

TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**SARS-COV-2 SPIKE PROTEİNİNİN EVRİMSEL YOLLARININ
DEĞİŞEBİLİRLİK HARİTALARI İLE TAHMİNİ**

YÜKSEK LİSANS TEZİ

RABİA KAŞIKCI

Biyomedikal Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Ersin Emre ÖREN

NİSAN 2025



TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

Rabia Kaşıkçı

İMZA



ÖZET

Yüksek Lisans

SARS-COV-2 SPIKE PROTEİNİNİN EVRİMSEL YOLLARININ

DEĞİŞEBİLİRLİK HARİTALARI İLE TAHMİNİ

Rabia Kaşıkçı

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Biyomedikal Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Ersin Emre Ören

Tarih: Nisan 2025

Virüsler, bulaşıcı ajanlar olarak insanlarda hastalıklara ve ölümlere yol açabilmektedir. Bu olumsuz etkileri önlemek için aşılar ve ilaçlar geliştirilmiş olsa da virüsün genetik yapısında meydana gelen mutasyonlar sebebiyle virüs, bağışıklık sisteminden kaçma ve konak hücreye daha iyi bağlanma gibi özellikler kazanarak tedavilere karşı dirençli hale gelebilir. Küresel COVID-19 pandemisi, yeni SARS-CoV-2 varyantları ortaya çıktıkça virüsün genetik materyalinde meydana gelen değişikliklerin hızlı, güvenilir ve etkili bir şekilde izlenmesi ve tahmin edilmesi gerekliliğini ortaya koymuştur. Virüsün yeni genetik yapısına uygun ilaç ve aşılardan geliştirilmesi maliyetli ve zaman alıcıdır. Bu nedenle, SARS-CoV-2'nin genomunda meydana gelebilecek mutasyonları ve bu mutasyonların etkilerini henüz gerçekleşmeden önce tahmin etmek ve gerekli önlemleri almak kritik önem taşımaktadır. Bu tezde, COVID-19 pandemisine yol açan SARS-CoV-2 virüsünün evrimsel süreci incelenmiş ve gelecekte meydana gelebilecek önemli mutasyonların tahmin edilmesi amaçlanmıştır. Bu amaç doğrultusunda, öncelikle virüsün insan hücrelerini enfekte etmesinden sorumlu Spike (S) proteini bölgesine ait gen dizileri GISAID veri tabanından temin edilmiş ve ardından bu diziler çeşitli çoklu hizalama yöntemleri (Clustal Omega, MAFFT, Toffee, vb.) ile

hizalanmıştır. Daha sonra bu hizalamalar, çeşitli skor fonksiyonları (Karlin, Sander, Valder, vb.) kullanılarak her bir amino asit için değişebilirlik skorları hesaplanmıştır. Elde edilen değişebilirlik skorları arasındaki korelasyon analiz edilerek S proteininde korunmaya ve değişmeye yatkın bölgeler belirlenmiştir. Hesaplanan değişebilirlik skorları ve deneysel mutasyon oranı kullanılarak, rastgele yürüyüş yöntemi ile gelecekte meydana gelebilecek mutasyonlar tahmin edilmiştir. Bu çalışmanın, SARS-CoV-2 virüsünün evrimsel dinamiklerinin anlaşılmasına katkı sağlaması beklenmektedir.

Anahtar Kelimeler: Sekans analizi, Skorlama fonksiyonları, Mutasyon haritaları Spike proteini.



ABSTRACT

Master of Science

FORECASTING OF THE EVOLUTIONARY PATHWAYS OF THE SARS-COV-2 SPIKE PROTEIN THROUGH VARIABILITY MAPS

Rabia Kaşıkçı

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Biomedical Engineering Science Programme

Supervisor: Dr. Ersin Emre Ören

Date: April 2025

Viruses are known as infectious agents and can cause diseases and deaths in humans. Although vaccines and drugs have been developed to prevent these effects, mutations occurring in the genetic structures of viruses may enable them to acquire characteristics such as evading the immune system and binding more effectively to host cells, thereby becoming resistant to existing treatments. The global COVID-19 pandemic has highlighted the necessity of monitoring and predicting genetic changes in the virus quickly, reliably, and effectively as new SARS-CoV-2 variants emerge. Developing new drugs and vaccines that match the virus's updated genetic content is costly and time-consuming. Therefore, predicting possible mutations in the SARS-CoV-2 genome and their potential effects before they occur, and taking the necessary precautions, is of great importance.

In this thesis, the aim is to study the evolutionary process of the SARS-CoV-2 virus, which causes COVID-19 pandemic, and to predict significant future mutations. To achieve this, genetic sequences of the spike protein region, which is responsible for the virus infecting human cells, were first obtained from the GISAID database. These sequences were then aligned using various multiple sequence alignment methods (such as Clustal Omega, MAFFT, and T-Coffee). Subsequently, these alignments were used

to calculate mutability scores for each amino acid based on different scoring functions (such as Karlin, Sander, Valdar, etc.). By examining the correlation between the obtained scores, regions of the virus that are either conserved or prone to change were identified. Using the calculated mutability scores and experimental mutation rates, potential future mutations were predicted through a random walk method. The studies conducted in this thesis are expected to contribute to the prediction of the evolutionary trajectory of the SARS-CoV-2.

Keywords: Sequence analysis, Scoring functions, Mutation maps, Spike protein.



TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Dr. Öğr. Üyesi Ersin Emre Ören'e teşekkür ederim. Biyonanotasarım Laboratuvarı çalışma arkadaşlarıma özellikle hiç yardımlarını esirgemeyen Zeynep Köker'e ve Çağlanaz Akın Gültaktı'ya teşekkür ederim. Her zaman bana destek oldukları ve yanımda olduklarını hissettirdikleri için anneme, babama ve Ömer Tahir Özütemiz'e teşekkür ederim. Yüksek lisans eğitimim boyunca sağladıkları burs imkânı için TOBB ETÜ'ye, kıymetli tecrübelerinden faydalandığım TOBB ETÜ Biyomedikal Mühendisliği Bölümü öğretim üyelerine ve destekleriyle her zaman yanımda olan aileme ve arkadaşlarıma çok teşekkür ederim. Son olarak bu tezde yer alan kısmi nümerik hesaplamaların gerçekleştirildiği TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'ne ve 2210/A Yurt İçi Genel Yüksek Lisans Burs Programı kapsamında 24 ay boyunca bana burs sağlayan TÜBİTAK'a ayrıca çok teşekkür ederim.



İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET.....	v
ABSTRACT	vii
TEŞEKKÜR	ix
İÇİNDEKİLER	xi
ŞEKİL LİSTESİ.....	xiii
ÇİZELGE LİSTESİ.....	xix
KISALTMALAR	xxi
SEMBOL LİSTESİ	xxiii
1. GİRİŞ	1
1.1 Tezin Amacı	1
1.2 Literatür Araştırması	1
1.2.1 Biyoinformatik çalışma alanı ve veri incelemesi	1
1.2.2 Genetik kod (DNA, RNA)	2
1.2.3 Mutasyon ve çeşitleri	4
1.2.4 Protein yapı ve fonksiyonu	5
1.2.5 Antibiyotik ve antiviral ilaçların keşfi ve tasarım	6
1.2.6 Virüs evrimi	8
1.2.7 SARS-CoV-2	9
1.2.8 Deneysel çalışmalar	16
1.2.8.1 Spike proteinin antiviral ilaçlara karşı direnç gelişimi	17
1.2.8.2 Spike proteinin bağışıklığa karşı direnç gelişimi	18
1.2.9 Modelleme çalışmaları	20
1.2.9.1 Protein sekans hizalaması	20
1.2.9.2 Filogenik ağaç oluşturma	21
2. MATEMATİKSEL MODELLEME VE SAYISAL YÖNTEMLER	25
2.1 SARS-CoV-2 Proteininin Evrimsel İlişkinin İncelenmesi	25
2.2 Amino asitler Arası İlişkiler ve Skorlama Matrisleri	26
2.2.1 Amino asit frekanslarının incelenmesi	27
2.2.2 Mutasyon analizleri	28
2.3 Entropi Hesabı ve Skorlama Fonksiyonları	30
2.3.1 Entropi hesaplaması	31
2.3.2 Skorlama fonksiyonların kullanımı	32
2.3.3 Metotlar arası korelasyon	36
2.4 Proteinlerin Evrimsel Değişimlerinin Tahmini	36
2.4.1 SARS-CoV-2 S proteininin mutasyon hızının hesaplanması	37
2.4.2 Rastgele yürüyüş (random walk) yöntemi	38
2.4.3 Tahminlerin değerlendirilmesi	43
3. MODELLEME SONUÇLARI VE YORUMLAR	45
3.1 SARS-CoV-2 Veri Setinin Analizi	46
3.2 Metotlar Arası Korelasyon Değerlendirmesi	54
3.3 Mutasyon Haritalarının Oluşturulması	63
3.4 Tahmin Yöntemlerinin Performans Sonuçları	65

3.5 Sonuların Doğrulanması	72
3.6 Gelecek Mutasyonların Tahmini	78
4. SONUÇ VE ÖNERİLER.....	85
KAYNAKLAR.....	89
EKLER.....	97
ÖZGEÇMİŞ.....	Hata! Yer işareti tanımlanmamış.



ŞEKİL LİSTESİ

Sayfa

- Şekil 1.1: Santral Dogma. DNA'dan RNA sentezi transkripsiyon adı verilen adımda gerçekleşir. Ardından RNA, translasyon yoluyla amino asit zincirine çevrilir. Bu zincir, doğru şekilde katlanarak işlevsel bir proteine dönüşür. 2
- Şekil 1.2: mRNA kodonları ile bunlara karşılık gelen amino asitler arasındaki ilişkiyi gösteren standart genetik kodon tablosu. Merkezden dışa doğru okunarak, üçlü nükleotid dizileri (kodonlar) protein sentezi sırasında hangi amino asidin ekleneceğini belirler. Protein sentezinin başlaması ve sona ermesi açısından önemli olan başlatıcı kodon (AUG, üçgen ile gösterilmiştir) ve durdurucu kodonlar (UAA, UAG, UGA; kare ile işaretlenmiştir) tablo üzerinde gösterilmiştir. 3
- Şekil 1.3: Antibiyotiklerin çalışma mekanizması. Bakteriyel DNA'dan RNA polimeraz aracılığıyla mRNA sentezlenir ve bu mRNA, ribozomal alt birimlerle birlikte protein sentezini başlatır. tRNA'lar amino asitleri taşıyarak polipeptit zincirini oluşturur. Ancak kloramfenikol ve tetrasiklin gibi antibiyotikler, bu süreci engelleyerek ribozomal fonksiyonları bozar. Bu müdahale, protein sentezini durdurarak bakterinin işlevsiz hale gelmesine ve ölümüne neden olur. 7
- Şekil 1.4: Koronavirüs ailesinin şeması. SARS-CoV-2, Nidovirales takımına bağlı Coronaviridae ailesi ve Coronavirinae alt ailesi altında yer alan Betacoronavirus cinsine ait soylar arasında sınıflandırılmaktadır. 9
- Şekil 1.5: Spike (S) proteininin üç boyutlu yapısı. (a) Kristal yapısı çözümlenmiş Spike protein trimeri (PDB ID: 6VXX), her biri farklı renkle (kırmızı, mavi ve sarı) gösterilmiş üç alt birimden (protomeren) oluşmaktadır ve tamamı kapalı konformasyondadır. (b) Tek bir Spike protomerinin kapalı konformasyondaki yapısı sunulmuştur. Bu yapıda, reseptör bağlanma bölgesi kırmızı, N-terminal bölge mavi, bağlayıcı (konnektör) bölge yeşil ve 614. pozisyon sarı renk ile vurgulanmıştır (Smaoui & Yahyaoui, 2021). 10
- Şekil 1.6: SARS-CoV-2'nin yapısı ve konak hücreye giriş mekanizması. Virüs yüzeyinde yer alan Spike (S), Membran (M) ve Envelop (E) proteinleri ile iç kısımdaki Nükleokapsid protein (N) ve ribonükleik asit (RNA) yapıları belirtilmiştir. Spike proteinleri, konak hücre yüzeyinde bulunan ACE-2 reseptörüne bağlanarak viral girişin gerçekleşmesini sağlar. 11
- Şekil 1.7: SARS-CoV-2 varyantların ortaya çıkış zaman çizelgesi. Kırmızı renkle gösterilen varyantlar, endişe verici varyantları (VOC), mavi renkle gösterilenler ise ilgilenilen varyantları (VOI) temsil etmektedir. 13
- Şekil 1.8: SARS-CoV-2 varyantlarına ait S proteini mutasyonlarının dağılımını gösteren ısı haritası. Kırmızı renk belirli bir mutasyonun varlığını, mavi renk ise yokluğunu göstermektedir. Alpha, Beta, Gamma, Delta ve Omicron gibi varyantlar farklı mutasyon profilleri sergilemektedir; Omicron, önceki varyantlardan önemli ölçüde ayrıldığını gösteren yüksek sayıda S proteini mutasyonuna sahiptir. 14

Şekil 1.9: SARS-CoV-2 virüsünün Spike (S) proteinine ait amino asit dizilimini içeren FASTA formatında bir örnek gösterimi. “>” işaretiyle başlayan ilk satır, sekansa dair açıklayıcı bilgileri içerirken (örneğin, virüs adı, izolasyon tarihi ve veri tabanı kimliği gibi), devamındaki satırlar ise proteini oluşturan amino asit dizisini göstermektedir.	21
Şekil 1.10: Jalview yazılımı ile oluşturulmuş filogenetik ağaç gösterimi. Farklı diziler arasındaki evrimsel ilişkiler, ağaç yapısı üzerinden görselleştirilmiştir. Dallanma noktaları, ortak ataları temsil ederken, dalların uzunlukları diziler arasındaki evrimsel mesafeyi yansıtmaktadır.	23
Şekil 2.1: Amino asitlerin mRNA üzerindeki kodonlara göre dağılımını gösteren tablo. Her bir amino asit, kendisine karşılık gelen bir veya daha fazla üçlü baz dizisi (kodon) ile şifrelenmektedir. Genetik kodun yapısı gereği, bazı amino asitler birden fazla kodon tarafından kodlanabilir.	28
Şekil 2.2: Kodon tablosundaki kodonların sayılmasıyla elde edilen amino asit frekans dağılımı. Her bir amino asit, kendisini kodlayan kodon sayısına göre temsil edilmiştir. Leucine (L), Arginine (R) ve Serine (S) gibi amino asitler çok sayıda kodon tarafından şifrelendiği için daha yüksek frekansla temsil edilmiştir.	28
Şekil 2.3: Kodon tablosu temel alınarak oluşturulmuş amino asitler arası mutasyon dönüşüm matrisi. Her hücrede yer alan sayılar, bir amino asidin başka bir amino aside kaç farklı kodon mutasyonu üzerinden dönüşebileceğini, yani kaç farklı şekilde dönüşüm gerçekleşebildiğini göstermektedir.	29
Şekil 2.4: Sekans dizilim örneği.	33
Şekil 2.5: Amino asit sayısının, bir proteinin mutasyona uğrama olasılığı üzerindeki etkisini gösteren grafik. Grafik, amino asit sayısı arttıkça mutasyon olasılığının da yükseldiğini göstermektedir.	38
Şekil 2.6: Zamanın (yıl bazında) mutasyona uğrama olasılığı üzerindeki etkisini gösteren grafik. Grafik, mutasyon olasılığının zamanla birlikte hızla arttığını ve kısa sürede maksimum seviyeye ulaştıktan sonra sabit kaldığını göstermektedir.	38
Şekil 2.7: Rastgele yürüyüş algoritması ile bir düzlemde gerçekleştirilen iki boyutlu hareketin temsili görselleştirmesi. Yeşil nokta hareketin başlangıç konumunu, kırmızı nokta ise rastgele adımlar sonucunda ulaşılan bitiş konumunu göstermektedir.	39
Şekil 2.8: Rastgele yürüyüş algoritması ile elde edilen farklı tahmin sayıları arasındaki korelasyon matrisi. Her hücrede iki farklı tahmin seti arasındaki korelasyon katsayısı gösterilmektedir. Renk skalası, korelasyon değerlerini ifade etmektedir: koyu kırmızı tonlar 1’e yakın yüksek korelasyonları, mavi tonlar ise %97’in altındaki korelasyonları temsil etmektedir.	41
Şekil 3.1: Çalışma adımları şeması.	45
Şekil 3.2: Veri seti dağılımı.	46
Şekil 3.3: COVID-19’a ait sekansların bazı ülkelere göre toplam dağılımı. Grafikte, GISAID veri tabanı üzerinden elde edilen sekans sayıları ülkelere göre sınıflandırılmıştır. ABD, İngiltere ve Hindistan gibi ülkeler yüksek sayıda sekans paylaşımı ile ön plana çıkarken; veri katkısı daha sınırlı olan ülkeler grafik sonunda yer almaktadır. Bu dağılım, genomik gözetim kapasitesi ve uluslararası veri paylaşımı açısından ülkeler arasındaki farklılıkları yansıtmaktadır.	47
Şekil 3.4: Altı farklı skorlama fonksiyonu ve dört skorlama matrisi kullanılarak yapılan analizlerde, işlenen sekans sayısının işlem süresi (gün cinsinden) üzerindeki etkisini gösteren grafik. Sekans sayısı arttıkça geçen süre doğrusal olmayan	

- şekilde artmakta olup, ikinci dereceden polinomsal bir modelle ($R^2 = 0.9994$) yüksek doğrulukla temsil edilmiştir. 48
- Şekil 3.5: SARS-CoV-2 S proteinine ait mutasyon pozisyonlarının frekans dağılımı. Üst grafikte, her bir pozisyonda gözlemlenen mutasyon sıklığı ısı haritası şeklinde görselleştirilmiştir; renk skalası yüksek frekansları (örneğin 614. pozisyon) sarıya doğru vurgulamaktadır. Alt panelde ise mutasyon frekansları pozisyon bazında sayısal olarak gösterilmiştir. Grafik, belirli pozisyonların (örneğin 614) diğerlerine göre daha sık mutasyona uğradığını ve bunların evrimsel baskı altında olabileceğini ortaya koymaktadır. 50
- Şekil 3.6: D614G mutasyonunun, farklı veri setleri arasındaki korelasyon yapısına etkisini gösteren ısı haritaları. Sol panel, D614G mutasyonu taşıyan sekanslara ait korelasyon matrisini; sağ panel ise bu mutasyonu taşımayan sekanslara ait korelasyon matrisini göstermektedir. Renk skalası, iki set arasındaki korelasyon katsayısını ifade etmektedir (1: yüksek korelasyon, 0: düşük korelasyon). 51
- Şekil 3.7: Farklı setlerde gözlemlenen S proteini mutasyonlarının dağılımını gösteren grafik. Yatay ekseninde mutasyon pozisyonları, dikey ekseninde ise veri setleri yer almaktadır. Kırmızı noktalar, ilgili mutasyonun ilgili sette gözlemlendiğini ifade etmektedir. Örneğin T95I mutasyonu tüm setlerde görülmektedir. 52
- Şekil 3.8: 2019 ve 2021 yılına ait protein sekanslarından türetilen 20 setin, benzerlik ilişkilerine göre oluşturulan ve 5 gruba ayrılan filogenetik ağacı. Y ekseninde kümeler arası uzaklıklar, kırmızı çizgide ise ayırım eşiği gösterilmiştir. 53
- Şekil 3.9: Farklı veri setlerine ait gözlemlenmiş amino asit frekanslarının karşılaştırılması. Grafik, 1, 2, 4, 6 ve 11 numaralı setlerde yer alan amino asitlerin frekans dağılımını göstermektedir. 53
- Şekil 3.10: Kullanılan skorlama fonksiyonları ve skorlama matrisleri (BLOSUM62, GONNET, PAM120, PET91) arasında hesaplanan korelasyonları gösteren ısı haritası. Grafik, her bir skorlama fonksiyonu-matris kombinasyonu ile olan ilişkisini görselleştirmektedir. Kırmızı tonlar yüksek korelasyonu, mavi tonlar ise düşük korelasyonu ifade etmektedir. 55
- Şekil 3.11: Entropi değeri ile altı farklı mutasyon skorlama fonksiyonu (BNT2, BNT3, Hogervorst, Karlin, Sander, Valdar) arasında hesaplanan korelasyon matrisi. Renk skalası, korelasyon katsayısını ifade etmektedir; kırmızı tonlar yüksek pozitif korelasyonu, mavi tonlar ise düşük korelasyonu göstermektedir. 56
- Şekil 3.12: Dört farklı skorlama matrisinin (BLOSUM62, PAM120, GONNET, PET91) benzerliklerine göre oluşturulmuş filogenik ağacı 57
- Şekil 3.13: S proteinine ait mutasyona uğrama skorlarının farklı skorlama fonksiyonlarına (BNT2, BNT3, Hogervorst, Karlin, Sander ve Valdar) göre dağılımı. Yatay ekseninde mutasyona uğrama skorları, dikey ekseninde ise bu skorların frekansı gösterilmektedir. 58
- Şekil 3.14: S proteinine ait mutasyona uğrama skorlarının küçükten büyüğe sıralanarak çizildiği karşılaştırmalı grafik. Farklı skorlama fonksiyonları (BNT2, BNT3, Hogervorst, Karlin, Sander ve Valdar) kullanılarak elde edilen skorlar pozisyon bazında sıralanmıştır. 58
- Şekil 3.15: Gözlem sayısına göre tahmin edilen pozisyon ve amino asit sayılarının değişimi. Yatay ekseninde kullanılan gözlem sayısı, dikey ekseninde ise tahmin edilen toplam pozisyon ve amino asit sayıları gösterilmektedir. 59
- Şekil 3.16: Seçilen beş farklı veri setine (Set_01, Set_02, Set_04, Set_06, Set_11) ait amino asit pozisyonlarına göre hesaplanan entropi değerlerinin dağılım grafiği. Her bir grafik, ilgili setteki amino asit pozisyonlarının çeşitliliğini ve varyasyon seviyesini yansıtmaktadır. Yüksek entropi değerleri, ilgili pozisyonda daha fazla

mutasyon çeşitliliği olduğunu göstermekte olup; bu bölgeler evrimsel baskı, fonksiyonel önem veya varyant oluşumu açısından dikkat çekici olabilir.....	61
Şekil 3.17: Seçilen beş farklı veri setine ait amino asit pozisyonlarına göre hesaplanan entropi değerlerinin birbiriyle olan korelasyon matrisi.	62
Şekil 3.18: Bütün setler için BNT3 fonksiyonu kullanılarak hesaplanan mutasyona uğrama skorlarının S proteini boyunca dağılımı. Grafik üzerinde, en yüksek varyasyon gösteren ilk 15 amino asit pozisyonu vurgulanmıştır.	64
Şekil 3.19: Set-2 için BNT3 skorlama fonksiyonu, PAM120 skorlama matrisi ve gözlemlenmiş amino asit frekanslarının kombinasyonu kullanılarak yapılan tahminlerin sonuçları.	67
Şekil 3.20: Set-4 için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahminlerin sonuçları.	68
Şekil 3.21: Set-11 için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.	68
Şekil 3.22: SARS-CoV-2 S proteininin 3B yapısı üzerinde Omicron varyantı için tespit edilen bölgelerinin görselleştirilmesi. (a) Mor renk ile gösterilen bölgeler, skor fonksiyonlarına dayalı olarak yapılan tahminler ile tespit edilen amino asit pozisyonlarını temsil etmektedir. (b) Kırmızı renk ile gösterilen alanlar ise bu pozisyonlarda tespit edilen spesifik amino asit değişimlerini göstermektedir. ..	70
Şekil 3.23: Delta varyantına ait analizlerde kullanılan veri setlerinin benzerliğine göre oluşturulan filogenik ağaç. Dikey ekseninde kümeler arası mesafe ölçütü yer almakta olup, kırmızı çizgi belirli bir eşik değeri üzerinde kümelerin ayrımını göstermektedir ve toplam dört ana kümeye ayrılmıştır.	73
Şekil 3.24: Set-3 için BNT3 skorlama fonksiyonu, PAM120 skorlama matrisi ve dönüşüm frekansı kombinasyonu kullanılarak yapılan tahminlerin sonuçları. ..	75
Şekil 3.25: Set-3 için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.	75
Şekil 3.26: Set-6 için Hogervorst skorlama fonksiyonu, PET91 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.	76
Şekil 3.27: SARS-CoV-2 S proteininin 3B yapısı üzerinde Delta varyantı için tespit edilen bölgelerinin görselleştirilmesi. (a) Mor renk ile gösterilen bölgeler, skor fonksiyonlarına dayalı olarak yapılan tahminler ile tespit edilen amino asit pozisyonlarını temsil etmektedir. (b) Kırmızı renk ile gösterilen alanlar ise bu pozisyonlarda tespit edilen spesifik amino asit değişimlerini göstermektedir. ..	77
Şekil 3.28: Gelecek tahminler için yapılan çalışmadaki mutasyon skorların dağılımı.	79
Şekil 3.29: Gelecek tahminler için yapılan çalışmadaki mutasyona uğrama skorların grafiği.	79
Şekil 3.30: Birinci grup sekanslar için BNT3 skorlama fonksiyonu, PAM120 skorlama matrisi ve gözlemlenmiş amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.	81
Şekil 3.31: Birinci grup sekanslar için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak elde edilen tahmin sonuçları.	82

