, *Virology*, cilt 170(1), ss. 346-351
- [212] Matrosovich, M., Zhou, N., Kawaoka, Y. ve Webster, R. (1999). The Surface Glycoproteins of H5 Influenza Viruses Isolated from Humans, Chickens, and Wild Aquatic Birds Have Distinguishable Properties, *Journal of Virology*, cilt 73(2), ss. 1146-1155
- [213] Gao, Y., Zhang, Y., Shinya, K., Deng, G., Jiang, Y., Li, Z., Guan, Y., Tian, G., Li, Y., Shi, J., Liu, L., Zeng, X., Bu, Z., Xia, X., Kawaoka, Y. ve Chen, H. (2009). Identification Of Amino Acids in HA And PB2 Critical for The Transmission of H5N1 Avian Influenza Viruses in A Mammalian Host, *Plos Pathology*, cilt 5(12), ss. 1-11
- [214] Chen B.J., Leser, G.P., Jackson, D. ve Lamb, R.A. (2008). The Influenza Virus M2 Protein Cytoplasmic Tail Interacts with The M1 Protein and Influences Virus Assembly at The Site of Virus Budding, *Journal of Virology*, cilt 82(20), ss. 10059-10070
- [215] Chassey, B., Aublin-Gex, A., Ruggieri, A., Meyniel-Schicklin, L., Pradezynski, F., Davoust, N., Chantier, T., Tafforeau, L., Mangeot, P.E., Ciancia, C., Perrin-Cocon, L., Bartenschlager, R., Andre, P. ve Lotteau, V. (2013). The Interactomes of Influenza Virus NS1 And NS2 Proteins Identify New Host Factors and Provide Insights for ADAR1 Playing a Supportive Role in Virus Replication, *Plos Pathology*, cilt 9(7), ss. 1-14
- [216] Bullido, R., Gomez-Puertas, P., Saiz, M.J. ve Portela, A. (2001). Influenza A Virus NEP (NS2 Protein) Downregulates RNA Synthesis of Model Template RNAs, *Journal of Virology*, cilt 75(10), ss. 4912-4917
- [217] Sinha, S. (2017). The Fibonacci Numbers and Its Amazing Applications, *International Journal of Engineering Science Invention*, cilt 6(9), ss. 7-14
- [218] Parsaud-Sharma, D. ve O'Leary, J.P. (2015). Fibonacci Series, Golden Proportions, and the Human Biology, *Austin Journal of Surgery*, cilt 2(5), ss. 1-6
- [219] Perez, J.C. (1991). Chaos DNA and Neuro-Computers: A Golden Link, *Speculations in Science and Technology*, cilt 14(4), ss. 336-346

- [220] Negadi, T. (2015). A Mathematical Model for the Genetic Code(s) Based on Fibonacci Numbers and their q-Analogues, *NeuroQuantology*, cilt 13(3), ss. 259-272
- [221] Nimbe, P., Frimpong, S.O. ve Opoku, M. (2014). An Efficient Strategy for Collision Resolution in Hash Tables, *International Journal of Computer Applications*, cilt 99(10), ss. 35-41
- [222] Sanders, J.M., Monogue, M.L., Jodlowski, T.Z. ve Cutrell, J.B. (2020). Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19), *Journal of American Medical Associations*, cilt 323(18), ss. 1824-1836
- [223] Kaplan, R.M. ve Mistein, A. (2021). Influence of a COVID-19 Vaccine's Effectiveness and Safety Profile on Vaccination Acceptance, *Proceedings of the National Academy of Sciences of the United States of America*, cilt 118(10), ss. 1-5
- [224] Dey, L., Chakraborty, S. ve Mukhopadhyay, A. (2020). Machine Learning Techniques of Sequence-Based Prediction of Viral-Host Interactions Between SARS-CoV-2 and Human Proteins, *Biomedical Journal*, cilt 43(5), ss. 438-450
- [225] Barman, R.K. Saha, S. ve Das, S. (2014). Prediction of Interactions between Viral and Host Proteins Using Supervised Machine Learning Methods, *Plos One*, cilt 9(11), ss. 1-10
- [226] WenXing, X., Ni, H., QiuTing, J., Farooq, A.B.U., ZeQiang, W., YanSu, S., ChengChun, W., LiPing, W. ve GuoPing, W. (2009). Probe Binding to Host Proteins: A Cause for False Positive Signals in Viroid Detection by Tissue Hybridization, *Virus Research*, cilt 145(1), ss. 26-30
- [227] AlQuraishi, M. (2021). Machine Learning in Protein Structure Prediction, *Current Opinion in Chemical Biology*, cilt 65, ss. 1-8
- [228] Lyu Z., Wang, Z., Luo, F., Shuai, J. ve Huang, Y. (2021). Protein Secondary Structure Prediction with a Reductive Deep Learning Method, *Frontiers in Bioengineering and Biotechnology*, cilt 9, ss. 1-8
- [229] Messina, F., Giombini, E., Agrati, C., Vairo, F., Bartoli, T.A. ve diğ. (2020). COVID-19: Viral-Host Interactome Analyzed by Network Based-Approach Model to Study Pathogenesis Of SARS-CoV-2 Infection, *Journal of Translational Medicine*, cilt 18, ss. 1-10
- [230] Llabres, M. ve Valiente, G. (2020). Alignment of Virus-Host Protein-Protein Interaction Networks by Integer Linear Programming: SARS-CoV-2, *Plos One*, cilt 15(12), ss. 1-13
- [231] Liu-Wei, W., Kafkas, Ş., Chen, J., Dimonaco, N.J., Tegner, J. ve Hoehndorf, R. (2021). DeepViral: Prediction of Novel Virus-Host Interactions from Protein Sequences and Infectious Disease Phenotypes, *Bioinformatics*, cilt 37(17), ss. 2722-2729
- [232] Du, H., Chen, F., Liu, H. ve Hong, P. (2021). Network-Based Virus-Host Interaction Prediction with Application to SARS-CoV-2, *Patterns*, cilt 2(5), ss. 1-11
- [233] Ben-Hur, A. ve Noble, W.S. (2006). Choosing Negative Examples for The Prediction of Protein-Protein Interactions, *BMC Bioinformatics*, cilt 7, ss. 1-6
- [234] Striemer, G.M. ve Akoğlu, A. (2009). Sequence alignment with GPU: Performance and design challenges, *IEEE International Symposium on Parallel and Distributed Processing*, Roma, İtalya
- [235] Mandrekar, J.N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment, *Biostatistics for Clinicals*, cilt 5(9), ss. 1315-1316