Şekil 3.32: İkinci grup sekanslar için BNT3 skorlama fonksiyonu, PAM120 skorlama matrisi ve gözlemlenmiş amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.	83
Şekil 3.33: İkinci grup sekanslar için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.	84





ÇİZELGE LİSTESİ

Sayfa

Çizelge 1.1: SARS-CoV-2’de görülen mutasyonlar ve ilaçlara olan direnci.....	18
Çizelge 1.2: S Proteininde görülen mutasyonlar ve etkisi.	19
Çizelge 2.1: SARS-CoV-2 S protein sekansların 3 aylık dağılımları.	26
Çizelge 2.2: Amino asitlerin dönüşüm frekansları.....	30
Çizelge 2.3: Wuhan referans sekansına (EPI_ISL_402124) ait kodonların ve bunlardan türetilen amino asitlerin dağılımı.....	42
Çizelge 2.4: Referans sekanstan elde edilen frekans	43
Çizelge 3.1: Sekans ve zaman dağılımı.....	49
Çizelge 3.2: Sekanslardaki silinme miktarı.....	54
Çizelge 3.3: Bütün setlerin tüm hesaplamaları için en değişken 15 bölge.....	60
Çizelge 3.4: Korunan bölgelerin listesi.	60
Çizelge 3.5: Bütün setler için entropi değerlerine göre en fazla değişen 15 bölge.	62
Çizelge 3.6: Omicron varyantı için yapılan tahminlerin performans karşılaştırması.	69
Çizelge 3.7: Entropi kullanılarak yapılan mutasyon pozisyonu tahminleri.	70
Çizelge 3.8: Entropi kullanılarak yapılan amino asit mutasyonu tahminleri.	71
Çizelge 3.9: Delta varyantı için yapılan hesaplamada en değişken bölgeler.	74
Çizelge 3.10: Delta varyantı için yapılan tahminlerin performans karşılaştırması.	76
Çizelge 3.11: Gelecek için kullanılacak verinin sekans dağılımı.	78
Çizelge 3.12: Gelecek tahminler için yapılan çalışmadaki en değişken 15 bölge.	78
Çizelge 3.13: Gelecek varyantlar için en değişken 15 pozisyon.....	80



KISALTMALAR

ACE2	: Anjiyotensin dönüştürücü enzim 2
BLOSUM	: Blok Değiştirme Matrisi (Block Substitution Matrix)
DNA	: Deoksiribonükleik asit
ED	: Eşik Değeri
E	: Envelope
FDA	: Gıda ve İlaç İdaresi (Food and Drug Administration)
GISAID	: Tüm Grip Verilerinin Paylaşılmasına İlişkin Küresel Girişim
K2P	: Kimura 2-Parametresi
HIV	: İnsan Bağışıklık Yetmezliği Virüsü (Human Immunodeficiency Virus)
IC₅₀	: % 50 İnhibitör Konsantrasyonu
mRNA	: Mesajcı Ribonükleik asit
tRNA	: Haberci Ribonükleik asit
rRNA	: Ribozomal Ribonükleik asit
M	: Membran
MSA	: Çoklu Sekans Hizalama (Multiple Sequence Alignment)
N	: Nükleokapsid
NJ	: Neighbor-Joining
NMR	: Nükleer Manyetik Rezonans
ORF	: Nükleer Manyetik Rezonans Açık Okuma Çerçevesi
PAM	: Noktasal Kabul Edilen Mutasyon (Point Accepted Mutation)
PCR	: Polimeraz Zincir Reaksiyonu
PDB	: Protein Data Bankası
PET91	: PAM benzeri Evrimsel Ağaç (PAM-like Evolutionary Tree)
PIR	: Protein Bilgi Kaynağı (Protein Information Resource)
RdRp	: RNA'ya bağımlı RNA polimerazlar
RMSD	: Kare Ortalamanın Karekökündeki Sapma (Root Mean Square Deviation)
RNA	: Ribonükleik asit
SARS-COV2	: Şiddetli Akut Solunum Sendromu Koronavirüs 2
SCOP	: Proteinlerin Yapısal Sınıflandırılması
SIFT	: Toleranslıdan Toleranssızların Sınıflandırılması (Sorting Intolerant From Tolerant)
SNAP	: Kabul Edilmeyen Polimorfizmlerin Taranması (Screening for Non-Acceptable Polymorphisms)
S	: Spike
VOI	: İlgiilenilen Varyantlar (Variant of Interest)
VOC	: Endişe Verici Varyantlar (Variant of Concern)
kryo-EM	: kriyo Elektron Mikroskobu
WHO	: Dünya Sağlık Örgütü (World Health Organization)



SEMBOL LİSTESİ

Bu çalışmada kullanılmış olan simgeler açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklama
B	Mutasyon matrisi
$Corr$	Korelasyon
Cov	Kovaryasyon
d	Mesafe
f_a^{norm}	Normalize skorlama fonksiyonu
f_{BNT2}	BNT2 skorlama fonksiyonu
f_{BNT3}	BNT3 skorlama fonksiyonu
$f_{Hogervorst}$	Hogervorst skorlama fonksiyonu
f_{Sander}	Sander skorlama fonksiyonu
f_{Valdar}	Valdar skorlama fonksiyonu
H	Mutasyon hızı
$Hist$	Histogram
L	Sekans uzunluğu (Amino asit dizilimi)
m	Skorlama matrisi
M	Modifiye skorlama matrisi
$mean$	Ortalama
N	Sekans sayısı
P	Beklenen frekans
Q	Doğal frekans
s	Amino asit
$stdev$	Standart sapma
w	Ağırlık faktörü (Katsayı)



1. GİRİŞ

1.1 Tezin Amacı

Virüslerin genomlarının ayrıntılı olarak analiz edilmesi, bu organizmaların evrimsel süreçlerinin anlaşılması, hastalıklara yol açan olan türlere karşı etkili aşılar geliştirilmesi ve sebep olabilecekleri olası salgınların önüne geçilebilmesi için önemlidir. Bu tez kapsamında, 2020 yılında pandemiye sebep olan SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) virüsünün genomu üzerinde çalışılarak, gelecekte ortaya çıkabilecek mutasyonların tahmin edilmesi amaçlanmaktadır. Bu amaçla, öncelikle Spike (S) proteini üzerindeki mutasyona uğrama ihtimalleri daha yüksek olan pozisyonları belirlemek adına iki farklı yöntem kullanılmıştır. Birincisi, her pozisyon için mutasyona uğrama skorları çıkaran, pozisyona bağlı olarak geliştirilmiş skrolama fonksiyonları sonuçları ve ikinci olarak pozisyon üzerinde gerçekleşen çeşitliliğinin bilgisini veren, pozisyona özel entropi değerleri kullanılmıştır. Bu fonksiyonların sonuçları, hesaplanan farklı amino asit frekansı ve hesaplanan mutasyon hızı kullanılarak gelecekte oluşabilecek mutasyonlar tespit edilmeye çalışılmıştır. Sonuçların korelasyonları ve tahmin performansları incelenerek karşılaştırmalar yapılmış, yöntemler arasındaki farklar belirlenmiş ve en etkili tahmin yapabilen yaklaşım ortaya konulmuştur.

1.2 Literatür Araştırması

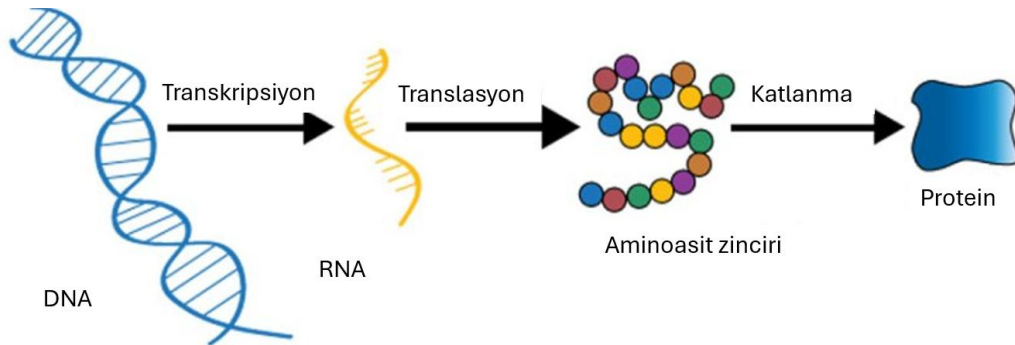
1.2.1 Biyoinformatik çalışma alanı ve veri incelemesi

İnsan genom projesinin ortaya atılması ile birlikte, biyolojik veri miktarı gün geçtikçe artmıştır ve hala artmaya devam etmektedir. Elde edilen bu büyük verinin doğru bir şekilde analiz edilmesi gerekliliği, biyoinformatik alanına olan ihtiyacı ve talebi artırmıştır. Biyoinformatik, biyolojik verilerin elde edilmesi, yorumlanması ve ilgili araçlarla uygulanması için gerekli olan bir alandır (Bayat, 2002). Biyoinformatik alanında biyolojik veriler işlenirken; bilgisayar bilimi ve istatistik gibi bilimlerden faydalanılmaktadır. Biyoinformatik ile ilgilenen uzmanlar bu alanlardan yararlanırken

veya büyük biyolojik verileri işlerken de yeni yöntemler geliştirebilmektedir. Özellikle genetik verilerin anlamlandırılması ve işlenebilir hale gelmesi için algoritmik hesaplamalar gereklidir. Biyoinformatik alanında genetik verinin çeşitliliğine ve araştırma amacına göre yeni algoritmalar ve araçlar geliştirilmektedir. (Mathura & Kanguane, 2009; Saparov & Zech, 2025). Genetik veri, DNA ve RNA dizilimlerinden oluşmakta olup, bu verilerden elde edilen protein sekansları genetik bilginin bir ifadesidir. Bu verilerin işlenmesi, gelecek virüslerin evrimi ve dolayısıyla aşı tasarımı, ilaç tasarımı gibi alanlarda önem arz etmektedir. Bu nedenle, bu alandaki ilerlemeler genetik verilerin anlamlı biçimde kullanılmasına ve bu verilerin sağlık ve bilim alanlarındaki çeşitli uygulamalarda değerli bilgiler sunmasına olanak sağlamaktadır.

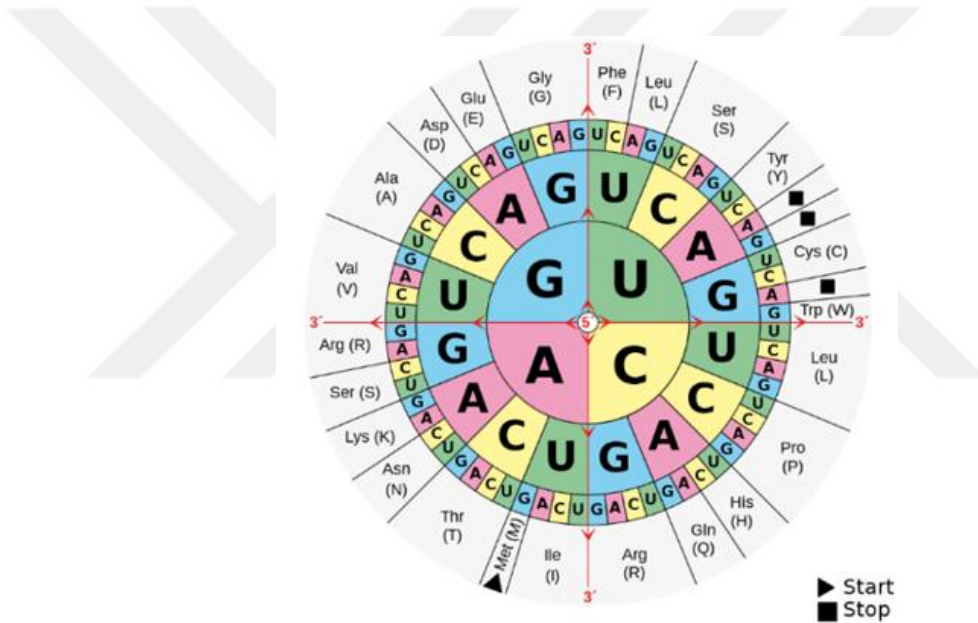
1.2.2 Genetik kod (DNA, RNA)

Genetik materyalin temel taşıyıcısı olan DNA veya RNA, nükleotit adı verilen birimlerden meydana gelir. Her bir nükleotit; bir azotlu baz, beş karbonlu bir şeker ve bir fosfat grubundan oluşur ve çift sarmallı yapılardan meydana gelmektedir. DNA'daki azotlu bazlar Adenin (A), Timin (T), Guanin (G) ve Sitozin (C) olarak adlandırılır. RNA, DNA'da olduğu gibi Adenin (A), Timin (T), Guanin (G) ve DNA'dan farklı olarak Urasil (U) bazlarından oluşur ve genetik bilginin işlenmesinde rol alan bir diğer önemli moleküldür. Haberci RNA (mRNA), taşıyıcı RNA (tRNA) ve ribozomal RNA (rRNA) olmak üzere RNA üç farklı grupta incelenir. Bu RNA türleri, genetik bilgiden protein sentezine uzanan süreçte önemli roller üstlenir. Bu biyolojik süreç 'santral dogma' (merkezi dogma) olarak adlandırılır (Mathura & Kanguane, 2009; Şekil 1.1).



Şekil 1.1: Santral Dogma. DNA'dan RNA sentezi transkripsiyon adı verilen adımda gerçekleşir. Ardından RNA, translasyon yoluyla amino asit zincirine çevrilir. Bu zincir, doğru şekilde katlanarak işlevsel bir proteine dönüşür.

Santral dogma, biyolojik sistemlerde genetik bilgi yolunu ifade eden bir terimdir. İlk olarak 1957 yılında Francis Crick tarafından ifade edilmiştir ve genel hatları ile DNA'nın RNA çeşitlerini ve RNA'ların proteinleri oluşturduğu şeklinde özetlenebilir. Genetik bilgi DNA'dan alınır, mesajcı RNA (mRNA)'ya taşınır ve sonrasında genetik bilgi ile amino asit sentezi yapılarak proteinler oluşur. Protein oluşum evresinde, mRNA'da bulunan üçlü nükleotitler ribozom tarafından amino asitlere dönüştürülür. Bu dönüştürülen üçlü nükleotitlere kodon adı verilir ve kodonlar farklı kombinasyonlar ile toplam 20 farklı amino asidi temsil eder ve bir amino asit birden fazla kodon tarafından üretilebilmektedir. Kodonların hangi amino asitlere dönüştüğünü gösteren bir kodon tablosu mevcuttur (Şekil 1.2). Genetik bilginin hangi proteini sentezleyebileceği kodon tablosu kullanılarak doğru bir şekilde çözümlenir.



Şekil 1.2: mRNA kodonları ile bunlara karşılık gelen amino asitler arasındaki ilişkiyi gösteren standart genetik kodon tablosu. Merkezden dışa doğru okunarak, üçlü nükleotid dizileri (kodonlar) protein sentezi sırasında hangi amino asidin ekleneceğini belirler. Protein sentezinin başlaması ve sona ermesi açısından önemli olan başlatıcı kodon (AUG, üçgen ile gösterilmiştir) ve durdurucu kodonlar (UAA, UAG, UGA; kare ile işaretlenmiştir) tablo üzerinde gösterilmiştir.

Kodon tablosunda görüldüğü üzere bir amino asit birden fazla kodon yolu ile sentezlenebilir. Örneğin; Lizin amino asidi 3 Adenin (AAA) ile oluşabilirken aynı zamanda 2 Adenin, 1 Guanin (AAG) ile de oluşabilir. Bu gibi örnekler sayesinde genetik çeşitlilik oluşurken, bu genetik çeşitlilik, protein sentezindeki çeşitliliğe ve dolayısıyla organizmanın adaptasyonuna da katkı sağlanmaktadır (Mathura & Kangueane, 2009).

1.2.3 Mutasyon ve çeşitleri

Mutasyonlar, organizmanın genetik materyalinde rastgele meydana gelen ve gelecek nesillere aktarılabilen değişimlerdir. Bu değişimler, organizmanın biyolojik işleyişinde aksamalara yol açabileceği gibi, bazı durumlarda virüslerde veya organizmalarda avantaj sağlayan özelliklerin ortaya çıkmasına da sebep olabilmektedir (Url-1). Mutasyonlar, çeşitli nedenlerle meydana gelebilir. DNA replikasyonu sırasında oluşan hatalar mutasyonlara yol açabileceği gibi, tek nükleotidlerin yapısını değiştiren kimyasal maddeler ve radyasyon gibi mutajen etkenler de mutasyonlara yol açabilir. Mutasyonlar meydana geldikten sonra organizmada belirgin değişikliklere yol açabileceği gibi, organizmanın hayat döngüsü için etki yaratmayabilir. Ayrıca, mutasyonların büyük bir kısmı evrim sürecine katkı sağlarken, somatik hücrelerde meydana gelen mutasyonlar gibi bir sonraki nesle aktarılmayan mutasyonların evrime doğrudan bir etkisi veya katkısı bulunmamaktadır. Mutasyon çeşitleri farklı başlıklar altında incelenebilir (İbrahim, 2021). Bunlar, yer değiştirme mutasyonları, yerleştirme mutasyonları ve silinme mutasyonlarıdır.

Yer değiştirme mutasyonları: Genetik kodda yer alan bir amino asidin, başka bir amino aside dönüşmesinden meydana gelen mutasyonlardır.

- **Sessiz mutasyonlar:** Bu tür mutasyonlarda, genetik koddaki tek değişim üretilen proteini etkilememektedir. Örneğin, AAG kodonunda meydana gelen mutasyon sonucunda AAA kodonu oluşması buna örnek verilebilir. AAG kodonu da AAA kodonu da Lizin amino asidini oluşturmaktadır.
- **Anlamsız Mutasyonlar:** Anlamsız mutasyonlarda meydana gelen mutasyon sebebi ile üretilen kodon stop kodonudur. Örneğin, AGG kodonunun mutasyonla UAG'ye dönüşmesi, Arjinin amino asidi yerine bir stop kodonu oluşmasına yol açabilir.
- **Yanlış Anlamlı Mutasyonlar:** Bu tür mutasyonlar sonucunda oluşan protein farklı bir proteine dönüşür. Örneğin, AGG Arjinin üretirken mutasyon sonucu ACG kodonu oluşur ve bunun sonucu Triptofan üretilir.

Yerleştirme mutasyonları: Bu tür mutasyonlarda genetik materyalin sıralamasına bir kodon çifti girmektedir. Örneğin, CCGA olan DNA dizi içeresine ATT girerek CCGATTA dizisi oluşmaktadır.

Silinme mutasyonları: Silinme mutasyonları, genetik dizi içerisindeki kodon veya kodonların silinmesinden kaynaklanır. Örneğin GCATTT dizi içerisinde bir

nükleotidin silinmesi sonucu GCATT oluşması bu tür mutasyonlara örnek verilebilir (Smith & Doe, 2021).

1.2.4 Protein yapı ve fonksiyonu

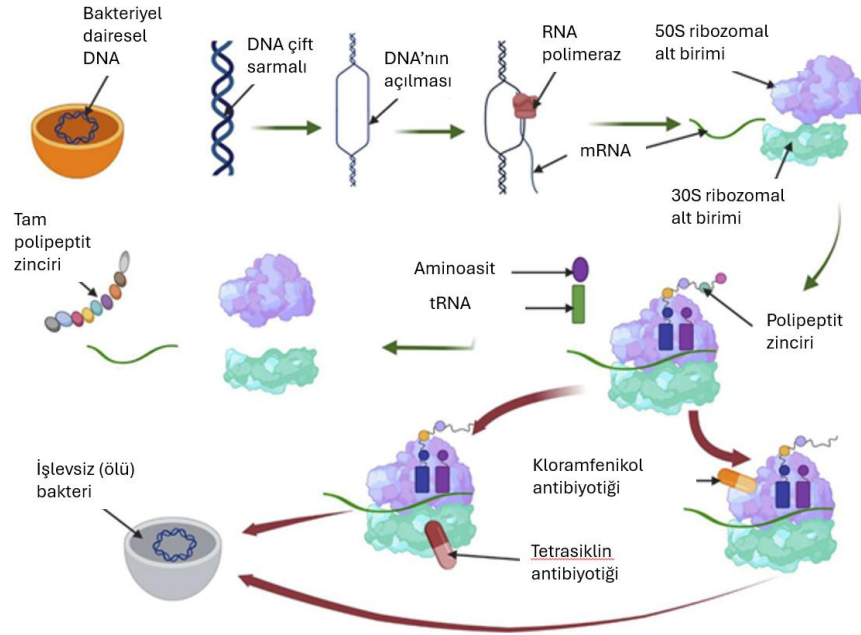
Proteinler, canlı organizmalardaki en karmaşık ve işlevsel açıdan en önemli biyomoleküllerden biridir. Önceki bölümlerde de belirtildiği üzere, proteinlerin yapı taşları amino asitlerdir ve 20 farklı amino asidin çeşitli şekillerde dizilmesiyle farklı proteinler oluşmaktadır. Amino asitler, birbirlerine peptit bağlarıyla bağlanarak farklı katlanmalar ile farklı dizilere ve konformasyonlara sahip proteinleri ve dolayısıyla farklı işlevleri üstlenen protein yapılarını oluştururlar. Proteinler peptit bağlarıyla oluştukları için sıklıkla polipeptit olarak da adlandırılmaktadır. Proteinler, çeşitli ligandlara farklı afinitelerle bağlanır ve proteinlerin biyolojik işlevi, diğer moleküllere bağlanma yeteneklerine dayanır (Berg, Tymoczko, & Stryer, 2002). Bu bağlanma, zayıf veya kuvvetli etkileşimler yoluyla gerçekleşir ve sonucunda bir özgüllük oluşarak protein belirli bir işlev için kullanılabilir. Örneğin, proteinler konak hücre reseptörlerine bağlanarak hücrel etkileşimlerde rol alabilir veya antikorlar olarak bakterilere bağlanarak bağışıklık sisteminin işleyişine katkı sağlayabilir. Proteinler, farklı afinitelere sahip olarak çeşitli ligandlara bağlanma kapasitesine sahiptir. Ligandlara bağlandığı bölgeler, belirli bir amino asit diziliminden oluşur. Buna ek olarak, protein veya ligand molekülündeki amino asit diziliminde meydana gelen herhangi bir değişiklik, bağlanma sürecini önemli ölçüde etkileyebilir (Url-2). Genel olarak protein yapısı dört farklı düzeyde incelenmektedir: birincil, ikincil, üçüncül ve dördüncül yapı. Birincil yapı, proteinlerin amino asit dizilimi ile ilgilidir ve bu seviyede meydana gelen herhangi bir değişiklik, ikincil, üçüncül ve dördüncül yapıların bütünlüğünü etkileyebilmektedir. İkincil yapı, polipeptit zincirinin omurga atomları arasındaki etkileşimler sonucu oluşan yerel katlanmış yapılardır. Bu katlanmalar, hidrojen bağları sayesinde alfa sarmallar ve beta tabakalar gibi düzenli yapılardan oluşur. Üçüncül yapı, ikincil yapıdaki katlanmaların daha kompleks hale gelmesiyle oluşur ve çeşitli kimyasal bağlarla stabilize edilir. Proteinlerin dördüncül yapısı ise, birden fazla polipeptit zincirinin bir araya gelerek üç boyutlu fonksiyonel bir yapı oluşturmasıyla meydana gelir. Proteinlerin üç boyutlu yapısının belirlenmesi, ilaç ve aşı geliştirme süreçleri açısından büyük önem taşımaktadır. Bu yapının deneysel olarak belirlenmesi için X-ışını kristalografisi, nükleer manyetik rezonans (NMR) ve kriyo-elektron mikroskobu (kriyo-EM) gibi yöntemler kullanılmaktadır. X-

ışını kristalografisi, protein molekülünün genel şekline odaklanır ve kristaldeki atomların konumlarını ve düzenini belirlemek için X-ışınlarından yararlanır. Kristale yönlendirilen bu ışınlar, farklı yoğunluklarda kırınım desenleri oluşturarak yapısal bilgilerin elde edilmesini sağlar. NMR spektroskopisi, atomik yapı hakkında bilgi sağlayan bir tekniktir ve manyetik rezonansa dayalı olarak çalışır. Farklı kimyasal grupların belirli frekanslarda rezonansa girmesiyle elde edilen sinyaller, proteinin atomik düzeyde yapısını ortaya çıkarmak için kullanılır. Kriyo-elektron mikroskobu (kriyo-EM) yöntemi ise, örneklerin kristalleşmeden donmuş amorf su fazında sabitlenmesini sağlar. Bu sayede proteinler doğal hallerine en yakın şekilde korunarak detaylı bir şekilde incelenebilir (Url-3). Protein yapılarını belirlemek için kullanılan yöntemler, hastalık araştırmaları başta olmak üzere birçok kritik alanda büyük önem taşımaktadır. Örneğin, orak hücre anemisi, kistik fibröz gibi hastalıklar, proteinin birincil yapısında meydana gelen mutasyonların yapısal seviyelerde değişimlere yol açması sonucu ortaya çıkmaktadır (Url -2). Bu sebeple proteinlerin yapılarının ve fonksiyonlarının bilinmesi önem arz etmektedir. Biyoinformatik alanında da proteinlerin dizilimleri, katlanmaları ve üç boyutlu yapısı, farklı moleküllere bağlanması gibi birçok alanın yararı için çalışılmaktadır (Y. Xu, Xu, & Liang, 2007).

1.2.5 Antibiyotik ve antiviral ilaçların keşfi ve tasarım

Virüsler, insan sağlığı için ciddi bir tehdit oluşturarak ölümcül hale gelebilmektedir. Bu nedenle, antiviral ve antibiyotik gibi ilaçların geliştirilmesi büyük önem taşımaktadır. Bazı virüsler için aşı geliştirmek oldukça zor olduğundan, bu tür enfeksiyonlarla mücadelede etkili antiviral ilaçların geliştirilmesi önemli bir gereklilik olarak öne çıkmaktadır (De Clercq, 2002). 1928 yılında Sir Alexander Fleming, antibiyotik araştırmalarının temelini atan önemli bir keşfe imza atmış ve penisilini bulmuştur (Davies & Davies, 2010). 1942 yılında ilk kez saflaştırılan penisilin, 1945'ten itibaren ise yaygın olarak kullanılmaya başlanmıştır. Özellikle 1950'ler ile 1970'ler arasındaki dönem, antibiyotik araştırmalarının altın çağı olarak kabul edilmekte olup, bu süreçte streptomisin, tetrasiklin, kloramfenikol ve eritromisin gibi pek çok antibiyotik sınıfı keşfedilmiş ve klinik kullanıma girmiştir (Adedeji, 2016; Aminov, 2010; Clardy, Fischbach & Walsh, 2006). Antibiyotikler, farklı mekanizmalarla etki göstererek mikroorganizmaların büyümesini ve çoğalmasını engellemektedir (Şekil 1.3). Örneğin, florokinolonlar grubu antibiyotikler, bakteri DNA'sının replikasyonunu durdurarak etki göstermektedir. Kloramfenikol gibi

antibiyotikler ise protein biyosentezini engelleyerek organizmaların protein üretimini hedef almaktadır. Vankomisin ise hücre duvarına saldırarak etki gösteren antibiyotikler arasında yer almaktadır. Ayrıca, trimetoprim gibi ilaçlar, folik asit metabolizmasını inhibe ederek DNA sentezi için gerekli olan folik asit üretimini engeller. Bu çeşitli etki mekanizmaları sayesinde farklı antibiyotikler geliştirilmiş ve günümüzde de yeni antibiyotiklerin keşfedilmesine yönelik çalışmalar devam etmektedir (Halawa vd., 2024; Jangra vd., 2025).



Şekil 1.3: Antibiyotiklerin çalışma mekanizması. Bakteriyel DNA'dan RNA polimeraz aracılığıyla mRNA sentezlenir ve bu mRNA, ribozomal alt birimlerle birlikte protein sentezini başlatır. tRNA'lar amino asitleri taşıyarak polipeptit zincirini oluşturur. Ancak kloramfenikol ve tetrasiklin gibi antibiyotikler, bu süreci engelleyerek ribozomal fonksiyonları bozar. Bu müdahale, protein sentezini durdurarak bakterinin işlevsiz hale gelmesine ve ölümüne neden olur.

Antiviral ilaçların çalışma mekanizması, antibiyotiklerden farklıdır. Antibiyotikler bakterileri doğrudan öldürürken, antiviraller genellikle virüsleri tamamen yok etmek yerine onların çoğalmasını engellemeye odaklanır (Kausar vd., 2021). Virüsler, konak hücrenin mekanizmasını kullanarak çoğaldıkları için, onları hedef alan ilaçların geliştirilmesi oldukça zordur. İlk antiviral ilaç olan idoksuridin, 1963 yılında onaylanmış olup, bu tarihten itibaren antiviral ilaçlar sürekli olarak geliştirilmekte ve çeşitlendirilmektedir (Bule vd., 2019). Antiviral ilaçların çalışma mekanizmaları şu şekildedir. Özellikle SARS-CoV-2 ve diğer koronavirüslere karşı geliştirilen Remdesivir ve türevleri, virüsün RNA polimerazını hedef alarak replikasyonunu engeller. Ritonavir, Lopinavir grubu ise proteaz inhibitörleri olarak çalışırlar (Cao vd.,

2020). Virüsün konak hücreye girişini engelleyen bir çalışma mekanizmasına sahip antiviral ilaç grubu da vardır (Hoffmann vd., 2020). İmmün modülatörlerin ait olduğu grup ise vücudun bağışıklık yanıtını güçlendirerek virüs replikasyonunu baskılayarak çalışır. Nükleozit ve Nükleotid Analogları grubu ise virüsün genetik materyalinin sentezini durdurarak çalışan bir gruptur (Kausar vd., 2021; Aiello vd., 2022).

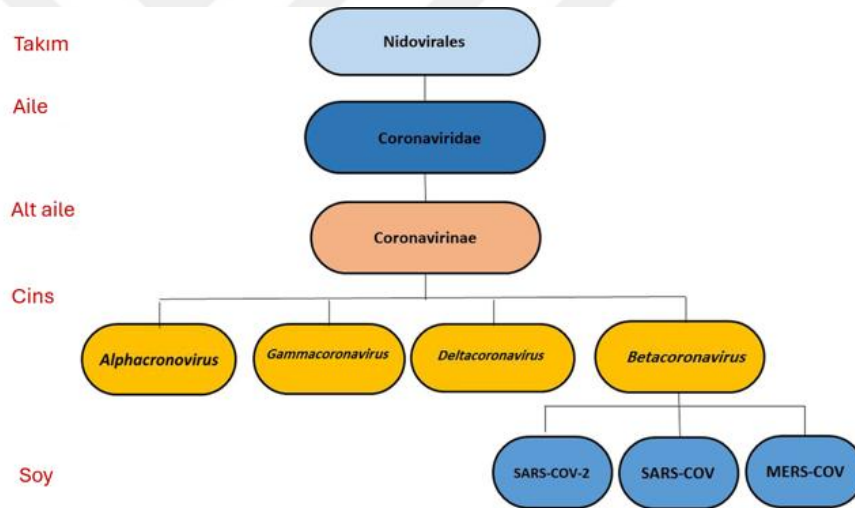
1.2.6 Virüs evrimi

Virüslerin ortaya çıkması ve gelişmesi ile ilgili farklı teoriler vardır. Bir görüşe göre, virüsler hücre içi parazitken, zamanla bağımsız çoğalma yeteneği kazanmışlardır. Bir diğer teoriye, virüslerin hücrelerden kaçan genetik yapılarının kendi kendilerine kopyalanmasıyla oluştuğunu öne sürmektedir. Son olarak, bazı araştırmalar virüslerin ilkel ve replikasyonla çoğalabilen moleküllerden evrimleştiğini savunmaktadır. Evrim süreçleri hala devam eden virüsler, farklı mekanizmalarla değişim geçirebilmektedir. Bu evrimsel süreçler; konak hücre değişimi, hastalık yapma gücü (virülans) seviyesinin değişimi, ilaçlara dirençli türlerin ortaya çıkışı ve bağışıklık sisteminden kaçış mekanizmaları gibi çeşitli amaçlar ile gerçekleşmektedir (Payne, 2017).

Virüsler, yeniden düzenleme, rekombinasyon ve mutasyon yoluyla genetik çeşitlilik kazanabilirler. Mutasyon oranları ise virüs türüne bağlı olarak büyük farklılıklar gösterebilir. Örneğin, HIV gibi bazı virüslerde her eklenen nükleotid başına mutasyon oranı 10^{-3} ile 10^{-4} arasında değişebilirken, Herpesvirüsler gibi daha düşük mutasyon hızına sahip virüslerde bu oran 10^{-8} ile 10^{-11} aralığında görülebilir (Cann, 2011). DNA veya RNA taşıyan yapılar olarak sınıflandırılan virüsler arasında, RNA virüslerinin mutasyon oranı daha yüksektir. Bunun nedenleri ise, RNA'ya bağımlı RNA polimerazlar (RdRp), DNA'ya bağımlı DNA polimerazlara kıyasla daha yüksek hata oranlarına sahip olmasıdır ve ayrıca DNA virüslerinin aksine, RNA virüslerinde genetik materyalin kopyalanması sırasında hata düzeltme mekanizmalarının bulunmamasıdır (Lauring, Frydman, & Andino, 2013). Bu durum, SARS-CoV-2 gibi RNA virüslerinin hızlı evrim geçirmesine ve genetik çeşitliliğin artmasına yol açmaktadır (LaTourrette & Garcia-Ruiz, 2022). Mutasyon hızı, virüslerin evrimsel süreçte ilerleyişini belirleyen en önemli faktörlerden biridir. Çünkü bir virüs evrim geçirdikçe yeni özellikler kazanır; örneğin, ilaç direnci geliştirebilir veya bağışıklık sisteminden kaçma yeteneğini artırabilir. Bu değişimler, virüsün yayılma kapasitesini ve hastalık yapma potansiyelini doğrudan etkileyerek, insan sağlığı açısından önemli sonuçlar doğurabilir.

1.2.7 SARS-CoV-2

Koronavirüsler, memeli canlılar ile kuşları da enfekte edebilen tek sarmallı pozitif polariteli RNA virüsleridir. Bu virüsler, Nidovirales takımına, Coronaviridae familyasına ve Orthocoronavirinae alt familyasına aittir. Orthocoronavirinae alt familyası, Alphacoronavirüsler, Betacoronavirüsler, Gammacoronavirüsler ve Deltacoronavirüsler olmak üzere dört grupta incelenmektedir. Alphacoronavirüsler ve Betacoronavirüsler genellikle memelileri enfekte ederken, Gammacoronavirüsler ve Deltacoronavirüsler ise kuşları da içeren geniş bir konakçı çeşitliliğine sahiptir (Şekil 1.4) (Mingaleeva vd., 2022; Saberian vd., 2022). Avian enfeksiyöz bronşit virüsü (IBV) ise tarihte keşfedilen ilk koronavirüstür. İnsanları enfekte eden koronavirüsler ise 1966-1967 yıllarında keşfedilmiştir. Daha sonraki yıllarda, HCoV-NL63 (Hollanda, 2004) ve HCoV-HKU1 (Hong Kong, 2004) gibi virüs türleri de ortaya çıkmış ve tanımlanmıştır.



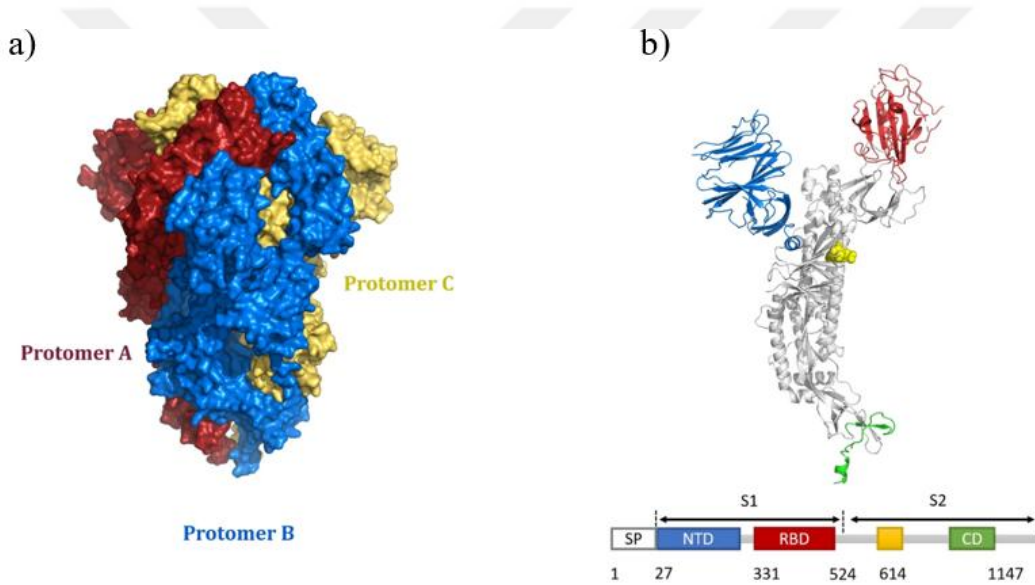
Şekil 1.4: Koronavirüs ailesinin şeması. SARS-CoV-2, Nidovirales takımına bağlı Coronaviridae ailesi ve Coronavirinae alt ailesi altında yer alan Betacoronavirus cinsine ait soylar arasında sınıflandırılmaktadır.

2012'de Suudi Arabistan'da tespit edilen orta doğu solunum sendromu MERS-CoV, yüzde 35'e varan ölüm oranına sebep olduğu rapor edilmiştir (Url-4). Koronavirüslerin farklı türler arasında geçiş yapabildiği ve bu geçişler sonucunda yeni koronavirüs türlerinin ortaya çıktığı birçok vaka rapor edilmiştir. Bu duruma örnek olarak, 2019 yılında Wuhan, Çin'de keşfedilen ve küresel bir pandemiye sebebiyet veren SARS-CoV-2 virüsü gösterilebilir. SARS-CoV-2'nin ortaya çıkmasında türler arası geçiş etkisi çok büyüktür (Crits-Christoph vd., 2024). SARS-CoV-2'nin genom yapısı ise

diğer SARS-CoV'lar ile yaklaşık %80 oranında benzerken yarası koronavirüslerine de yüksek benzerlik gösterdiği görülmüştür (Saberian vd., 2022).

1.2.7.1 SARS-CoV-2 yapısı

SARS-CoV-2, Beta-coronavirus cinsi tek sarmallı pozitif anlamlı olan bir RNA virüsüdür. Dizileme işlemlerinin sonucu olarak, SARS-CoV-2 genomu yaklaşık olarak 29,9 kilobaz (kb) yani 30.000 nükleotid uzunluğunda olduğu rapor edilmiştir (Lu vd., 2020). Bu bilgilere ek olarak SARS-CoV-2 genomu 12'si fonksiyonel olan toplam 13-15 açık okuma çerçevesinden (ORF) oluşmaktadır (Naqvi vd., 2020). SARS-CoV-2 yapısal olarak, 4 farklı kısımdan oluşur. Bu yapısal kısımlar Spike (S)(Şekil 1.5), Zarf (Envelope, E), Zar (Membran, M) ve Nükleokapsid (N) proteinleridir.



Şekil 1.5: Spike (S) proteininin üç boyutlu yapısı. (a) Kristal yapısı çözümlenmiş Spike protein trimeri (PDB ID: 6VXX), her biri farklı renkle (kırmızı, mavi ve sarı) gösterilmiş üç alt birimden (protomerdan) oluşmaktadır ve tamamı kapalı konformasyondadır. (b) Tek bir Spike protomerinin kapalı konformasyondaki yapısı sunulmuştur. Bu yapıda, reseptör bağlanma bölgesi kırmızı, N-terminal bölge mavi, bağlayıcı (konnektör) bölge yeşil ve 614. pozisyon sarı renk ile vurgulanmıştır (Smaoui & Yahyaoui, 2021).

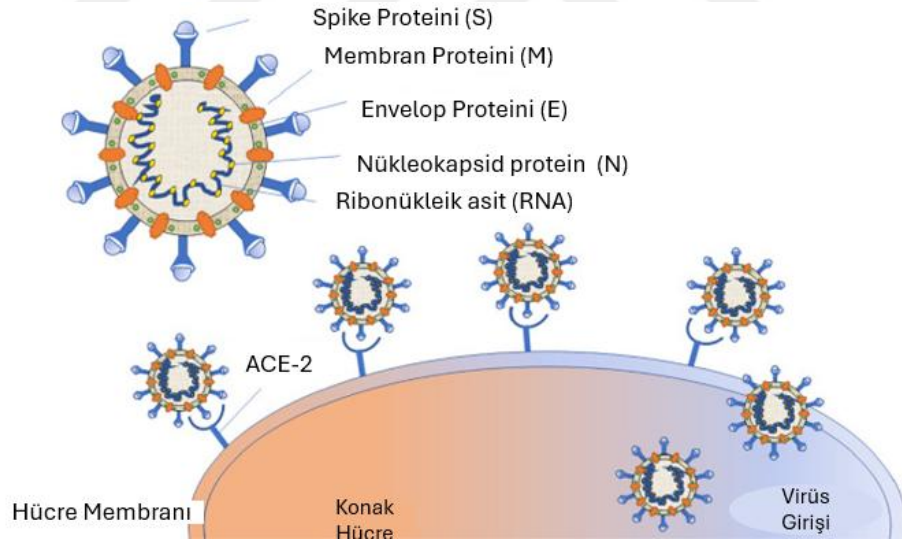
Spike (S) Proteini: Konakçı hücreye bağlanmada görev alan S proteini, 1273 amino asitten oluşur. S1 ve S2 olmak üzere 2 alt bölümden oluşur ve bu S proteini virüs ile bağlanma süresince farklı şekillerde hareket edebilir. Yapısal olarak S1 (1-686) bağlanma alanı ve S2 füzyon alanı (687–1273) içermektedir. 1 ile 686 residue arasını kapsayan S1 bağlanma kısmı, kapalı pozisyondan açık pozisyona gelerek, SARS-CoV-2 virüsünün, konak hücrede bulunan ACE-2 (Angiotensin Converting Enzyme

2) reseptörüne bağlanır. S proteininin 687–1273 arasındaki S2 bölgesi, virüsün konak hücre reseptörüne bağlanmasının ardından, viral ve hücre zarlarının kaynaşmasını sağlayarak virüsün genetik materyalinin hücre içerisine aktarılmasında rol almaktadır. (Şekil 1.5) (Tao vd., 2021; Naqvi vd., 2020).

Zarf (Envelope, E): Zarf proteini, küçük viral proteinlerden oluşan bir yapıdır ve virionların birleşmesine ve salınmasına yardımcı olur. Bu protein, virüsün konak hücrelerle etkileşimini ve enfeksiyon sürecini başlatmasını sağlar.

Zar (Membran, M): M proteinleri, 222 amino asitten oluşan ve RNA paketlenmesinde görev alan bir yapısal proteindir. Bu protein, virionun yapısal bileşenlerinden biridir ve viral partikülün olgunlaşmasında önemli bir rol oynar.

Nükleokapsid (N): Viral RNA transkripsiyonunun ve replikasyonunun arttırılmasına yardımcı olan nükleokapsid proteini, viral genom ile M proteini arasında etkileşime girerek viral birleşmeyi sağlamaktadır. Bu şekilde, virüsün genetik materyalinin paketlenmesinde ve bulaşma yeteneğinin arttırılmasında önemlidir (Tao vd., 2021; Naqvi vd., 2020) (Şekil 1.6).



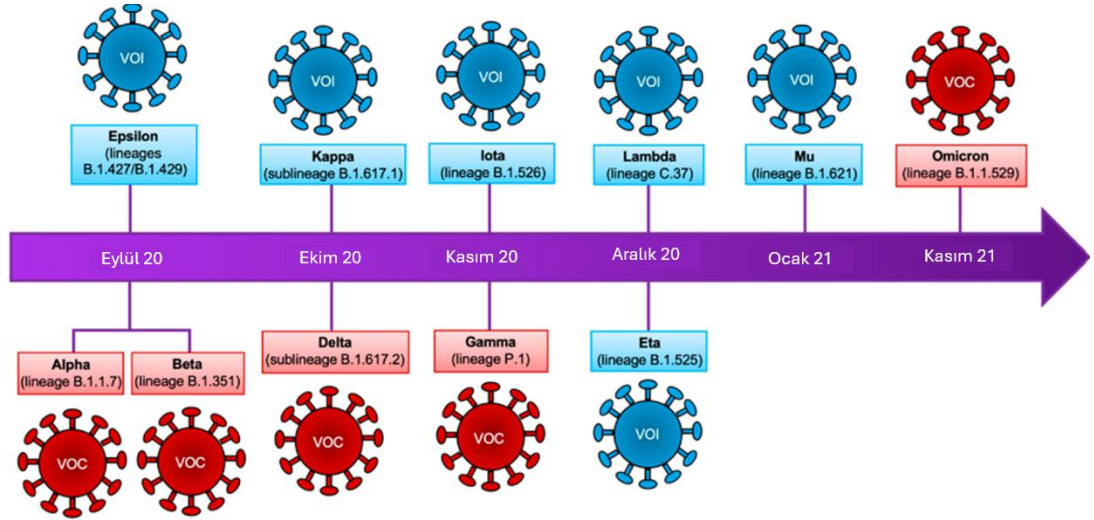
Şekil 1.6: SARS-CoV-2'nin yapısı ve konak hücreye giriş mekanizması. Virüs yüzeyinde yer alan Spike (S), Membran (M) ve Envelop (E) proteinleri ile iç kısımdaki Nükleokapsid protein (N) ve ribonükleik asit (RNA) yapıları belirtilmiştir. Spike proteinleri, konak hücre yüzeyinde bulunan ACE-2 reseptörüne bağlanarak viral girişin gerçekleşmesini sağlar.

1.2.7.1 Evrimi ve varyantları

Birçok RNA virüsü gibi, koronavirüsler hızlı bir şekilde değişebilirler, evrimsel süreçleri aylar veya yıllar sürebilir ve bu süreçler gözlemlenebilir. Virüs evrimi, bulaşma olayları, bağışıklık grafikleri ve insan hareketleri gibi çeşitli ekolojik

dinamiklerle birlikte gerçekleşir. Bu evrimsel süreçte, mutasyonların üretilmesi ve yeni genetik materyalin yayılma hızı önemli bir rol oynar (Markov vd., 2023; Url-5). Koronavirüslerin çeşitliliği ve evrimi birkaç faktöre bağlıdır. Mutasyon oranı, virüsler için evrimin temelini oluşturur. RNA virüsü olması, konak hücrede hayatta kalma avantajları gibi sebepler, mutasyon oranını etkiler. Evrimsel değişim oranı ve rekombinasyon, virüsün evrimini büyük ölçüde etkileyen faktörlerdir. Bu faktörlerin etkisiyle SARS-CoV-2'de farklı mutasyonlar ortaya çıkmaktadır. Bu mutasyonlar, virüsün farklı varyantlarını oluşturarak çeşitlilik yaratmaktadır (Guruprasad vd., 2023; Akkız, 2022). Varyant ifadesi, bir canlı organizmanın genetik kodunda meydana gelen birden çok mutasyon sonucunda ortaya çıkan yeni bir formu tanımlar. SARS-CoV-2 virüsü de bir dizi mutasyona uğramıştır ve bu mutasyonlar bir araya gelerek çeşitli varyantların ortaya çıkmasına neden olmuştur. Bu varyantlar, virüsün immün sistemden kaçma veya ilaçlara karşı direnç geliştirme gibi özellikler kazanmasına sebebiyet verir. SARS-CoV-2 virüsü, ortaya çıktığı tarihten itibaren farklı mutasyonlar geçirerek birçok varyantı oluşturmuştur. WHO (Dünya Sağlık Örgütü), SARS-CoV-2 varyantlarını sınıflandırmak için iki ana kategori oluşturmuştur: Endişe Verici Varyantlar (VOC- Variant of Concern) ve İlgilenilen Varyantlar (VOI- Variant of Interest). Endişe verici varyantlar daha yüksek bulaşıcılık daha şiddetli hastalık ve mevcut tedavilere direnç göstermek gibi özelliklere sahip olan varyantlardır. İlgilenilen varyantlar genellikle genetik özellikleri, bulaşıcılığı veya hastalık şiddeti gibi özellikler nedeniyle dikkat çekmektedir. Alfa, Beta, Gama, Delta ve Omicron endişe verici (VOC) olarak gruplandırılırken Lambda ve Mu varyantları ilgilenilen varyantlar (VOI) olarak sınıflandırılabilir (Şekil 1.7) (Flores-Vega vd., 2022; Shah & Woo, 2022; Guruprasad vd., 2023; Url-4).

Alfa varyantı: Eylül 2020'de Birleşik Krallıkta ortaya çıkan ilk varyanttır. Bu varyant S proteininde toplam 8 farklı mutasyon ($\Delta 69-70$ silinme, $\Delta 144$ silinme, N501Y, A570D, P681H, T716I, S982A ve D1118H) olmak üzere genomunda toplam 17 farklı mutasyon bulundurmaktadır. Bu mutasyonlardan bazıları S proteinin daha iyi bağlanmasını sağlarken bazıları da virüsün, immün sistemden kaçmasına sebep olmuştur. Örneğin, N501Y mutasyonu virüsün, konak hücrede bulunan ACE-2 reseptörüne bağlanma afinitesini arttırmıştır. H69/V70 silinmesi ise virüsün immün sistemden kaçmasına katkı sağlamıştır (Flores-Vega vd., 2022; Choi & Smith, 2021).



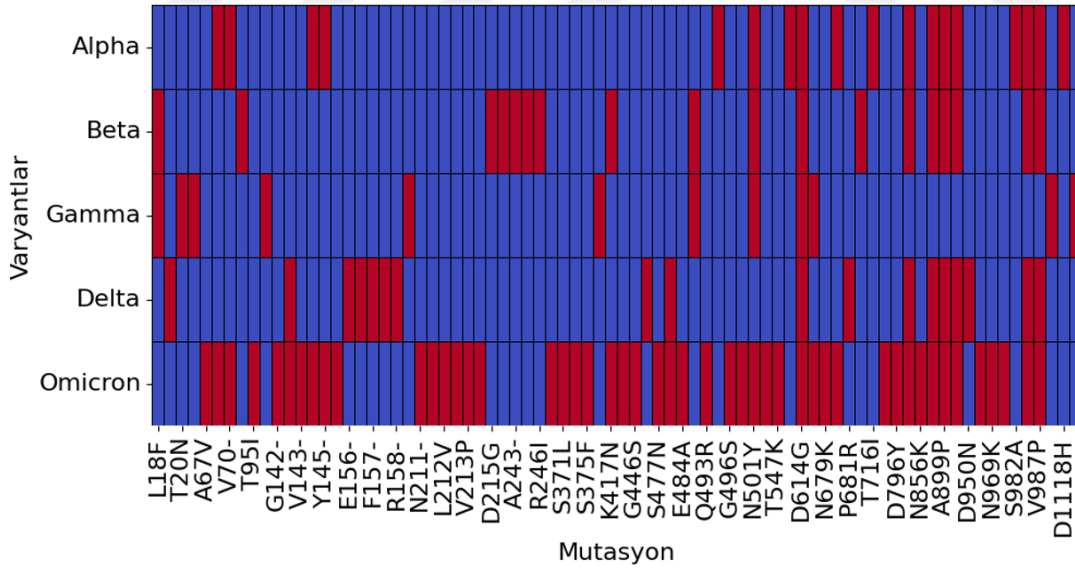
Şekil 1.7: SARS-CoV-2 varyantlarının ortaya çıkış zaman çizelgesi. Kırmızı renkle gösterilen varyantlar, endişe verici varyantları (VOC), mavi renkle gösterilenler ise ilgilenilen varyantları (VOI) temsil etmektedir.

Beta varyantı: Güney Afrika'da Eylül 2020'de ortaya çıkan Beta varyantı, S proteinini üzerinde 9 farklı mutasyon barındırmaktadır. Bu mutasyonlar arasında L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G ve A701V bulunmaktadır. Bu genetik değişiklikler, virüsün konak hücreye daha etkili bir şekilde bağlanmasına ve bulaşıcılık oranının artmasına neden olmuştur. Özellikle D614G mutasyonu, virüsün bulaşıcılığını arttırarak baskın hale gelmiş ve bundan sonraki tüm varyantlarda yerini korumuştur (Jangra vd., 2021; Şekil 1.8).

Gamma Varyantı: Gamma varyantı, Kasım 2020'de Japonya ve Brezilya'da ortaya çıkan bir SARS-CoV-2 varyantıdır. Bu varyant, S proteinini üzerindeki 10 farklı mutasyon (L18F, T20N, P26S, D138Y, R190S, H655Y, T1027I, V1176, K417T, E484K ve N501Y) bulundurmaktadır. Bu mutasyonlar sebebiyle virüsün bulaşılabilirliği ve konak hücreyi yeniden enfekte etme oranı artmıştır. Özellikle, yapılan çalışmalar Gamma varyantının, monoklonal antikor tedavisinin etkisini önemli ölçüde azalttığı tespit edilmiştir (U.S. Food and Drug Administration, 2021).

Delta Varyantı: Delta Varyantı, Ekim 2020'de Hindistan'da tanımlanan bir SARS-CoV-2 varyantıdır. S proteininde 11 farklı mutasyon içermektedir. Bunlar sırası ile T19R, T95I, G142D, 156del, 157del, R158G, L452R, T478K, D614G, P681R ve D950N mutasyonlarıdır. Delta varyantı, aşılınmış bireylerde dahi yüksek bulaşma kapasitesi göstermesiyle öne çıkmaktadır. Bu etkiyi destekleyen önemli bir mutasyon, L452R mutasyonudur. Bu mutasyonun virüsün konak hücrelere bağlanma afinitesini artırdığı yapılan çalışmalarda gözlemlenmiştir (Deng vd., 2021).

Omicron Varyantı: Omicron varyantı, Kasım 2021’de Botsvana, Afrika’da tanımlanmıştır. S proteininde 30’dan fazla mutasyon içermektedir (Zhang vd., 2022). Bu mutasyonlar sebebi ile diğer virüslere kıyasla yüksek bulaşıcılık, bağışıklık direncinde artma gibi özellikler göstermiştir (Flores-Vega vd., 2022). Omicron’un sahip olduğu yüksek sayıda mutasyon ve önceki varyantlardan belirgin şekilde farklılık göstermesi, kökeninin yalnızca insan kaynaklı olmayabileceğini düşündürmektedir. Yapılan bir çalışmada, Omicron varyantında görülen mutasyonların insan hücre ortamında evrimleşen varyantlardaki mutasyonlarla karşılaştırıldığında, fare hücre ortamında gözlemlenen mutasyon profilleriyle yüksek düzeyde benzerlik gösterdiği bildirilmiştir. Bu bulgular, SARS-CoV-2’nin insanlardan farelere bulaştığını, burada evrimsel olarak değişime uğradıktan sonra tekrar insanlara geçtiğini öne sürmektedir (Wei vd., 2021). Bir diğer çalışmada ise, Omicron alt varyantlarının yalnızca insanlardaki ACE-2 reseptörüne değil, aynı zamanda fare, hamster gibi bazı evcil ve yabani memelilerdeki ACE-2 reseptörlerine de yüksek bağlanma kapasitesi gösterdiği ortaya konmuştur (Zhao vd., 2023). Bu durum, Omicron’un türler arası geçiş yapabilme potansiyeline sahip olduğunu ve evrimsel sürecinde birden fazla konak türünün rol oynamış olabileceğini desteklemektedir.



Şekil 1.8: SARS-CoV-2 varyantlarına ait S proteini mutasyonlarının dağılımını gösteren ısı haritası. Kırmızı renk belirli bir mutasyonun varlığını, mavi renk ise yokluğunu göstermektedir. Alpha, Beta, Gamma, Delta ve Omicron gibi varyantlar farklı mutasyon profilleri sergilemektedir; Omicron, önceki varyantlardan önemli ölçüde ayrıldığını gösteren yüksek sayıda S proteini mutasyonuna sahiptir.

Varyantlarda gözlemlenen mutasyonların ısı haritası Şekil 1.8’de gösterilmiştir. Burada da gösterildiği gibi Omicron varyantı diğer varyantlardan daha fazla mutasyona sahipken her varyant kendine ait mutasyonlar içermektedir (Url 6).

1.2.7.2 SARS-CoV-2 için ilaç geliştirilmesi

SARS-CoV-2'nin küresel bir pandemiye yol açması, ilaç geliştirme çalışmalarını acil bir gereklilik haline getirmiştir. Geliştirilen ilaçlar, farklı çalışma mekanizmalarına sahip olup, virüsün yayılmasını ve etkisini azaltmaya yönelik çeşitli yöntemlerle tasarlanmaya çalışılmıştır. SARS-CoV-2'nin tedavisinde kullanılan ilaçlardan biri olan Remdesivir, daha önce Ebola gibi farklı virüslerin tedavisinde de kullanılmıştır. Ayrıca, Remdesivir SARS-CoV-2'nin bir hayvan modeli üzerinde yapılan çalışmada, SARS-CoV-2 üzerinde etkinliği gösterilen ilk antiviral ilaç olarak bilinmektedir. Remdesivir 'in çalışma mekanizması, viral genom içine entegre olarak, yeni oluşacak viral RNA zincirlerinin erken sonlanmasını sağlaması üzerine kuruludur. Bu sayede virüsün çoğalması engellenerek enfeksiyonun yayılımı kontrol altına alınmaya çalışılmaktadır. Favipiravir, RNA sentezini durdurmaya yönelik bir mekanizmaya sahiptir. Nükleotid analogu olarak işlev görerek, sahte bir nükleotid gibi tanınır ve viral RNA sentezi sırasında genoma dahil edilir. Bu süreç, erken zincir sonlandırmasına neden olarak viral replikasyonu engeller ve virüsün yayılmasını önlemeye yardımcı olur. Bu ilaçlar, SARS-CoV-2'nin ACE-2 reseptörüne bağlanmasını engelleyerek virüsün hücreye girişini önlemeye yönelik bir mekanizma ile çalışmaktadır (Aboul-Fotouh vd., 2023). İlaç tasarımlarına ek olarak, pasif immün terapilerden biri olan antikor bazlı tedaviler, SARS-CoV-2 hastalığının önlenmesi ve tedavi edilmesi için araştırılıp geliştirilmeye çalışılmıştır. Bu yöntemlerden biri olan İmmün Plazma Tedavisi, hastalığı geçirmiş bireylerden alınan plazmanın, enfekte olmuş bireylere enjekte edilmesiyle uygulanmaktadır. Bu yaklaşım hala araştırılmakta ve tartışılmakta olsa da bu tedaviyi alan bireylerin, yalnızca standart bakım gören hastalara kıyasla daha yüksek iyileşme oranlarına sahip olduğu ortaya konulmuştur (Klassen vd., 2021). Monoklonal antikor tedavisi ise hastalığa karşı üretilen antikorların insanlara enjekte edilmesi ile ortaya çıkan bir tedavi yöntemidir. Hastalığı geçirmiş, iyileşmiş bireylerden elde edilebileceği gibi laboratuvar ortamında da geliştirilebilmektedir (Phan vd., 2024). İlaçlar ve monoklonal antikor tedavilerinin yanı sıra, SARS-CoV-2 enfeksiyonunun önlenmesi amacıyla aşı geliştirme çalışmaları da hızla yürütülmüştür. Aşı çalışmaları ise 4 farklı grupta incelenebilir. Bunlar Nükleik

asit tabanlı aşı, viral vektör/adenovirüs aşıları, protein bazlı aşılar, tam inaktif (ölü) virüs aşıları. Nükleik asit tabanlı aşılar da SARS-CoV-2 virüsünün S proteininin genetik bilgileri iletilerek vücudun bağışıklık sisteminin bu proteine karşı antikor üretmesi teşvik edilir. BNT162b2 (Pfizer-BioNTech aşısı) ve Spikevax (Moderna aşısı), aşıları bu gruba girmektedir. Viral Vektör/Adenovirüs aşılar da zararsız bir adenovirüs, vektör olarak kullanılır ve SARS-CoV-2'ye ait genetik materyali taşımak için bu vektör modifiye edilir. JCOVDEN (Johnson & Johnson), Vaxzevria (AstraZeneca) ve Sputnik V aşıları bu gruba girmektedir. Protein tabanlı aşılar da ise, SARS-CoV-2'nin kendisi kullanılmaktadır ve bu protein bir kişinin bağışıklık sistemi tarafından algılandığında bağışıklık tepkisi oluşmaya başlar. Acil kullanım onayı almış tek protein bazlı aşı ise Nuvaxovid'dir. Tam inaktif virüs tabanlı aşılar ise, öldürülmüş ya da etkisiz hale getirilmiş (inaktif edilmiş) virüsleri içerir. Bağışıklık hücreleri, bu etkisiz hale getirilmiş virüsle karşılaştığında bağışıklık yanıtı geliştirerek hastalığa karşı koruma sağlar. BBIBP-CorV, CoronaVac (Sinovac) ve Covaxin, bu mekanizma ile çalışan aşılardır (Awadasseid vd., 2021; Niknam vd., 2022; Batool vd., 2025).

1.2.8 Deneysel çalışmalar

SARS-CoV-2 virüsünün yapısının ve enfeksiyon mekanizmasının anlaşılması, etkili ilaç ve aşı geliştirme çalışmalarının hız kazanmasına ve kesintisiz ilerlemesine öncülük etmiştir. Bu doğrultuda, viral proteinlerin sekansları ve yapısal modellemeleri üzerinde gerçekleştirilen detaylı analizler büyük önem taşımaktadır. Yapılan bu çalışmalar, bir yandan ilaçların hedef proteinler üzerindeki etki mekanizmalarının aydınlatılmasına katkı sağlarken, diğer yandan virüslerin yapısal ve genetik özelliklerinin kapsamlı bir şekilde ortaya konulmasına imkan tanımaktadır. Bazı araştırmalar doğrudan antiviral ajanların bağlanma bölgelerinin tanımlanmasına veya ilaç direncine yol açabilecek mutasyonların belirlenmesine odaklanırken; bazıları ise virüslerin yapısal ve genetik çeşitliliğini derinlemesine analiz ederek evrimsel süreçlerinin anlaşılmasını hedeflemektedir. Örneğin, S proteininin RBD (Receptor Binding Domain) bölgesinde bulunan 201 amino asit pozisyonunda gerçekleştirilen sistematik bir çalışmada, olası 3.819 tekli amino asit değişimi tek tek deneysel olarak sentezlenmiş ve oluşturulan varyantların 3 boyutlu yapılarındaki değişimlerin ACE2 reseptörüne bağlanma kapasiteleri üzerindeki etkileri detaylı bir şekilde analiz edilmiştir. Bu tez çalışmasında ise, yalnızca mevcut varyantlar ile sınırlı kalınmayarak, S proteini üzerinde oluşabilecek potansiyel mutasyonlar hesaplamalı biyoloji

kullanılarak geniş kapsamlı bir biçimde değerlendirilmiştir. Burada uygulanan yöntemler, deneysel çalışmalardan farklı olarak, çok daha büyük bir mutasyonel uzayın elde edilip incelenmesine ve gelecekte oluşabilecek varyasyonların sistematik olarak öngörülmesine olanak sağlamaktadır.

1.2.8.1 Spike proteinin antiviral ilaçlara karşı direnç gelişimi

S proteini, RNA virüsü olması, yüksek mutasyon hızına ve hızlı çoğalma yeteneğine sahip olması nedeniyle sıkça mutasyona uğrayarak farklı özellikler kazanmaktadır. Mutasyonlar sayesinde kazanılan farklı özelliklerden biri olan ilaçlara karşı direnç kazanma yetisi tedavi yöntemi geliştirilmesinin önüne geçebilmektedir. Günümüzde direnç testleri için 2 ana yöntem kullanılmaktadır, genotipik antiviral direnç testi ve fenotipik ilaç duyarlılık testi. Genotipik antiviral direnç testi, ilacın direncine sebebiyet veren mutasyonları tespit etmek amacıyla enzimatik genlerin dizileme sonucunun incelenmesini içerirken, fenotipik ilaç duyarlılık testi ise bir virüsün farklı konsantrasyonlarda bulunan ilaçlara karşı büyüme yeteneğini ölçmek için tasarlanmıştır. Genotipik antiviral direnç testinde ilgili bölgeler PCR ile çoğaltıldıktan sonra Sanger gibi yeni nesil dizileme yöntemleri kullanılarak mutasyonlar belirlenir. Elde edilen mutasyon listesi önceden bilinen direnç veri tabanları ile karşılaştırılması yapılır ve hangi ilaca karşı direnç geliştirildiği tespit edilmeye çalışılır. Hızlı olmasına rağmen direncin fenotip üzerindeki etkisi ölçülemez. Fenotipik ilaç duyarlılık testi içinse virüs, farklı dozlarda antiviral ilaçlara maruz kalır ve nasıl tepki gösterdiği incelenir. Virüs kültür içinde büyütülür, kültüre farklı dozlarda antiviral ilaç eklenir ve viral replikasyon hızı hesaplanarak dirençli olduğu ilaç belirlenir. Buna ek olarak IC50 (yarı maksimum inhibitör konsantrasyonu) ortaya çıkarılarak nicel bir değerlendirme yapılmaya çalışılır. Gerçek direnç mekanizmasını ortaya koysa da uzun sürmesi ve pahalı olması sebebiyle dezavantaja sahiptir (Strasfeld & Chou, 2010). SARS-CoV-2’de bazı mutasyonlar sayesinde virüs farklı mekanizmalar geliştirerek kendini ilaçlara karşı dirençli hale getirmiştir. Örneğin ilaç direnci için virüs, genellikle RNA polimerazı (RdRp – nsp12) ve ana proteazı (3CLpro – nsp5) mutasyon bölgesi olarak hedef almaktadır. Örneğin E802D/A, S759A, V792I, C799F, V166L mutasyonları RdRp üzerinde değişikliğe sebebiyet vererek SARS-CoV-2 için kullanılan Remdesivir ilacına karşı dirence yol açmıştır (Batool vd., 2025). S proteininde meydana gelen mutasyonlar ile aşılarla karşı direnç geliştirme ve virüsün konak hücreye bağlanmasını etkileme gibi mekanizmalar ile virüs dirençli hale gelmeye çalışmıştır (Mushebenge

vd., 2023). Hangi mutasyonların hangi ilaç üzerine etkisi olduğu Çizelge 1.1’de verilmiştir. Farklı ilaçlar, farklı mutasyonlardan etkilenebileceği gibi, aynı mutasyondan da etkilenme potansiyeline sahiptir.

Çizelge 1.1: SARS-CoV-2’de görülen mutasyonlar ve ilaçlara olan direnci.

Mutasyon	Viral Bölge	Hedef Protein / Enzim	İlaç Direnci Etkisi
E802D/A	Nsp12 (Replikasyon kompleksi)	RdRp (RNA Bağımlı RNA Polimeraz)	Remdesivir direnci
S759A	Nsp12	RdRp	Remdesivir direnci
V792I	Nsp12	RdRp	Remdesivir direnci
C799F	Nsp12	RdRp	Remdesivir direnci
V166L	Nsp12	RdRp	Remdesivir direnci
L167F	Nsp5 (Ana proteaz)	3CLpro	Paxlovid direnci
H172Q/F	Nsp5	3CLpro	Paxlovid direnci
S144A/M/F/G/Y	Nsp5	3CLpro	Paxlovid ve Ensitreivir direnci
H172Y	Nsp5	3CLpro	Paxlovid direnci
Q192 L/P/T/S/A/I/H/W/C/F	Nsp5	3CLpro	Paxlovid direnci
F140L	Nsp5	3CLpro	Paxlovid ve Ensitreivir direnci
T304I	Nsp5	3CLpro	Paxlovid ve Ensitreivir direnci
L50F + E166V	Nsp5	3CLpro	Paxlovid ve Ensitreivir direnci
G15S	Nsp5	3CLpro	Paxlovid ve Ensitreivir direnci
T211	Nsp5	3CLpro	Paxlovid ve Ensitreivir direnci
A173V	Nsp5	3CLpro	Paxlovid ve Ensitreivir direnci
M49L/I/T/V	Nsp5	3CLpro	Ensitreivir direnci

1.2.8.2 Spike proteinin bağışıklığa karşı direnç gelişimi

S, ilaçlara karşı direnç kazanabildiği gibi, bağışıklık sisteminden kaçarak veya ACE-2 reseptörüne bağlanma kapasitesini artırarak virüsün yayılımının artırılmasına yol açar (Harvey vd., 2021) ancak bağışıklık sistemi, S proteinini farklı mekanizmalarla hedef alarak virüsü engellemeye çalışır. Tip I ve III İnterferon (IFN) yanıtı mekanizmasında, RNA algılayıcı reseptörler, SARS-CoV-2’nin RNA’sını tanıyarak interferon üretimini başlatır. IFN sinyalizasyonu, virüs çoğalmasını engellemeye

alışan antiviral genlerin (ISG) aktivasyonuna yol aar. Nötraliye Edici Antikorlar, bağıřıklık sisteminin S proteinine karşı ürettiğı antikorlardır. Bu antikorlar, S proteininin ACE2'ye bağlanmasını engelleyerek virüsün hücreye girişini önlemeye alışır. T Hücre Tepkisi, S proteini üreten enfekte hücreleri hedef alarak virüsü ortadan kaldırmayı amaçlar. Bu mekanizmalar, bağıřıklık sisteminin SARS-CoV-2'yi kontrol altına almak için geliřtirdiğı doėal savunma yollarıdır (Minkoff & tenOever, 2023). Bu gibi mekanizmalar bağıřıklık sisteminin paraları olsa da S proteininde meydana gelen mutasyonlar sebebiyle virüs ilgili mekanizmalardan kaçmayı ok iyi başatabilmektedir. Bu mutasyonlar, S proteininin genetik yapısında meydana gelmektedir (Harvey vd., 2021; Shah & Woo, 2022). izelge 1.2'de, S proteininde gözlemlenen mutasyonlar ile bu mutasyonların potansiyel biyolojik etkileri gösterilmektedir. Her bir mutasyon, virüsün bulařıcılığı, konak hücreye bağlanma yeteneğı ve bağıřıklık sisteminden kaçma kapasitesi üzerinde farklı etkiler yaratmaktadır.

izelge 1.2: S Proteinde görülen mutasyonlar ve etkisi.

Mutasyon	İřlevi
D614G	Virüsün hücrelere girişini ve bulařıcılığını artırır.
N501Y	ACE2 reseptörüne bağlanma gücünü artırır.
S477N	ACE2'ye bağlanmayı artırır.
Y453F	ACE2'ye bağlanmayı artırır.
L452R	Bulařıcılığı artırır.
P681H/P681R	Bulařıcılığı artırır.
E484K	Antikorların virüsü nötraliye etme yeteneğini azaltır.
N439K	ACE2'ye bağlanmayı artırır ve bazı monoklonal antikorlardan kaçmaya yardımcı olur.
K417N/T	Antikor nötralizasyonunu azaltabilir.
Δ69-70 (H69del/V70del)	Antikorlardan kaçmaya yardımcı olabilir.
Δ144/Δ242-244	Antikorlardan kaçışı kolaylařtırır.
T478K	ACE2 bağlanmasını güçlendirir.
Q493K	ACE2 bağlanmasını güçlendirir.
Q498R	ACE2 bağlanmasını güçlendirir.
E484A	Nötralizasyon kaçışı sağlar.
N440K	Antikor bağlanmasını zayıflatır.
K417N	Antikor bağlanmasını zayıflatır.
Y505H	Antikor bağlanmasını zayıflatır.
G339D	Antikorlardan kaçışı destekler.
S371L	Antikorlardan kaçışı destekler.
S373P	Antikorlardan kaçışı destekler.
S375F	Antikorlardan kaçışı destekler.

1.2.9 Modelleme çalışmaları

Proteinlerde meydana gelen mutasyonların etkilerini gözlemleyebilmek ve ilgili üç boyutlu yapıyı ortaya çıkarmak için modelleme çalışmaları büyük önem taşımaktadır. Protein veya DNA sekans analizlerinin tek başına yeterli olmaması ve süreçlerin uzun sürmesi nedeniyle, proteinlerin davranışlarını ve birbirleriyle olan ilişkilerini incelemek amacıyla birçok çalışma gerçekleştirilmiştir.

1.2.9.1 Protein sekans hizalaması

Protein sekanslarının analiz edilmesi ve benzerliklerinin ortaya konulması, proteinlerin anlaşılması ve deneysel çalışmaların yapılması için büyük önem taşımaktadır (Apweiler vd., 2004). Proteinlerin anlaşılmasının ilk adımı olarak amino asit sekans analizi yer almaktadır. Amino asit sekansların analiz edilmesi için öncelikle hizalanması gerekmektedir. Bu hizalama sonucunda, aynı olan amino asitler alt alta sıralanırken, bir amino asidin bir sonraki nesilde değişmesi farklı bir harf ile, silinmesi ise boşluk ile gösterilmektedir (Haritha vd., 2018). Yapılan hizalamalar puanlandırılarak değerlendirilmektedir ve hizalama yöntemleri, en yüksek puanı alacak şekilde en uygun hizalamayı yapmayı hedeflemektedir. Bu işlemler, sekans uzunluğunun fazla olması nedeniyle zorlayıcı hale gelebilir. Bu süreci optimize etmek için sınıflandırma, kümeleme, regresyon, karar ağaçları, yapay sinir ağları ve gizli Markov modelleri gibi algoritmalarla desteklenen programlama tekniklerinden yararlanılmaktadır (Mathura & Kanguane, 2009). Sekans analizleri için çeşitli hizalama teknikleri kullanılmaktadır. Bu yöntemler sayesinde protein, DNA ve RNA sekans dizileri benzerliklerine göre hizalanarak evrimsel ilişkilerin incelenmesine olanak sağlanır. Çift dizi hizalama yöntemleri, iki dizinin karşılaştırılmasını sağlar. Yerel hizalama, sekans içerisindeki küçük benzer bölgelere odaklanırken, Global hizalama, tüm dizileri baştan sona karşılaştırır. Bu kapsamda, Needleman-Wunsch algoritması ve AVID, LAGAN, ACANA gibi yöntemler kullanılmaktadır. Çoklu dizi hizalama, iki veya daha fazla dizinin karşılaştırılmasıyla yapılan hizalama türüdür. Bu yöntem, daha kapsamlı analizler sunarak evrimsel ilişkilerin belirlenmesine yardımcı olur. ClustalOmega, MAFFT, ClustalW, T-Coffee, DIALIGN, PROBCONS ve SAGA, çoklu dizi hizalama için yaygın olarak kullanılan tekniklerdir (Wang vd., 2025). Hizalama için kullanılan amino asit sekansları genellikle FASTA formatında

saklanmaktadır. FASTA formatı, önce bir açıklama satırı ve ardından ilgili sekans bilgisini içeren bir çıktı yapısıdır (Şekil 1.9).

```
>Spike|hCoV-19/Wuhan/WIV04/2019|2019-12-30|EPI_ISL_402124|Original/1-1273
MFVFLVLLPLVSSQCVNLTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSG
TNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTLDSTQSLIVNNATNVVIVKVEFQFCNDPFLGVY
YHKNNKSWMESEFRVYSSANNCTFEYVSQPFMDLEGKQGNFKNLREFVFNIDGYFKIYSKHTPINLVRDL
PQGFSALEPLVDLPIGINITRFQTLALHRSYLTGDSGGWTAGAAAYVGYLQPRTFLLKYNGTITDA
VDCALDPLSETKCTLKSFTEKGIYQTSNFRVQPTESIVRFPNITNLCPGEVFNATRFASVYAWNRRKISN
CVADYSVLVNSASFSTFKCYGVSPKLNLCFTNVYADSFVIRGDEVQIAPGQTGKIADYNYKLPPDFTGC
VIAWNSNNLDSKVGNGNYLYRLFRKSNLKPFFERDISTEIQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVG
YQPYRVVLSFELLHAP-ATVCGPKKSTNLVKNKCVNFENGLTGTGVLTESNKKFLPFQQFGRDIADTTDA
VRDPQTLIELDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGNSVFTQRA
GCLIGAEHVNNSEYCDIPIGAGICASYQTQNSPRRARSVASQSIIAYTMSLGAENSVAYSNNISAIPTNFT
ISVTTEILPVSMTKTSVDCTMYICGDSSTECNLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKT
PPIKDFGGNFESQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIARDLCAQKFNGLTVLPP
LLTDEMIQAQYTSALLAGTITSGWTFGAGAAQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQ
DSLSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAIISSVLNDILSRDLKVEAEVQIDRLITGRQLSQTY
VTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLSFQPSAPHGVVFLHVTYVPAQEKNTTAP
AICHGKAHFPREGVFSNGTHWFVTQRNFYEPQIITDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEE
LDKYFKNHTSPDVLGDISGINASVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAAG
LIAIVMVTIMLCMTSCCCLKGCCSCGSCCKFDEDDSEPVLLKGVKLHYT
```

Açıklama Kısmı

**Aminoasit dizilimi/
Protein sekansı**

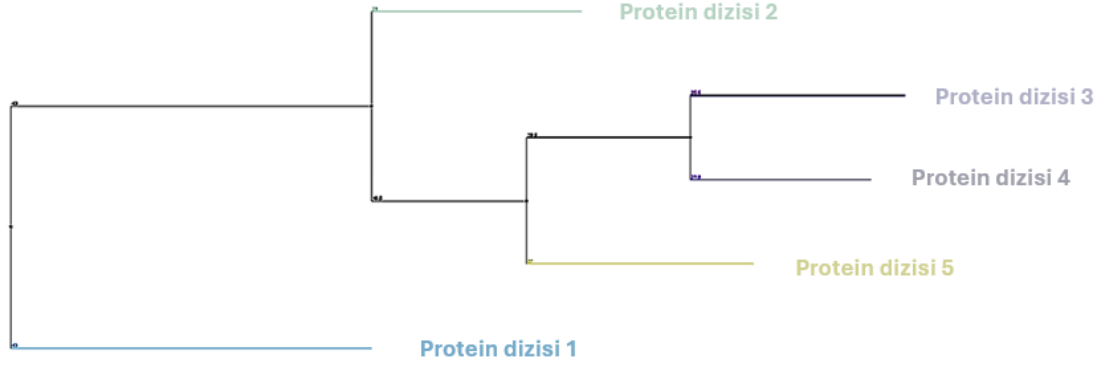
Şekil 1.9: SARS-CoV-2 virüsünün Spike (S) proteinine ait amino asit dizilimini içeren FASTA formatında bir örnek gösterimi. “>” işaretiyle başlayan ilk satır, sekansa dair açıklayıcı bilgileri içerirken (örneğin, virüs adı, izolasyon tarihi ve veri tabanı kimliği gibi), devamındaki satırlar ise proteini oluşturan amino asit dizisini göstermektedir.

Protein dizileri için kullanılan başlıca veri tabanları PIR (Protein Information Resource) güvenilir bir veri tabanı olup buna ek olarak SWISS-PROT ise yüksek düzeyde açıklamalar ve detaylı bilgiler içeren kapsamlı bir protein dizi veri tabanıdır. Swiss-Prot'un bilgisayar destekli ek veri tabanı olan TrEMBL ise büyük ölçekli protein dizilerini içeren otomatik olarak oluşturulmuş bir veri kaynağıdır. Bunlara ek olarak, proteinlerin üç boyutlu yapılarının incelenmesi için kullanılan veri tabanları da mevcuttur. PDB (Protein Data Bank), dünya çapında proteinlerin üç boyutlu yapılarının yer aldığı geniş bir veri tabanıdır. SCOP (Structural Classification of Proteins), proteinleri ikincil yapı özelliklerine göre sınıflandırıp inceleyen bir sistemdir. CATH ise protein yapılarının hiyerarşik olarak sınıflandırılmasını sağlayan bir diğer önemli yapısal veri tabanıdır (Url-7; Apweiler vd., 2004).

1.2.9.2 Filogenik ağaç oluşturma

Filogenetik ağaçlar, organizmalar veya genetik diziler arasındaki evrimsel ilişkileri belirlemek ve görselleştirmek için oluşturulan dallanmış diyagramlardır. Bu ağaçlar, ortak bir atadan türeyen türlerin zaman içinde nasıl dallandığını ve birbirleriyle olan ilişkilerini gösterir. Filogenetik analiz süreci, protein dizileri arasındaki evrimsel farklılıkları belirlemek için çoklu dizi hizalama (Multiple Sequence Alignment- MSA)

yapılmasıyla başlar. Protein dizi hizalama işlemi, organizmalar arasındaki homolog bölgeleri belirleyerek evrimsel değişikliklerin tespit edilmesine yardımcı olur. Daha sonra, hizalanmış protein dizileri arasındaki genetik mesafe hesaplanarak filogenetik ağacın yapısı belirlenir. Genetik mesafe hesaplama aşamasında Jukes-Cantor, Kimura 2-Parametre (K2P) gibi modeller kullanılarak diziler arasındaki mutasyon oranları belirlenir. Genetik mesafeler hesaplandıktan sonra ağacın oluşturulması için farklı algoritmalar kullanılır. Filogenetik ağaç oluşturma yöntemleri temelde üç gruba ayrılır. Mesafeye dayalı yöntemlerde, organizmalar arasındaki en kısa genetik mesafeler belirlenerek ağaç oluşturulur. Bu yöntemde yaygın olarak Neighbor-Joining (NJ) ve UPGMA (Unweighted Pair Group Method with Arithmetic Mean) algoritmaları kullanılır. Karakter temelli yöntemlerde ise maksimum olasılık (Maximum Likelihood- ML) ve maksimum parsimon (Maximum Parsimony- MP) yaklaşımları kullanılarak en az evrimsel değişiklik gerektiren ağaç belirlenir. Üçüncü bir yöntem olarak Bayesçi olasılık temelli yöntemlerde, evrimsel değişim modelleri kullanılarak en olası filogenetik ağaç tahmin edilir. Bayes yaklaşımı, genetik değişim oranlarını göz önünde bulundurarak olasılıklara dayalı bir ağaç tahmini sunar. Filogenetik ağacın oluşturulmasından sonra, elde edilen ağaçların yorumlanması ve görselleştirilmesi için çeşitli yazılımlar kullanılır. FigTree, iTOL (Interactive Tree of Life), PhyloDraw ve Jalview gibi araçlar, ağaçların düzenlenmesi, renklendirilmesi ve detaylandırılması için kullanılır. Ayrıca, filogenetik analiz sonuçları bilimsel doğrulama süreçlerinden geçmeli ve elde edilen ağaçların güvenilirliği istatistiksel yöntemlerle test edilmelidir (Machová Klimešová, 2021). Jalview programı ile elde edilmiş bir filogenetik ağaç Şekil 1.10'da gösterilmiş. Burada evrimsel süreçte protein dizisi 3 ve protein dizisi 4 birbirlerine benzediği bu benzerliğe kıyasla dizi 1 ile dizi 4 arasında benzerliğin daha az olduğu söylenebilir.



Şekil 1.10: Jalview yazılımı ile oluşturulmuş filogenetik ağaç gösterimi. Farklı diziler arasındaki evrimsel ilişkiler, ağaç yapısı üzerinden görselleştirilmiştir. Dallanma noktaları, ortak ataları temsil ederken, dalların uzunlukları diziler arasındaki evrimsel mesafeyi yansıtmaktadır.



2. MATEMATİKSEL MODELLEME VE SAYISAL YÖNTEMLER

2.1 SARS-CoV-2 Proteininin Evrimsel İlişkisinin İncelenmesi

Bu tez kapsamında incelenen SARS-CoV-2 virüsünün S proteinine ait genetik kodlar, GISAID (Url-8) adlı veri tabanından elde edilmiştir. Veri tabanındaki S proteinine ilişkin genetik kodlar direkt protein sekansları halinde sağlamaktadır. Güncel verilere ulaşmak adına, 2024'e kadar dönüştürülmüş protein dizileri içeren bir veri seti temin edilmiştir. Bu veri setinde yer alan protein sekansları, öncelikle birbirleriyle benzerlikleri açısından analiz edilmiştir. Aynı olan protein sekansları tespit edilerek veri setinden çıkarılmıştır. Böylece ana sette benzersiz protein dizileri olması sağlanmıştır. Protein sekanslarının okunması sırasında bazı hatalar meydana gelebilmektedir. Örneğin, 1273 amino asit içeren S proteinin içerisinde bazı sekanslar bilinmediği için protein ifade etmeyen X harfi girilmektedir. Bu nedenle, içerisinde bilinmeyen (X) içeren protein sekansları ana setten çıkarılmış ve beklenenden uzun veya kısa olan protein sekansları da veri setinden silinmiştir. S proteininin evrimsel sürecinin tahminlerinin doğruluğunu tespit edebilmek için, öncelikle veri setindeki protein sekansları yıllara göre sınıflandırılmıştır. SARS-CoV-2'nin 2019 yılının aralık ayında ortaya çıkması nedeniyle, bu yıla ait veri seti yalnızca üç protein sekansı içermektedir. 2020 ve 2021 yıllarına ait sekanslar eğitim (train) veri seti olarak, 2022 ve 2023 yıllarına ait sekanslar ise test veri seti olarak belirlenmiştir. Eğitim datasında yer alan sekansların dağılımı Çizelge 2.1'de verilmiştir. Sekans sayısının fazla olması analiz süresinin uzun olmasına sebebiyet vermektedir. Zaman yönetimin iyi ayarlanabilmesi adına, yıllara ait sekanslardan oluşan veri setleri 3 aylık dönemlere ayrılmıştır. Böylece, sekans dağılımları daha ayrıntılı bir şekilde analiz edilerek ana veri setinin oluşturulması için doğru bir yöntem izlenmeye çalışılmıştır. Çizelge 2.1'de veri setlerinin içerdiği sekans sayıları verilmiştir. Her periyottan rastgele protein sekansları seçilerek küçük alt kümeler (subsets) oluşturulmuştur. Bu alt kümeler, CLUSTAL OMEGA (Sievers vd., 2011) yazılımı kullanılarak çoklu dizi hizalamasına tabi tutulmuştur. Hizalama sürecinde, varsayılan ayarlar kullanılmıştır ve dizi benzerliklerinin analizinde, önce diziler arası uzaklıklara dayalı bir rehber filogenetik

ağaç oluşturulmuş, ardından bu yapı üzerinden yönlendirilen Hidden Markov Model (HMM) tabanlı hizalama yöntemi kullanılmıştır. HMM, dizilerdeki konumların olasılıksal geçişlerine dayalı bir model olup, özellikle belirsizlik içeren biyolojik verilerde güçlü hizalama performansı sunar. Elde edilen hizalama sonuçları Fasta formatında dışa aktarılmıştır.

Çizelge 2.1: SARS-CoV-2 S protein sekansların 3 aylık dağılımları.

Yıl	Birinci Periyot (İlk 3 ay)	İkinci Periyot (İkinci 3 ay)	Üçüncü Periyot (Üçüncü 3 ay)	Dördüncü Periyot (Dördüncü 3 ay)	Bilinmeyen	Toplam
19				3		3
20	885	2395	3501	12757	195	19734
21	27008	28971	48875	58737	1416	165007

Hizalanmış alt kümeler, referans sekans ile karşılaştırılarak alt kümeler içinde bulunan mutasyonların listesi çıkarılmıştır. Belirlenen mutasyonlar, ayrıca Jalview (Waterhouse vd., 2009) adlı uygulama ile görselleştirilmiştir. Ana veri setine hangi alt kümelerin dahil edileceğine karar verilmesi için, alt kümelerdeki mutasyonların sayıları ve türleri dikkate alınmıştır. Bunun sonucunda, alt kümeler arasındaki benzerlik katsayıları hesaplanmış ve kümelerin benzerlik ilişkilerine göre filogenik ağaçları oluşturulmuştur. Oluşturulan filogenik ağaçları sayesinde, alt kümeler arasındaki benzerlikler göz önünde bulundurularak ana gruplar belirlenmiştir.

2.2 Amino asitler Arası İlişkiler ve Skorlama Matrisleri

Evrimsel olarak, aynı atadan türeyen proteinler genellikle benzer amino asit dizilimlerine sahiptir. Ancak, bu dizilerde meydana gelen değişiklikler, çeşitliliğin oluşmasına yol açar. Bir amino asidin başka bir amino aside dönüşme olasılığı ise her amino asit çifti için farklıdır. Bu dönüşümlerin hesaplanmasında kodon tablosu kullanılır (Şekil 1.2) ve bu dönüşüm olasılıkları skorlama matrisleri ile sayısal bir şekilde ifade edilebilir. Skorlama matrisleri, amino asitler arasında bulunan evrimsel ilişkiyi modellemek için kullanılır. Bu modelleme amino asitler arasındaki gözlemlenmiş değişim frekansına dayanmaktadır. (Lipman vd., 1984). Matristeki skorlar, doğal gözlemlenen frekans ile beklenen frekansların karşılaştırılmasıyla

hesaplanır (Formül 2.1). Bu formüldeki $Q_{i,j}$ doğal frekansı, P_i i amino asidin beklenen frekansı, P_j j amino asidin beklenen frekansını ifade etmektedir.

$$S_{i,j} = \ln \left(\frac{Q_{i,j}}{P_i \times P_j} \right) \quad (2.1)$$

Bu formül, gözlemlenen değişimlerin beklenen değişimlere göre ne kadar anlamlı olup olmadığını gösterir. Bu hesaplamadan çıkan pozitif değerler amino asitlerin birbirine sıkça dönüştüğünü, negatif değerler ise nadiren dönüştüğünü ifade eder (Dayhoff vd., 1978). Fizikokimyasal özellikleri benzer olan amino asitler daha yüksek bir frekans ile değişmektedir yani benzerliği yüksek olan amino asitlerin birbirlerine dönüşme olasılığı, benzer olmayan amino asitlere göre daha yüksektir. Örneğin, küçük ve nötr bir amino asit olan Glisin (G), yapısal olarak benzeyen Alanin (A) ile daha yüksek değişim frekansı gösterirken, hacimce büyük ve aromatik bir amino asit olan Triptofan (W) ile yer değiştirme olasılığı çok daha düşüktür. İlk skrolama matrisi PAM skrolama matrisidir ve PAM1 100 amino asitlik bir dizide ortalama bir değişimi temsil eder. Daha uzak akraba amino asit dizilerini ifade etmek için PAM205 gibi farklı matrisler de geliştirilmiştir. Alternatif olarak, BLOSUM (Henikoff & Henikoff, 1992) matrisleri, protein ailelerinin korunmuş bölgelerindeki değişimlere odaklanır. Örneğin, BLOSUM62 matrisi, %62 benzerlikteki dizilerden türetilmiştir ve biyoinformatikte yaygın kullanıma sahiptir (Henikoff & Henikoff, 1992). Bunlara ek olarak, GONNET ve PET91 skrolama matrisleri, mevcut dizilere kıyasla daha kapsamlı veri setleri kullanılarak geliştirilmiş ve böylece çeşitli biyoinformatik analizlerde yaygın bir şekilde kullanılmaya başlanmıştır (Gonnet vd., 1992; Jones, Taylor & Thornton, 1992).

2.2.1 Amino asit frekanslarının incelenmesi

Amino asitlerin doğal frekansları, genetik kod tablosu kullanılarak nükleotid üçlülerinin (kodonların) dağılımına göre analiz edilmiş ve her bir kodon kombinasyonuna karşılık gelen amino asit frekansları sayısal olarak hesaplanmıştır (Şekil 2.1). Bu değerler, Formül 2.2 kullanılarak elde edilmiştir. Bu formülde f_i i nci

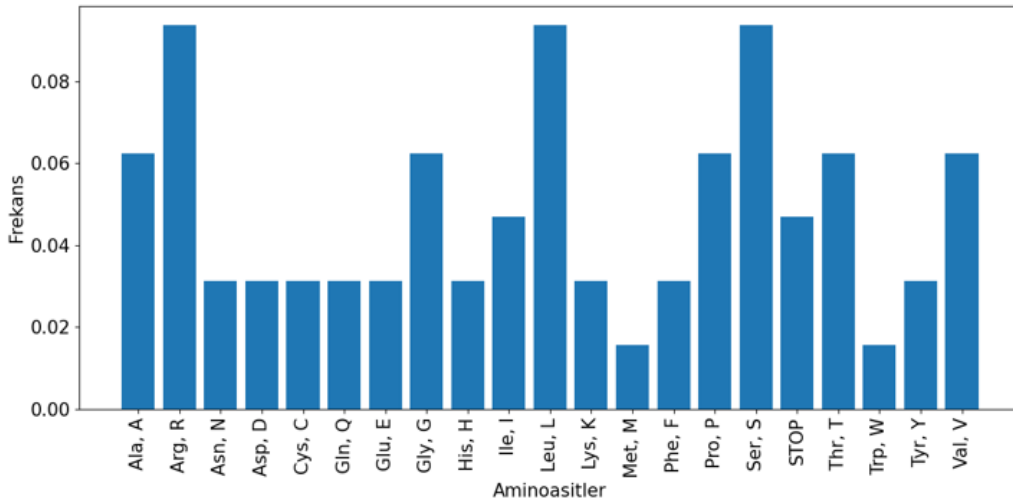
amino asidin frekansını, N_i bu amino asidi sentezleyen kodon sayısını ve son olarak N toplam kodon sayısını ifade etmektedir.

$$f_i = \frac{N_i}{N} \quad (2.2)$$

Ala, A	Arg, R	Asn, N	Asp, D	Asn or Asp, B	Cys, C	Gln, Q	Glu, E	Gln or Glu, Z	Gly, G	His, H	Ile, I	Leu, L	Lys, K	Met, M	Phe, F	Pro, P	Ser, S	Thr, T	Trp, W	Tyr, Y	Val, V	STOP
GCU	CGU	AAC	GAC	AAC	UGC	CAG	GAG	GAG	GGG	CAC	AUA	UUG	AAG	UUC	CCG	AGC	UCC	ACC	UGG	UAC	GUG	UAG
CGC	CGA	AAU	GAU	AAU	UGU	CAA	GAA	CAA	GGG	CAU	AUC	CUA	AAA	AUG	UUU	CCC	UCA	ACU	UUA	GUA	GUC	UGA
GCC	CGG	AAU	GAU	AAU	UGU	CAA	GAA	CAA	GGU	CAU	AUC	CUU	AAA	AUG	UUU	CCU	UCU	ACU	UGG	UUA	GUA	UGA
GCA	AGA	AAU	GAU	AAU	UGU	CAA	GAA	CAA	GGU	CAU	AUC	CUU	AAA	AUG	UUU	CCU	UCU	ACU	UGG	UUA	GUA	UGA
GCG	AGG	AAU	GAU	AAU	UGU	CAA	GAA	CAA	GGU	CAU	AUC	CUU	AAA	AUG	UUU	CCU	UCU	ACU	UGG	UUA	GUA	UGA

Şekil 2.1: Amino asitlerin mRNA üzerindeki kodonlara göre dağılımını gösteren tablo. Her bir amino asit, kendisine karşılık gelen bir veya daha fazla üçlü baz dizisi (kodon) ile şifrelenmektedir. Genetik kodun yapısı gereği, bazı amino asitler birden fazla kodon tarafından kodlanabilir.

Bu analizlere göre farklı çıkarımlar yapılmaktadır. Örneğin, Arjininin görülme olasılığı daha fazlayken, Triptofanın ortaya çıkma ihtimali daha düşüktür. Bu bilgiler bütün amino asitler için çıkarılmıştır (Şekil 2.2).



Şekil 2.2: Kodon tablosundaki kodonların sayılmasıyla elde edilen amino asit frekans dağılımı. Her bir amino asit, kendisini kodlayan kodon sayısına göre temsil edilmiştir. Leucine (L), Arginine (R) ve Serine (S) gibi amino asitler çok sayıda kodon tarafından şifrelendiği için daha yüksek frekansla temsil edilmiştir.

2.2.2 Mutasyon analizleri

Kodon tablosu kullanılarak nükleotitlerin birbirine dönüşüm süreçleri incelenmiştir. Bu inceleme, tek mutasyon, çift mutasyon ve üçlü mutasyon düzeylerinde ele alınmıştır. Her bir mutasyon türü için dönüşüm sayıları ayrı ayrı analiz edilerek,

toplam dönüşüm sayıları birleştirilmiştir. Bu verilerden oluşturulan bir matris, mutasyonların dönüşüm frekanslarını yansıtacak şekilde hesaplanmıştır. Şekil 2.3'teki matriste görüldüğü gibi, amino asitlerin birbirlerine dönüşüm sayıları bir ısı matrisi biçiminde sunulmuştur. Yüksek sayılar, dönüşümlerin daha kolay olduğunu, düşük sayılar ise dönüşümün daha zor olduğunu göstermektedir. Örneğin, N amino asidinin M'ye dönüşmesi için yalnızca 2 farklı yol bulunurken, N'nin R'ye dönüşmesi için 12 farklı yol mevcuttur. Bu da N'nin R'ye dönüşmesinin M'ye dönüşmesinden çok daha olası olduğunu gösterir. Bu sayede, hangi amino asitlerin birbirlerine daha sık dönüşme eğiliminde olduğu açık bir şekilde görülebilir. Formül 2.2 kullanılarak elde edilen frekanslar Çizelge 2.2'de verilmiştir. Bu formülde f_i i'nci amino asidin dönüşüm frekansını, N_i bu amino asidi sentezleyen toplam dönüşüm sayısını ve son olarak N toplam dönüşüm sayısını ifade etmektedir. Bu formül kullanılarak tüm amino asitler için dönüşüm frekansı Şekil 2.3'teki matris kullanılarak elde edilmiştir.

A	12	24	8	8	8	8	8	16	8	12	24	8	4	8	16	24	16	4	8	16	12
R	24	30	12	12	12	12	12	24	12	18	36	12	6	12	24	36	24	6	12	24	18
N	8	12	2	4	4	4	4	8	4	6	12	4	2	4	8	12	8	2	4	8	6
D	8	12	4	2	4	4	4	8	4	6	12	4	2	4	8	12	8	2	4	8	6
C	8	12	4	4	2	4	4	8	4	6	12	4	2	4	8	12	8	2	4	8	6
Q	8	12	4	4	4	2	4	8	4	6	12	4	2	4	8	12	8	2	4	8	6
E	8	12	4	4	4	4	2	8	4	6	12	4	2	4	8	12	8	2	4	8	6
G	16	24	8	8	8	8	8	12	8	12	24	8	4	8	16	24	16	4	8	16	12
H	8	12	4	4	4	4	4	8	2	6	12	4	2	4	8	12	8	2	4	8	6
I	12	18	6	6	6	6	6	12	6	6	18	6	3	6	12	18	12	3	6	12	9
L	24	36	12	12	12	12	12	24	12	18	30	12	6	12	24	36	24	6	12	24	18
K	8	12	4	4	4	4	4	8	4	6	12	2	2	4	8	12	8	2	4	8	6
M	4	6	2	2	2	2	2	4	2	3	6	2	0	2	4	6	4	1	2	4	3
F	8	12	4	4	4	4	4	8	4	6	12	4	2	2	8	12	8	2	4	8	6
P	16	24	8	8	8	8	8	16	8	12	24	8	4	8	12	24	16	4	8	16	12
S	24	36	12	12	12	12	12	24	12	18	36	12	6	12	24	30	24	6	12	24	18
T	16	24	8	8	8	8	8	16	8	12	24	8	4	8	16	24	12	4	8	16	12
W	4	6	2	2	2	2	2	4	2	3	6	2	1	2	4	6	4	0	2	4	3
Y	8	12	4	4	4	4	4	8	4	6	12	4	2	4	8	12	8	2	2	8	6
V	16	24	8	8	8	8	8	16	8	12	24	8	4	8	16	24	16	4	8	12	12
X	12	18	6	6	6	6	6	12	6	9	18	6	3	6	12	18	12	3	6	12	6
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X

Şekil 2.3: Kodon tablosu temel alınarak oluşturulmuş amino asitler arası mutasyon dönüşüm matrisi. Her hücrede yer alan sayılar, bir amino asidin başka bir amino aside kaç farklı kodon mutasyonu üzerinden dönüşebileceğini, yani kaç farklı şekilde dönüşüm gerçekleşebildiğini göstermektedir.

Çizelge 2.2: Amino asitlerin dönüşüm frekansları.

Amino asit	Amino asit Sayısı	Amino asit Frekans
A	252	0,063
R	378	0,094
N	126	0,031
D	126	0,031
C	126	0,031
Q	126	0,031
E	126	0,031
G	252	0,063
H	126	0,031
I	189	0,047
L	378	0,094
K	126	0,031
M	63	0,016
F	126	0,031
P	252	0,063
S	378	0,094
T	252	0,063
W	63	0,016
Y	126	0,031
V	252	0,063

2.3 Entropi Hesabı ve Skorlama Fonksiyonları

Bir protein, işlevini yerine getirebilmesi ve soyunu sürdürebilmesi için yapısal bütünlüğünü korumaya ya da yeni mutasyonlar ile evrimleşmeye eğilimlidir. Amino asitlerde meydana gelen her bir mutasyon, entropi ve “mutasyon skoru fonksiyonu” adı verilen matematiksel fonksiyonlarla ifade edilebilir. Bu bağlamda, literatürde daha önce tanımlanmış dört farklı mutasyon skorlama fonksiyonu (Karlin, Hogervorst, Sander ve Valdar) ile, tarafımızdan geliştirilen iki yeni fonksiyon (BNT2 ve BNT3) ayrıntılı olarak incelenmiştir. İncelenen bu fonksiyonlar, her bir pozisyon için mutasyonel değişebilirlik düzeyini yansıtan sayısal skorlar üretmektedir. Elde edilen bu değişebilirlik skorları, korelasyon analizleri aracılığıyla birbirleriyle karşılaştırılmış; fonksiyonlar arasında gözlenen benzerlikler ve farklılıklar belirlenmeye çalışılmıştır. Böylece, farklı skorlama yaklaşımlarının çıktıları arasındaki ilişkiler ortaya konmuş ve yöntemler arasındaki tutarlılık ya da ayrışma noktaları değerlendirilmiştir.

2.3.1 Entropi hesaplaması

Entropi kavramı ilk olarak 19. yüzyılda Clausius tarafından tanımlanmıştır ve genel anlamıyla düzensizlikle ilişkilendirilen bir kavramdır (Atkins & de Paula, 2010). Fizikten bilgi teorisine kadar birçok alanda kullanılan bu kavram, sistemlerin karmaşıklığı ve düzensizliği hakkında bilgi sağlar.

Termodinamiğin ikinci yasası kapsamında değerlendirilen ve izole sistemlerde entropi artışı ilkesine dayanan bu yaklaşım, entropinin temelini oluşturur. Söz konusu yasa, enerjinin iş yapma kapasitesinin zamanla azaldığını ve sistemlerin düzensizlik seviyelerinin arttığını ifade eder. Dolayısıyla entropi, evrende düzen ile düzensizlik arasındaki dengenin anlaşılmasını sağlayan temel bir fiziksel büyüklüktür. Bu bağlamda, entropi kavramı başlangıçta enerji dönüşümlerinin geri döndürülemezliğini (irreversibilite) ölçmek amacıyla kullanılmıştır. Bu yaklaşımda entropi değişimi formül 2.3'teki gibi tanımlanır. Bu formül bir sistemin entropi değişimini ifade etmektedir. Bu formüldeki dQ_{rev} sistemin çevresiyle değiş-tokuş ettiği tersinir (reversible) ısı miktarının diferansiyel bir parçasıdır. Bu ısının aktarımı sırasında sistemin mutlak sıcaklığı da T ile ifade edilir (Blundell & Blundell, 2016)

$$\Delta S = \int \frac{dQ_{rev}}{T} \quad (2.3)$$

Bu kavram, daha sonra istatistiksel mekanik çerçevesinde mikro düzeye indirgenmiş ve bir sistemin sahip olabileceği olası mikro durum sayısı ile ilişkilendirilmiştir. Bu bağlamda entropi, bir sistemin sahip olabileceği olası mikro durumların (Ω) sayısının logaritması ile ilişkilidir (Formül 2.4). Boltzmann sabiti k_B ile ifade edilir. Kısacası mikro durum sayısı arttıkça sistemin düzensizliği artar, dolayısıyla entropi de artar (Blundell & Blundell, 2016).

$$S = k_B \ln \Omega \quad (2.4)$$

Entropi kavramı, bilgi teorisine Claude Shannon tarafından 1948 tarihli "İletişimin Matematiksel Teorisi" adlı eseriyle kazandırılmıştır (Shannon & Weaver, 1949). Shannon'a göre, entropi iletilen veya incelenen bir mesajın içerdiği bilgi miktarını ölçer. Bu kavramın mantığı şu şekilde açıklanabilir: Bir olayın gerçekleşme olasılığı yüksek olduğunda, bu olay fazla bilgi vermez çünkü beklenen bir durumdur. Ancak, gerçekleşme olasılığı düşük olan bir olayın gerçekleşmesi daha fazla bilgi sağlar çünkü

beklenmeyen bir durumdur. Bu nedenle, entropi yüksek olasılıklı durumlar için düşüken, düşük olasılıklı durumlar için yüksektir. Bu şekilde, entropi kavramı iletilen bilginin beklenti derecesini ölçerek bilgi aktarımında önemli bir rol oynar. Bu yaklaşım özellikle veri sıkıştırma, bilgi aktarımı, makine öğrenimi ve biyoinformatik gibi alanlarda yaygın şekilde kullanılmaktadır (MacKay, 2003; Cover & Thomas, 2006). Shannon Entropinin hesaplaması formül 2.5'te gösterilmiştir. Burada $P(x_i)$ x olayının gerçekleşme olasılığını vermektedir.

$$H(x) = \sum P(x_i) \log_2(P(x_i)) \quad (2.5)$$

Bu çalışma kapsamında, Shannon entropi genetik değişimlerden bilgi elde etmek için kullanılmıştır. Örneğin, toplam 3199 farklı diziyi içeren bir veri setinde 371. pozisyonu ele alındığında, bu dizide R 1177 kez, H 1378 kez, L 1 kez ve P 643 kez gözlemlenmiştir. 371. pozisyon için Shannon entropisi aşağıdaki formülle hesaplanır:

$$H = \frac{1177}{3199} \log_2\left(\frac{1177}{3199}\right) + \frac{1378}{3199} \log_2\left(\frac{1378}{3199}\right) + \frac{1}{3199} \log_2\left(\frac{1}{3199}\right) + \frac{643}{3199} \log_2\left(\frac{643}{3199}\right) \quad (2.6)$$

371. pozisyon için entropi hesaplaması sonucunda elde edilen değer 1.056'dır. Bu hesaplamalar incelendiğinde, bir bölgenin Shannon entropisinin artmasıyla o bölgenin çeşitliliğinin arttığı ve dolayısıyla evrimsel süreçte çok değiştiği söylenebilir.

2.3.2 Skorlama fonksiyonların kullanımı

Skorlama fonksiyonları, Valdar ve çalışma arkadaşları (Valdar, 2002) tarafından geliştirilmiş çeşitli yöntemleri içermektedir. Bu fonksiyonlardan bir kısmı, frekans içeren veya benzeri matrislerle çalışan yöntemleri kapsar. Ancak, bu tür skorlama fonksiyonları diğer skorlama fonksiyonlarına göre daha yüksek hata payına sahip olma eğilimindedir. Çalışma kapsamında, en düşük hata oranına sahip skorlama fonksiyonları tespit edilmiş; bunlar, korelasyon analizleri ve yöntemler arasındaki karşılaştırmalarla incelenmiştir. Bu çalışmada ele alınan skorlama matrisleri iki ana grupta incelenmiştir. Bir grup, doğrudan skor matrislerini kullanırken, diğer grup ise hesaplamalara sekansların benzerliklerinden çıkardığı bir ağırlık katsayısı ekleyerek farklı bir yaklaşım benimsemiştir. Skorlama fonksiyonlarının çalışma prensibi şu şekildedir: N adet farklı, L uzunluğunda hizalanmış sekanslar üzerinde (Şekil 2.4) tüm pozisyonlar tek tek gezilir ve her pozisyona ilgili skorlama fonksiyonları uygulanır. Bu işlemin sonucunda, her pozisyon için bir skor elde edilir.

Sekans #	Primer Yapı
1	A B C X Y Z
2	A B C X Y Z
.	
.	
.	
N	A B C X Y Z
Pozisyon: 1 2 3 L	

Şekil 2.4: Sekans dizilim örneği.

Birinci grup skorlama fonksiyonları (Karlin, BNT2, BNT3, Hogervorst)

Karlin fonksiyonu, Formül 2.7 kullanılarak hesaplanmaktadır. Bu fonksiyon, bir sekansın belirli bir pozisyonundaki amino asitlerin çiftler halinde karşılaştırılmasıyla ilgili bir skor üretir. Formülde $s_i(x)$, i-inci sekansın x-inci pozisyonundaki amino asit ve $s_j(x)$, j-inci sekansın x-inci pozisyonundaki amino asidi göstermektedir. Öncelikle, M formülü kullanılarak $s_i(x)$ ve $s_j(x)$ amino asit çiftleri arasındaki karşılaştırma yapılır. Elde edilen sonuç, toplam çift sayısını normalize etmek için

$\frac{2}{N(N-1)}$ ile çarpılır. Burada N , toplam sekans sayısını ifade eder. Bu işlemler

sonucunda, her bir x pozisyonu için Karlin fonksiyonu ile bir skor elde edilir. Eğer bir pozisyonundaki tüm amino asitler birbirine tamamen eşitse, bu durumda fonksiyonun çıktısının normalizasyonunda 1 olur. Bu, o pozisyonundaki amino asitlerin birbirine tamamen benzer olduğu anlamına gelir. Eğer pozisyonundaki amino asitler farklılaşırsa, yani birbirinden farklı özelliklere sahip olursa, Karlin fonksiyonunun çıktısı 0'a yaklaşır. Bu, o pozisyonundaki amino asitler arasındaki farklılıkların arttığını gösterir.

$$f_{karlin} = \sum_i^N \sum_{j>i}^N M(s_i(x), s_j(x)) \times \frac{2}{N(N-1)} \quad (2.7)$$

$$M(a, b) = \frac{m(a, b)}{\sqrt{m(a, b)m(a, b)}} \quad (2.8)$$

Birinci gruba giren diğer skarlama fonksiyonları Hogervorst, BNT2 ve BNT3'tür. Bu fonksiyonlar, Karlin fonksiyonundaki gibi, her bir pozisyon için karşılaştırmalar yaparak bir skor elde ederler. Matematiksel olarak, $m(s_i(x), s_j(x))$, i-inci ve j-inci sekansların x-inci pozisyonlarındaki amino asitlerin karşılaştırmasından elde edilen sayısal değeri ifade eder. Bu skarlama fonksiyonları, Karlin fonksiyonu gibi, her bir x pozisyonu için bir skor üretir (Formül 2.9, Formül 2.10, Formül 2.11).

$$f_{hogervorst}(x) = \sum_i^N \sum_{j>i}^N \left[\frac{m(s_i(x), s_j(x)) [m(s_i(x), s_i(x)) + m(s_j(x), s_j(x))]}{2m(s_i(x), s_i(x))m(s_j(x), s_j(x))} \right] \quad (2.9)$$

$$f_{BNT2}(x) = \sum_i^N \sum_{j>i}^N \left[\frac{2m(s_i(x), s_j(x))}{m(s_i(x), s_i(x)) + m(s_j(x), s_j(x))} \right] \quad (2.10)$$

$$f_{BNT3}(x) = \sum_i^N \sum_{j>i}^N \left[\frac{|m(s_i(x), s_j(x))| m(s_i(x), s_j(x))}{m(s_i(x), s_i(x))m(s_j(x), s_j(x))} \right] \quad (2.11)$$

İkinci grup skarlama fonksiyonları (Sander, Valdar)

İkinci grup skarlama fonksiyonları, Sander ve Valdar fonksiyonlarını içerir. Bu skarlama fonksiyonları, birinci gruptan farklı olarak, sekanslar arasındaki benzerlikleri bir ağırlık katsayısı olarak formüle ederler.

$$d(s_i(x), s_j(x)) = \begin{cases} 1 & s_i(x) = s_j(x) \\ 0 & s_i(x) \neq s_j(x) \end{cases} \quad (2.12)$$

$$w_{ij} = 1 - \frac{1}{L} \sum_x^L d(s_i(x), s_j(x)) \quad (2.13)$$

$$\lambda = \left(\sum_i^N \sum_{j>i}^N d(s_i(x), s_j(x)) \right)^{-1} \quad (2.14)$$

$$f_{sander}(x) = \lambda \sum_i^N \sum_{j>i}^N w_{ij} \cdot m(s_i(x), s_j(x)) \quad (2.15)$$

Sander fonksiyonunda, protein sekansları arasındaki benzerlikler Formül 2.12’de ifade edilen $d(s_i, s_j)$ mesafe formülasyonu kullanılarak hesaplanır. Buradaki formülasyon sayesinde, tamamen aynı olan sekanslar 1 alırken, benzemeyen sekanslar 0'a kadar bir değer alır. Yani, sekanslar arasındaki benzerlik, 1 ile 0 arasında bir skora dönüştürülür. Karşılaştırılan iki sekans üzerinde yapılan hesaplamalar sonrasında, her bir elde edilen değer, λ formülü kullanılarak normalize edilir. Yani, tüm mesafe değerlerinin toplamına bölünerek bir skora dönüştürülür.

$$M(a, b) = \begin{cases} \frac{m(a, b) - \min(m)}{\max(m) - \min(m)} & a \neq \text{boşluk}, b \neq \text{boşluk} \\ 0 & a = \text{boşluk}, b = \text{boşluk} \end{cases} \quad (2.16)$$

$$d(s_i(x), s_j(x)) = 1 - \frac{1}{L} \sum_x^L M(s_i(x), s_j(x)) \quad (2.17)$$

$$w_i = \frac{1}{N-1} \sum_{i \neq j}^N d(s_i(x), s_j(x)) \quad (2.18)$$

$$f_{valdar}(x) = \frac{\sum_i^N \sum_{j>i}^N w_i w_j M(s_i(x), s_j(x))}{\sum_i^N \sum_{j>i}^N w_i w_j} \quad (2.19)$$

Valdar fonksiyonunda, protein sekansları arasındaki benzerlik hesaplaması, Sander fonksiyonunda kullanılan Formül 2.12 yerine, skarlama matrisine dayalı olarak Formül 2.16 ile gerçekleştirilmiştir. Mesafe hesaplamasında ise, elde edilen skorlar

sekans uzunluğu olan L değerine bölünerek belirlenmiştir (Formül 2.17). Valdar fonksiyonunda kullanılan katsayı, her sekans için ayrı ayrı hesaplanmaktadır. Sonuç olarak, Formül 2.18'de olduğu gibi, değerler N-1 ile bölünerek elde edilmiştir.

2.3.3 Metotlar arası korelasyon

Yöntemler arasındaki ilişkiyi incelerken, korelasyon formülü kullanılmıştır (Formül 2.23). Bu formül, kovaryansın standart sapmaya bölünmesiyle elde edilir. Formül 2.20'de ortalama hesabını ifade etmektedir ve x skarlama fonksiyonlarının sonucu ifade eder, N kaç adet skarlama fonksiyonu sonucu olduğunu göstermektedir.

$$mean(A) = \frac{x_1 + x_1 + + x_N}{N} \quad (2.20)$$

$$stdev(A) = \left(\frac{1}{m.(n-1)} \right) \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (A_{i,j} - mean(A))^2 \quad (2.21)$$

$$Cov(A, B) = \left(\frac{1}{m.n} \right) \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} ((A_{i,j} - mean(A))(B_{i,j} - mean(B))) \quad (2.22)$$

$$Corr(A, B) = \frac{Cov(A, B)}{(stdev(A)stdev(B))} \quad (2.23)$$

2.4 Proteinlerin Evrimsel Değişimlerinin Tahmini

Bu çalışma kapsamında, SARS-CoV-2 S proteininin gelecekteki olası mutasyonlarının tahmin edilmesi amaçlanmış olup, bu tahminlerin salgınların önlenmesi adına kritik bir öneme sahip olduğu bilinmektedir. Salgınlar meydana gelmeden önce alınacak önlemler, örneğin virüsün evrimsel sürecine göre gerekli ilaç ve aşılarda geliştirilmesi büyük önem taşımaktadır. Bu bölümde, önceki bölümde elde edilen skor ve entropi matrisleri kullanılarak, gelecekteki mutasyonların olasılıkları incelenmiştir. İşlem yapılırken, mutasyon hızının hesaplanması öncelikli olarak ele alınmış ve rastgele yürüyüş algoritması için amino asit frekansları göz önünde bulundurulmuştur.

2.4.1 SARS-CoV-2 S proteininin mutasyon hızının hesaplanması

Mutasyon hızı genetik koda ilişkin değişikliklerin bir sonraki nesle aktarılma süresi olarak tanımlanır. Bu hız virüslerin yayılım hızını, yaşam süresini ve diğer faktörleri etkileyen önemli bir parametredir. Bir virüsün mutasyon hızının yüksek olması, o virüsün adaptasyon sürecinin daha hızlı olduğunu gösterir ve bu durum, virüsün ölümcül hale gelme potansiyelini artırabilir. Yüksek mutasyon hızı virüsün hızlı bir şekilde evrimleşmesine olanak tanır, dolayısıyla daha dirençli ve ölümcül varyantların ortaya çıkmasına neden olabilir. Mutasyon hızının hesaplanabilmesi için öncelikle S proteininin nükleotid bazındaki mutasyon oranı gerekmektedir. Yapılan çalışmalar sonucunda, S geninin nükleotid mutasyon oranı, bölge başına yıllık 8.066×10^{-4} olarak belirlenmiştir (Wang vd., 2021). Formül 2.24 dikkate alınarak, önce nükleotid değişim oranı hesaplanmış ve bu değeri kullanarak protein değişim olasılığı belirlenmiştir. Bu formülasyonda, zamanı ifade etmek için t kullanılırken, sekans uzunluğunu ifade etmek için L kullanılır. Zaman olarak 1 yıl, sekans uzunluğu olarak 1273 alınmıştır. H_{aa} , bir nükleotidin değişme olasılığını, H_{pr} , bir proteinin değişme olasılığını vermektedir.

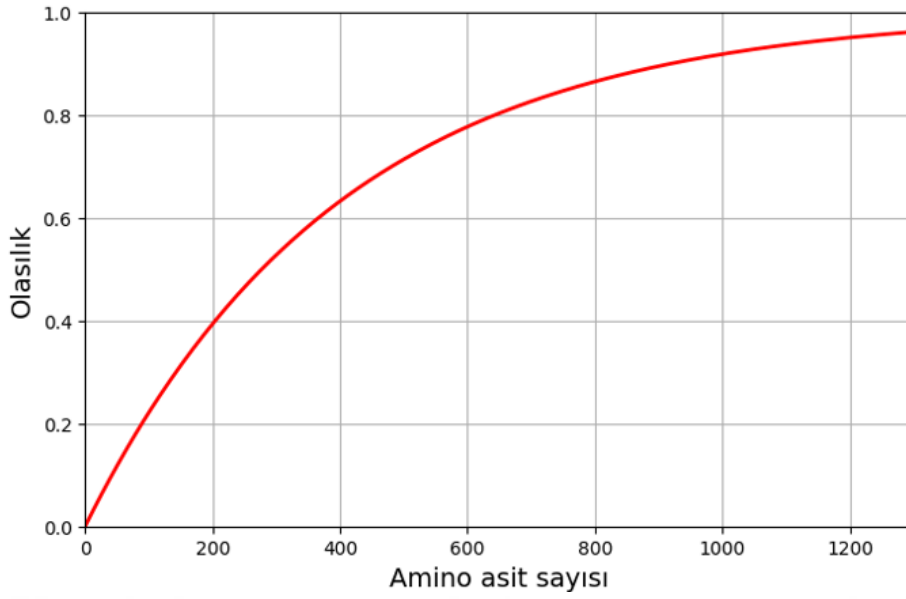
$$H_{aa} = 1 - (1 - H_{nuc})^3 \quad (2.24)$$

$$H_{pr} = 1 - \left[1 - (H_{aa})^z \right]^l \quad (2.25)$$

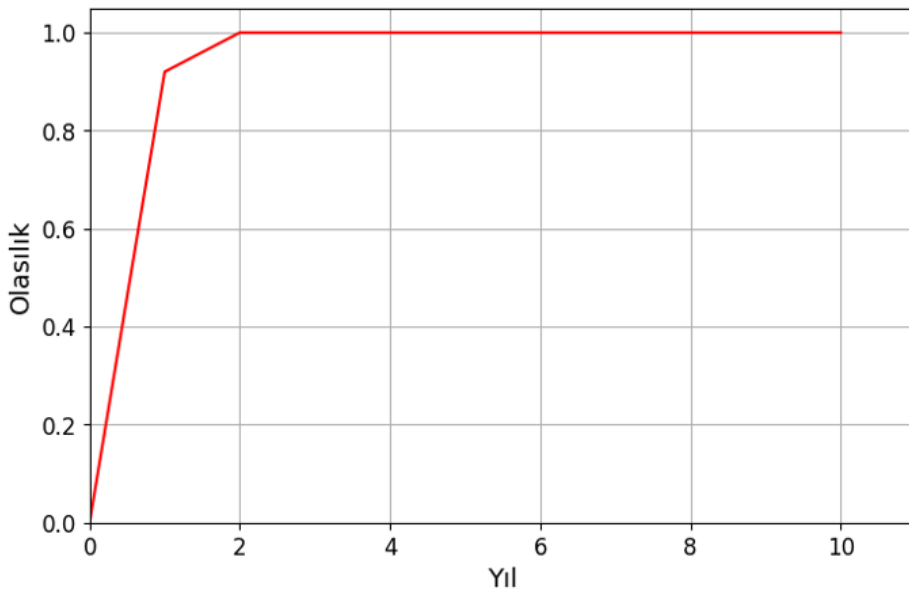
$$H_{aa} = 1 - (1 - 8.066 \times 10^{-4})^3 \quad (2.26)$$

$$H_{pr} = 1 - ((1 - 0.0024)^1)^{1273} \quad (2.27)$$

Mutasyon olasılığı sekans uzunluğu ve geçen zamana bağlı olarak değişir. Şekil 2.5'te görüldüğü gibi, amino asit sayısı arttıkça mutasyon olasılığı da artmaktadır. Geçen zaman, mutasyon olasılığını etkileyen önemli bir faktördür. Şekil 2.6'da görüldüğü gibi, yıl arttıkça mutasyon olasılığı da artmakta ve olasılık 1'e yaklaşmaktadır.



Şekil 2.5: Amino asit sayısının, bir proteinin mutasyona uğrama olasılığı üzerindeki etkisini gösteren grafik. Grafik, amino asit sayısı arttıkça mutasyon olasılığının da yükseldiğini göstermektedir.



Şekil 2.6: Zamanın (yıl bazında) mutasyona uğrama olasılığı üzerindeki etkisini gösteren grafik. Grafik, mutasyon olasılığının zamanla birlikte hızla arttığını ve kısa sürede maksimum seviyeye ulaştıktan sonra sabit kaldığını göstermektedir.

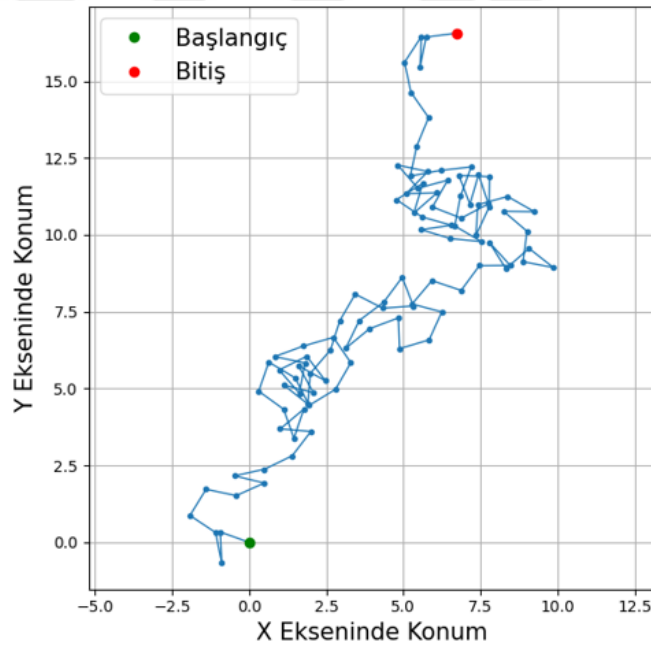
2.4.2 Rastgele yürüyüş (random walk) yöntemi

Rastgele yürüyüş (Random walk) algoritması, bir sistem veya nesnenin rastgele bir yön seçerek ilerlemesini temsil eden bir olasılık modelidir. Bu algoritma ilk başta fiziksel bir sistemde bulunan parçacık hareketlerini açıklayabilmek için geliştirilmiştir fakat ilerleyen zamanlarda istatistik, matematik ve benzeri alanlarda da kullanılan bir

yöntem haline gelmiştir (Casella, Fienberg, & Olkin, 2011; Grimmer & Stirzaker, 2001). Rastgele yürüyüş algoritması genellikle Markov zinciri yapısı içerisinde tanımlanır. Markov zinciri, bir sistemin olası durumları arasında geçiş yaptığı ve her geçişin yalnızca mevcut duruma bağlı olarak belirlendiği bir olasılık modelidir. Bu yapı, gelecekteki bir durumun yalnızca mevcut duruma bağlı olduğunu ve önceki adımlardan etkilenmediğini ifade eder (Norris, 1998). Bu özelliğe 'hafızasızlık' (memorylessness) denir ve biyolojik süreçlerin modellenmesinde sıklıkla kullanılmasının temel nedenlerinden biridir (Codling, Plank, & Benhamou, 2008). Bu rastgele hareket Formül 2.28 gösterildiği gibi ifade edilir.

$$X_n = X_{n-1} + \xi_n \quad (2.28)$$

Burada ξ_n , rastgele seçilen bir adımı temsil etmektedir ve genellikle -1, +1 değerlerini alır. Bu durum parçacığın veya sistemin sağa ya da sola hareket edeceğini ifade eder (Şekil 2.7).



Şekil 2.7: Rastgele yürüyüş algoritması ile bir düzlemde gerçekleştirilen iki boyutlu hareketin temsili görselleştirmesi. Yeşil nokta hareketin başlangıç konumunu, kırmızı nokta ise rastgele adımlar sonucunda ulaşılan bitiş konumunu göstermektedir.

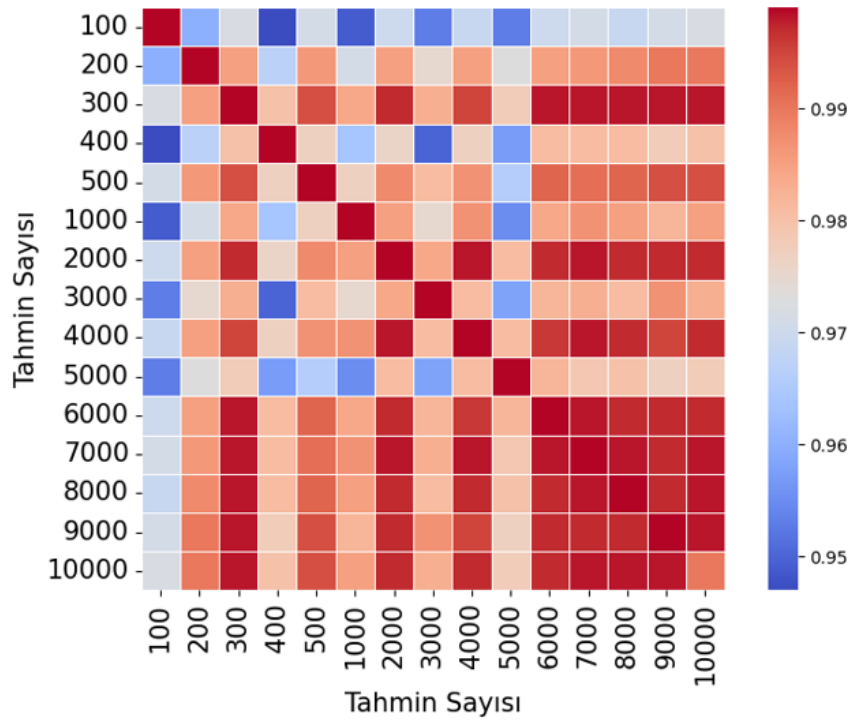
Bu çalışma kapsamında SARS-CoV-2 S proteini üzerindeki mutasyonların olasılıksal olarak hangi pozisyonlarda ve ne sıklıkla gerçekleşebileceğini modellemek amacıyla bir sistem kurulmuştur. Sistem içerisindeki tüm değişkenler, bir toplam vektörü olarak ifade edilir. Toplam vektörü, sistemdeki değerlerin ardışık bir şekilde toplanmasıyla

elde edilir. Rastgele yürüyüş algoritmasında, her bir adım, 0 ile toplam vektörün maksimum değeri arasında bir değer alır. Bu yöntem, sistemin dinamiklerini anlamak ve gelecekteki durumları tahmin etmek için etkili bir araçtır. Formül 2.29'da görülen vektörde sistemin çalışma prensibi gösterilmiştir. Burada pozisyon bazında üretilen mutasyon skorları toplanarak ilerlenir. Rastgele üretilen sayı belirli bir aralığa düştüğünde, o aralıkla ilişkilendirilen pozisyonda uygun amino asit mutasyonu tahmin edilir.

$$\begin{aligned}
 x &= \begin{bmatrix} x_1 \\ x_1 + x_2 \\ x_1 + x_2 + x_3 \\ . \\ . \\ . \\ x_1 + x_2 + x_3 + \dots + x_N \end{bmatrix} \quad x = \begin{bmatrix} Toplam_1 \\ Toplam_1 + Toplam_2 \\ Toplam_1 + Toplam_2 + Toplam_3 \\ . \\ . \\ . \\ Toplam_1 + Toplam_2 + Toplam_3 + \dots + Toplam_N \end{bmatrix} \\
 Aralık &= \begin{bmatrix} 0 < A < Toplam_1 \\ Toplam_1 < A < Toplam_2 \\ Toplam_2 < A < Toplam_3 \\ . \\ . \\ . \\ Toplam_3 < A < Toplam_4 \end{bmatrix} \quad (2.29)
 \end{aligned}$$

S proteininin evrim sürecinde meydana gelebilecek mutasyonlarını tespit etmek için rastgele yürüyüş (random walk) algoritmasında iki farklı değişken dikkate alınır. Bunlardan ilki, skor fonksiyonlarından çıkan skorlar veya entropi hesaplamasından elde edilen skorlar; ikincisi ise amino asitlerden elde edilen frekanslardır. Bu iki değişken, algoritmanın her adımında dikkate alınarak sistemin dinamiklerini ve evrimsel sürecini modellemek için kullanılır. İlk değişken olan skorlar, mutasyonun gerçekleşeceği pozisyonu belli eder, kullanılan amino asit frekansları ise o pozisyonda hangi amino asidin mutasyonunun gerçekleşeceğini gösterir. Örneğin, 200 amino asit uzunluğundaki bir protein dizisinde rastgele yürüyüş algoritmasının sonucu 112 R olarak belirlenmişse, bu verilen protein sekans dizinin 112. pozisyonundaki amino asidin Arjinin (R) ile değişeceğini ifade eder. Bu çalışma prensibi, rastgele yürüyüş

algoritmasının sürekli tekrarlanmasıyla işler. Her iterasyonda, farklı pozisyonlar ve amino asit değişiklikleri değerlendirilir. Süreç, toplam vektörünü kullanarak devam eder ve SARS-CoV-2'deki evrimsel dinamikler bu şekilde modellenir. Mutasyon hızı hesaplaması bölümünde de belirtildiği gibi, virüsün yılda bir mutasyon geçireceği düşünülerek devam edilmiştir. Tekrar edilecek iterasyon sayısının belirlenmesi için farklı sayıda tekrarlar gerçekleştirilmiş ve çıkan mutasyon sonuçları arasındaki ilişki analiz edilmiştir. Şekil 2.8'de sunulan sonuçlara göre, tahmin sayısının tahmin edilen bölgeler üzerinde belirgin bir etkisi olmadığı; her bir tahmin sayısının neredeyse aynı bölgelere işaret ettiği görülmektedir. Zaman faktörü de dikkate alındığında, tahmin sayılarının mutasyon üzerinde belirgin bir etkisinin olmadığı için tahmin sayısı seçilebilecek maksimum tahmin sayısı olan 2000 olarak seçilmiştir.



Şekil 2.8: Rastgele yürüyüş algoritması ile elde edilen farklı tahmin sayıları arasındaki korelasyon matrisi. Her hücrede iki farklı tahmin seti arasındaki korelasyon katsayısı gösterilmektedir. Renk skalası, korelasyon değerlerini ifade etmektedir: koyu kırmızı tonlar 1'e yakın yüksek korelasyonları, mavi tonlar ise %97'in altındaki korelasyonları temsil etmektedir.

Rastgele yürüme algoritmasının ikinci parametresi olan amino asit frekansı, bu tez kapsamında 4 farklı şekilde alınmıştır. Bunlar; doğal frekanslar, gözlemlenen frekanslar, referans frekansları ve dönüşüm frekanslarını ifade etmektedir. Doğal frekanslar, amino asitlerin kodon tablosundaki dağılımlarından hesaplanmıştır. Her bir amino asidi ifade eden kodonlar sayılmış ve bu verilerden bir frekans değeri elde

edilmiştir. Gözlemlenen frekans ise kullanılan veri setinden elde edilen frekanstır. Bu frekans, veri setinde bulunan her bir amino asidin kaç kez geçtiğinin sayılmasıyla hesaplanmıştır. Son olarak referans frekans, referans olarak kullanılan sekans üzerinden elde edilen frekanstır. Bu yöntemde, referans sekans içerisinde bulunan amino asitler sayılarak frekans değerleri hesaplanmıştır (Çizelge 2.4). Bu çalışma kapsamında, Wuhan'da ortaya çıkan ve EPI_ISL_402124 koduyla tanımlanan referans sekans kullanılmıştır (Okada vd., 2020).

Çizelge 2.3: Wuhan referans sekansına (EPI_ISL_402124) ait kodonların ve bunlardan türetilen amino asitlerin dağılımı.

A	R	N	D	C	Q	E
GCA 27	AGA 20	AAT 54	GAC 19	TGT 28	CAG 16	GAG 14
GCT 42	CGT 9	AAC 34	GAT 43	TGC 12	CAA 46	GAA 34
GCG 2	AGG 10					
GCC 8	CGG 2					
	CGC 1					
G	H	I	L	K	M	F
GGT 47	CAT 13	ATA 18	CTT 36	AAA 38	ATG 14	TTT 59
GGG 3	CAC 4	ATT 44	TTA 28	AAG 23		TTC 18
GGC 15		ATC 14	TTG 20			
GGA 17			CTA 9			
			CTC 12			
			CTG 3			
P	S	T	W	Y	V	X
CCA 25	TCT 37	ACA 40	TGG 12	TAC 14	GTT 48	TAA 1
CCC 4	AGT 17	ACC 10		TAT 40	GTC 21	
CCT 29	TCC 12	ACT 44			GTG 13	
	TCA 26	ACG 3			GTA 15	
	TCG 2					
	AGC 5					

Çizelge 2.4: Referans sekanstan elde edilen frekans

Amino asit	Amino asit Sayısı	Amino asit Frekans
A	79	0.062
R	42	0.033
N	88	0.069
D	61	0.048
C	40	0.035
Q	62	0.049
E	48	0.038
G	83	0.065
H	17	0.013
I	76	0.06
L	108	0.085
K	61	0.048
M	14	0.011
F	77	0.061
P	58	0.046
S	99	0.078
T	97	0.077
W	12	0.009
Y	54	0.043
V	97	0.076

Kodon tablosu dönüşüm frekansı, kodon tablosundaki amino asitler dikkate alınarak, mutasyonla hangi amino asidin hangi amino aside değiştiği hesaplanarak elde edilmiştir. Burada, üçlü kodonlardaki bir, iki ve üç mutasyon durumları göz önünde bulundurularak sayım yapılmıştır (Şekil 2.3).

2.4.3 Tahminlerin değerlendirilmesi

Yapılan hesaplamalar sonucunda elde edilen tahminlerin test edilmesi gerekmektedir. Geliştirilen modelin performansını değerlendirebilmek amacıyla, bu tezde veriler 2022 öncesi ile 2022 ve sonrası olarak ikiye ayrılmıştır. 2022 sonrası gelişen Omicron varyantı tahmin edilmeye çalışılmıştır. 2022 öncesi veriler eğitim seti olarak, 2022 ve sonrası veriler ise test verisi olarak kullanılmıştır. Eğitim veri seti, tüm hesaplamalarda kullanılmış ve doğrulama veri setindeki mutasyonlar ile karşılaştırmalar yapılmıştır; böylece modelin performansı test edilmiştir. Elde edilen tahminler iki farklı şekilde incelenmiştir. İlk olarak, skor fonksiyonları ve entropi kullanılarak elde edilen pozisyon tahminlerinin karşılaştırması yapılmıştır. Bu aşamada, tahmin edilen pozisyonların doğruluk oranı hesaplanmıştır. Ardından, doğru tahmin edilen

pozisyonlar üzerinden, bu pozisyonlarda amino asitlerin doğru tahmin edilme oranı belirlenmiştir. Bu oran, aşağıda verilen formül kullanılarak hesaplanmıştır.

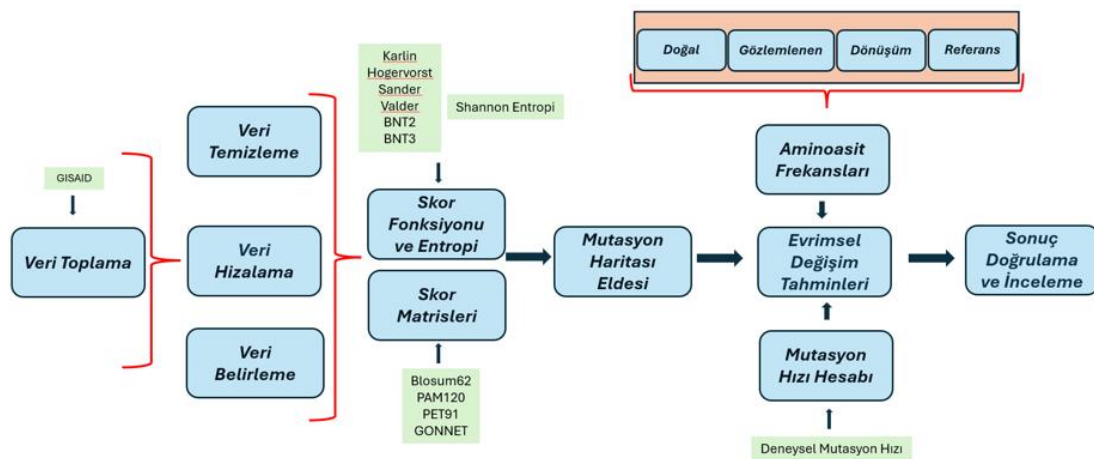
$$\text{Doğruluk Oranı} = \frac{\text{Doğru tahminler}}{\text{Doğru tahminler} + \text{Yanlış tahminler}} \quad (2.30)$$

Bu formül, tahmin edilen verilerin doğruluk oranını belirleyerek, doğru ve yanlış tahminlerin oranlarını net bir şekilde ortaya koymaktadır.



3. MODELLEME SONUÇLARI VE YORUMLAR

Bu tez kapsamında matematiksel modelleme bölümünde anlatılan yöntemler ve teknikler kullanılarak, SARS-CoV-2'nin S proteinine ait mutasyonların tahmin edilmesi amaçlanmıştır. Bu amaç doğrultusunda, öncelikle S proteininin sekans verileri bir önceki bölümde anlatıldığı gibi GISAID (Url-8) veri tabanından elde edilmiştir. Bu sekansların yıllara ve aylara göre dağılımı analiz edilmiş ve tüm veri setinin kullanılması zaman açısından uygun olmadığı için, rastgele seçilmiş veri setleri oluşturulmuştur. Tahmin algoritması için en uygun seti seçmek adına, veri setleri arasındaki benzerlikler değerlendirilerek filogenetik ağaç oluşturulmuş ve bu ağaç üzerinden farklı küçük alt setler elde edilmiştir. Bu alt setlere skor fonksiyonları uygulanarak, farklı skora matrisleri (*BLOSUM62*, *PAM120*, *PET91*, *GONNET*) ile mutasyona uğrama skoru değerleri hesaplanmıştır. Bu değerlere ek olarak her set için pozisyona özel entropi hesabı yapılmıştır. Ardından, bu skor değerleri ve 4 farklı amino asit frekansı (2.4.2.1) kullanılarak gelecekte oluşabilecek mutasyonları tahmin etmeye yönelik analizler gerçekleştirilmiştir. Tez kapsamında izlenen adımların şeması Şekil 3.1'de verilmiştir.



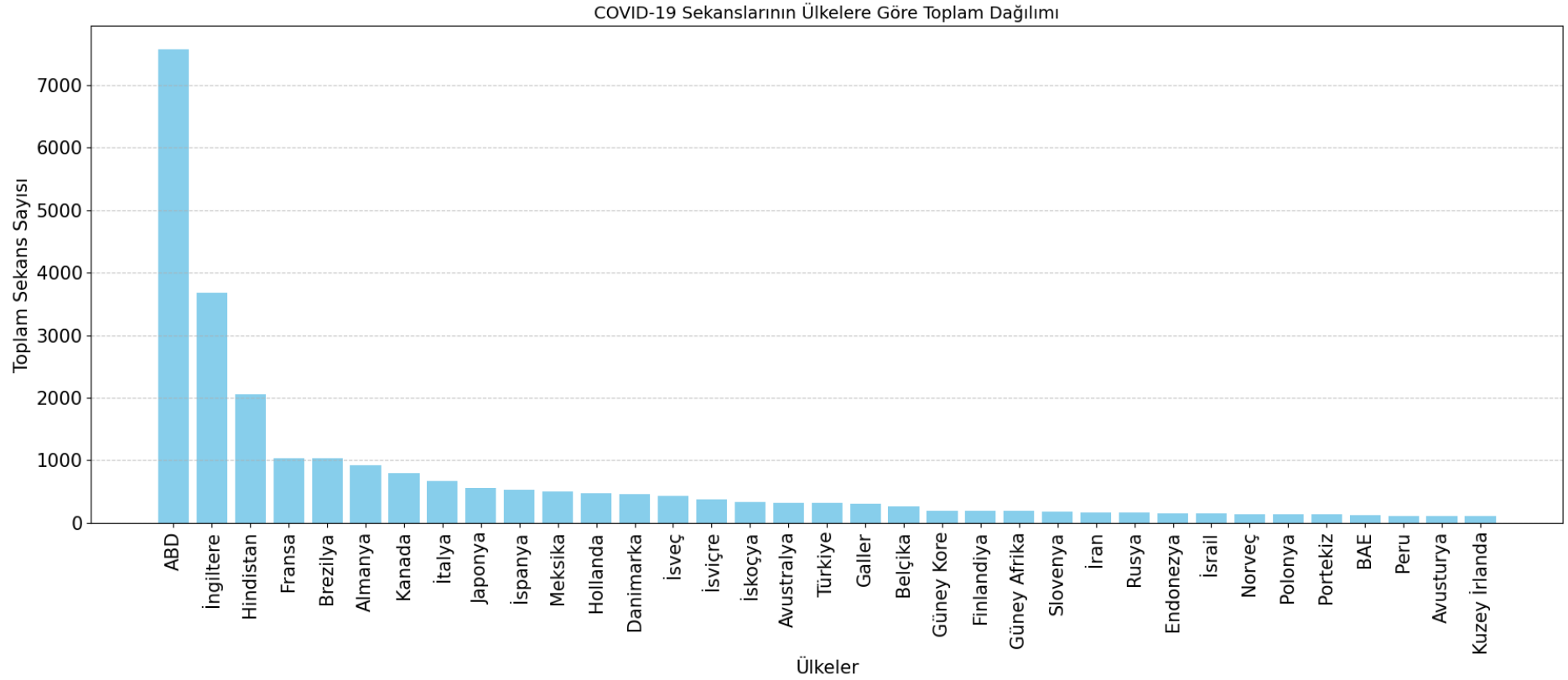
Şekil 3.1: Çalışma adımları şeması.

3.1 SARS-CoV-2 Veri Setinin Analizi

SARS-CoV-2'nin S proteininin gelecekteki mutasyonlarını tahmin edebilmek için geliştirilecek yöntem ve algoritmaların temelini, ilk olarak incelenecek veri setinin doğru şekilde hazırlanması, analiz edilmesi ve detaylı bir şekilde incelenmesi oluşturmaktadır. GISAID (Url-8) veri tabanından elde edilen sekanslar, bugüne kadar rapor edilmiş S protein sekanslarını içermektedir. İncelenen protein sekans verileri, 2019 yılına ait kayıtların oldukça sınırlı olduğunu, ancak 2021 yılına doğru veri sayısında belirgin bir artış olduğunu göstermektedir. Bu durumun temel sebebi, SARS-CoV-2 virüsünün 2019 yılının sonlarına doğru ortaya çıkmış olmasıdır. Sekansların ülkelere göre dağılımları incelendiğinde ABD'den gelen sekanslar veri setinde belirgin bir baskınlık göstermektedir. ABD'yi sırasıyla İngiltere ve Hindistan takip etmektedir. Bu durum, seçilen alt veri kümelerinin ülkelere göre dağılımında ABD'nin en yüksek orana sahip olduğunu, İngiltere ve Hindistan'ın ise önemli bir paya sahip olduğunu göstermektedir. Bu dağılımda, ülkelerin gelişmişlik seviyeleri etkili olmuş olabilir. Örneğin, ABD ve İngiltere gibi gelişmiş ülkelerde laboratuvar altyapılarının ileri seviye olması ve sekanslama çalışmalarına daha fazla yatırım yapıyor olması, bu ülkelere elde edilen sekans sayısının diğer ülkelere kıyasla yüksek olması ile sonuçlanmış olabilir. Hindistan'ın bu listede yer almasının sebebi ise, ülkedeki yüksek vaka sayıları nedeniyle sekanslama çalışmalarının artmış olması olabilir (Şekil 3.3). Sekans verilerinin artan hacmi, eğitim ve tahmin süreçlerinin doğruluğunu tespit edebilmek için verilerin 2022 öncesi ve 2022 sonrası olarak iki gruba ayrılmasını gerektirmiştir. Bu bağlamda, 2022 öncesine ait veriler eğitim seti, 2022 ve sonrası veriler ise test seti olarak belirlenmiştir. Elde edilen eğitim verisi kullanılarak, 2021 yılının sonunda ortaya çıkan Omicron varyantı ve bu varyantın mutasyonları üzerinde doğrulama çalışmaları gerçekleştirilmiştir (Şekil 3.2). Bu sayede, modelin yeni varyantları ne ölçüde doğru tahmin edebildiği değerlendirilmiştir.

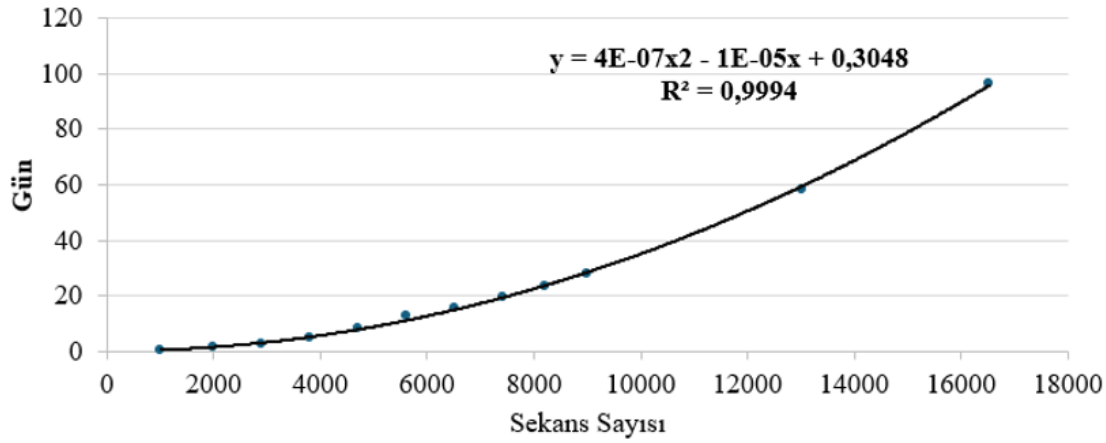


Şekil 3.2: Veri seti dağılımı.



Şekil 3.3: COVID-19’a ait sekansların bazı ülkelere göre toplam dağılımı. Grafikte, GISAID veri tabanı üzerinden elde edilen sekans sayıları ülkelere göre sınıflandırılmıştır. ABD, İngiltere ve Hindistan gibi ülkeler yüksek sayıda sekans paylaşımı ile ön plana çıkarken; veri katkısı daha sınırlı olan ülkeler grafik sonunda yer almaktadır. Bu dağılım, genomik gözetim kapasitesi ve uluslararası veri paylaşımı açısından ülkeler arasındaki farklılıkları yansıtmaktadır.

Sekanslara uygulanacak analiz yöntemlerinin hesaplama yükü göz önünde bulundurulduğunda, toplamda 184744 adet SARS-CoV-2 S protein sekansının her biri için altı farklı mutasyon skorlama fonksiyonu ve dört farklı evrimsel skor matrisi kullanılarak hesaplama yapılması hedeflenmiştir. Ancak bu işlem, kullanılan donanım olan 12 çekirdekli CPU ve 64 GB RAM kapasitesine sahip bir bilgisayarda yaklaşık 13.648,34 gün (yaklaşık 37 yıl) sürecektir. Mevcut hesaplama kaynakları ve tez süresinin sınırlılığı göz önünde bulundurulduğunda, tüm protein sekansların işlenmesi tez kapsamında mümkün değildir. Sekanslar ile işlem süresi arasındaki ilişki incelendiğinde, Şekil 3.4'te gösterildiği gibi bir grafik elde edilmiştir. Sekans sayısının artmasıyla birlikte işlem süresi de artmaktadır.



Şekil 3.4: Altı farklı skorlama fonksiyonu ve dört skorlama matrisi kullanılarak yapılan analizlerde, işlenen sekans sayısının işlem süresi (gün cinsinden) üzerindeki etkisini gösteren grafik. Sekans sayısı arttıkça geçen süre doğrusal olmayan şekilde artmakta olup, ikinci dereceden polinomsal bir modelle ($R^2 = 0.9994$) yüksek doğrulukla temsil edilmiştir.

Bu hesaplama, tüm işlemleri bir arada çalıştırmanın ne kadar büyük bir zaman gerektirdiğini göstermektedir. Bu tür uzun süreli hesaplamaların gerçekleştirilmesinin kullanılan bilgisayar donanımı sebebiyle zaman ve kaynak açısından mümkün olmadığı değerlendirilerek tez kapsamında veri setlerinin belirlenmesinde daha hızlı sonuç alınabilecek alternatif bir yaklaşım benimsenmiştir. İlk adım olarak, verinin zaman içindeki dağılımlarını daha iyi anlaşılması adına 3 aylık periyotlarla sekans sayıları incelenmiştir (Çizelge 2.1). Bu analiz, verinin zamana göre nasıl dağıldığını ve sekans yoğunluklarının hangi dönemlerde arttığını göstermektedir. Bu deneyler, farklı büyüklükteki veri setlerinde rastgele seçimin etkisini test etmek ve en uygun parametreleri belirlemek amacıyla gerçekleştirilmiştir.

Çalışma süreci şu şekilde ilerlemektedir: öncelikle her periyottan seçilecek sekans sayısı belirlenmektedir. Örneğin, her periyottan 600 sekans seçilmesi planlandığında, 2019 yılından 3 sekans, 2020 yılı için ilk periyottan 600, ikinci periyottan 600, üçüncü periyottan 600, dördüncü periyottan 600 ve bilinmeyen kısımdan 195 sekans seçilerek toplamda 2595 sekans elde edilmektedir. 2021 yılı için ise 5 farklı periyottan 600'er sekans seçilerek toplamda 3000 sekans elde edilmektedir. Böylece oluşturulan veri seti, toplamda 5598 sekans içermektedir. Bu işlem, farklı sayılarda tekrar edilerek çeşitli alt veri kümeler oluşturulmakta ve analiz edilmektedir. Deneylerin farklı tekrar sayıları ve sekans sayılarına göre yapılması en uygun veri veya veri setlerini seçmek amacıyla yapılmıştır. Süre hesaplamaları, Şekil 3.3'teki grafik göz önüne alınarak yapılmıştır. Tekrar sayısı ve sekans sayısının artmasıyla birlikte deney sürelerinde doğrusal bir artış gözlemlenmiştir (Çizelge 3.1).

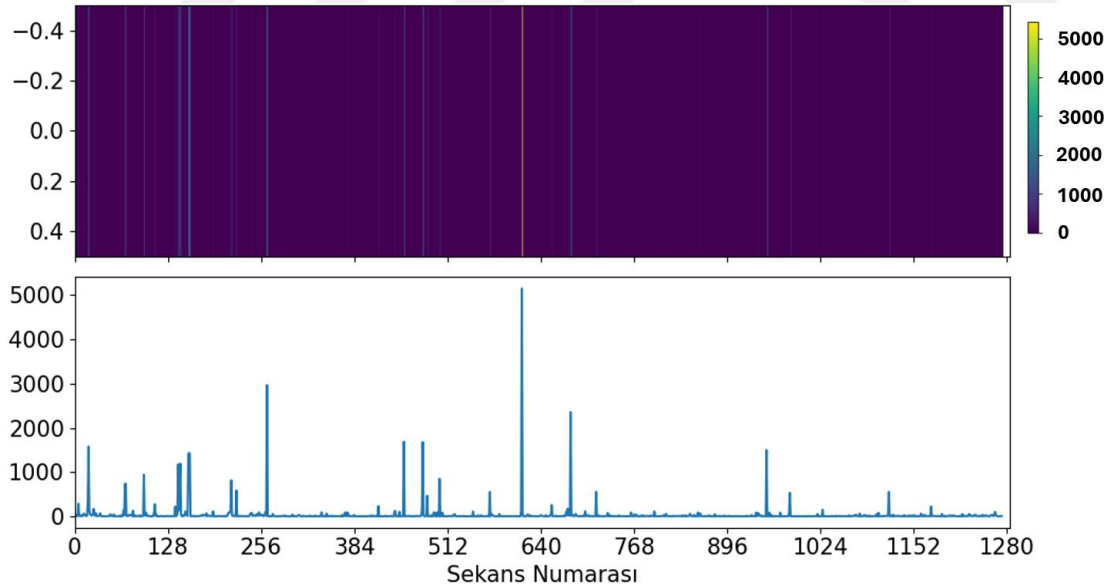
Çizelge 3.1: Sekans ve zaman dağılımı.

1 Skor Matrisi x 1 Skor Fonksiyonu					
Periyottaki Sekans Sayısı	Geçen Süre(gün)	5 Tekrar(gün)	10 Tekrar (gün)	15 Tekrar (gün)	20 Tekrar (gün)
100	0.015	0.075	0.15	0.225	0.3
200	0.064	0.32	0.64	0.96	1.28
300	0.119	0.595	1.19	1.785	2.38
400	0.207	1.035	2.07	3.105	4.14
500	0.351	1.755	3.51	5.265	7.02
600	0.528	2.64	5.28	7.92	10.56
700	0.646	3.23	6.46	9.69	12.92
800	0.822	4.11	8.22	12.33	16.44
900	0.98	4.9	9.8	14.7	19.6
1000	1.163	5.815	11.63	17.445	23.26
1500	2.429	12.145	24.29	36.435	48.58

Özellikle 600 sekans 20 tekrar, deneylerin yürütülmesinde gerekli zaman açısından uygun bir değer olarak belirlenmiştir. Bu periyot için süreler detaylı şekilde verilmiştir. 600 sekans seçilerek yürütülen deneylerde, sekanslar arasındaki farklılıkların ve benzerliklerin analiz edilebilmesi için setlerin mutasyonları incelenmiş, mutasyon sayılarına odaklanılmış ve setler arasında korelasyon analizi gerçekleştirilmiştir.

Mutasyonların İncelenmesi: 20 adet ve her periyottan 600 sekans içeren veri setlerindeki mutasyonlar referans sekansa (Wuhan, EPI 402124) göre belirlenmiştir. Bu mutasyonların her birinin verideki sıklığı ve çeşitliliği hesaplanmıştır (Şekil 3.7).

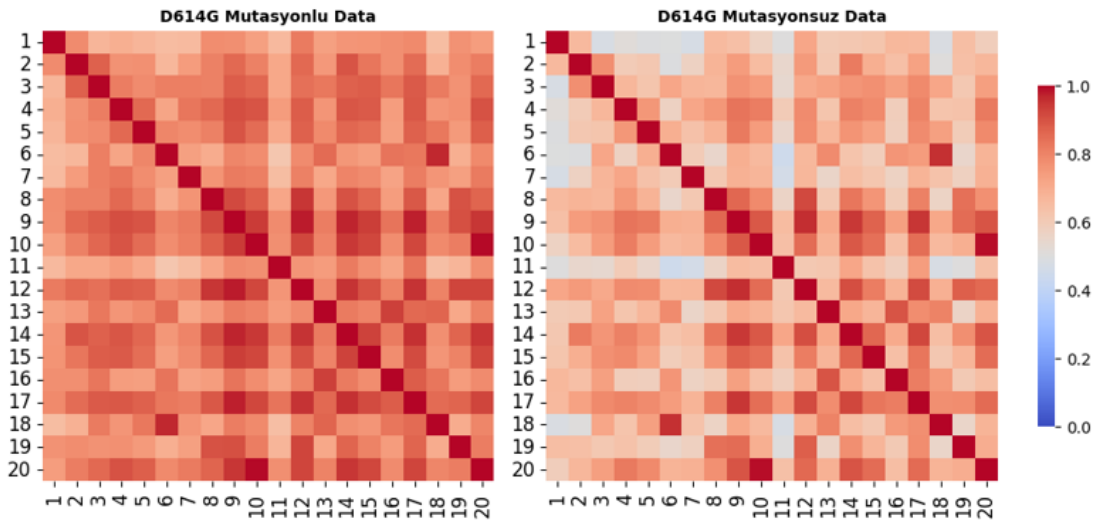
Mutasyon Sayılarının İncelenmesi: Her veri içindeki pozisyon bazında mutasyon sayısı sayılmıştır (Şekil 3.7). Bu, mutasyonların frekansları ve çeşitliliği analiz edilmiştir. Her set için yapılan çalışmada, Şekil 3.5'teki gibi mutasyon pozisyon dağılımı grafikleri elde edilerek setlerdeki mutasyon dağılımları incelenmiştir. Bu grafik bir ısı haritası ve çizgi grafiğinden oluşmaktadır. Isı grafiğinde, mutasyon sayısı fazla olan bölgeler sarı renkte az sayıda mutasyon içeren bölgeler lacivert renkte olacak şekilde gösterilmektedir. Şekil 3.5'teki çizgi grafiği ise pozisyonlarda toplam görülen mutasyon sayısını vermektedir. Örneğin 614 pozisyonunda sarı renk ile belirgin olduğu görülmektedir. Çizgi grafiğine bakıldığında bu pozisyonda setlerde toplam 5000 üzerinde mutasyon görülmektedir. Bu durum, 614. pozisyonun veri setleri arasında baskın olduğunu ortaya koymaktadır. 614. pozisyondaki baskınlığın temel nedeni, pandeminin ilk dönemlerinde S proteini üzerinde ortaya çıkan D614G mutasyonunun SARS-CoV-2'nin ACE2 reseptörüne bağlanma yeteneğini artırarak virüsün bulaşıcılığını önemli ölçüde yükseltmesidir. Bu mutasyonun sağladığı avantaj sayesinde virüs, küresel ölçekte baskın hale gelmiş ve sonraki tüm varyantlarda bu mutasyon yaygın olarak gözlemlenmiştir (Arora, Pöhlmann, & Hoffmann, 2021; Korber vd., 2020).



Şekil 3.5: SARS-CoV-2 S proteinine ait mutasyon pozisyonlarının frekans dağılımı. Üst grafikte, her bir pozisyonda gözlemlenen mutasyon sıklığı ısı haritası şeklinde görselleştirilmiştir; renk skalası yüksek frekansları (örneğin 614. pozisyon) sarıya doğru vurgulamaktadır. Alt panelde ise mutasyon frekansları pozisyon bazında sayısal olarak gösterilmiştir. Grafik, belirli pozisyonların (örneğin 614) diğerlerine göre daha sık mutasyona uğradığını ve bunların evrimsel baskı altında olabileceğini ortaya koymaktadır.

Korelasyon Analizi: Korelasyon analizleri, her tekrarda rastgele seçilmiş 600 sekans içeren veri setleri arasındaki ilişkileri anlamak amacıyla yapılmıştır. Bu analizler, rastgele seçilen setler arasında bir ilişki veya bağlantı olup olmadığını belirlemeyi amaçlamaktadır.

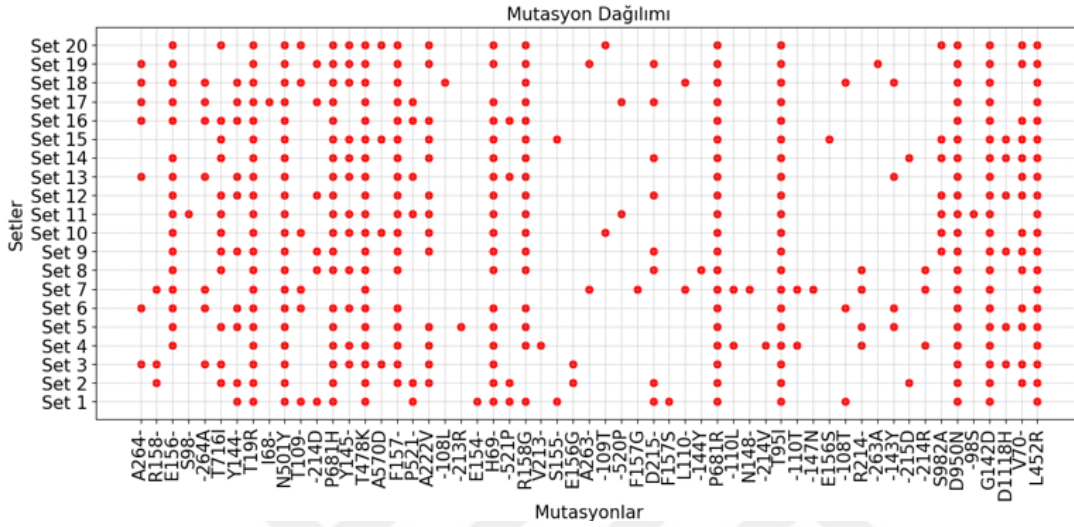
İlk aşamada yapılan analizler, rastgele seçilmiş veri setlerindeki mutasyonların genel dağılım üzerindeki etkisinin minimal olduğunu göstermiştir. Korelasyon değerlerinin büyük çoğunluğunun %90'ın üzerinde çıkması, seçilen veri setlerinin birbirine oldukça benzer olduğunu ortaya koymuştur (Şekil 3.6). Ancak bu sonuç, D614G mutasyonunun veri setindeki baskınlığıyla ilişkilendirilebilir. Seçilen 20 set, incelendiğinde toplam 5.598 sekans içeren her setin 5000'den fazlasında D614G mutasyonunun bulunması, korelasyon analiz sonuçlarına ağır bir etki yapabileceğini işaret eder. Bu nedenle, ikinci bir korelasyon analizi yapılmış ve bu aşamada D614G mutasyonu setlerden çıkarılmıştır. Yeni analizle, setler arasındaki ilişkiler daha net şekilde incelenebilmiştir. Yapılan analizler göstermiştir ki, baskın mutasyonu (D614G) veri setinden çıkardıktan sonra dahi, diğer mutasyon dağılımları için bile (Şekil 3.6) korelasyon değerleri yüksekliklerini korumuştur.



Şekil 3.6: D614G mutasyonunun, farklı veri setleri arasındaki korelasyon yapısına etkisini gösteren ısı haritaları. Sol panel, D614G mutasyonu taşıyan sekanslara ait korelasyon matrisini; sağ panel ise bu mutasyonu taşımayan sekanslara ait korelasyon matrisini göstermektedir. Renk skalası, iki set arasındaki korelasyon katsayısını ifade etmektedir (1: yüksek korelasyon, 0: düşük korelasyon).

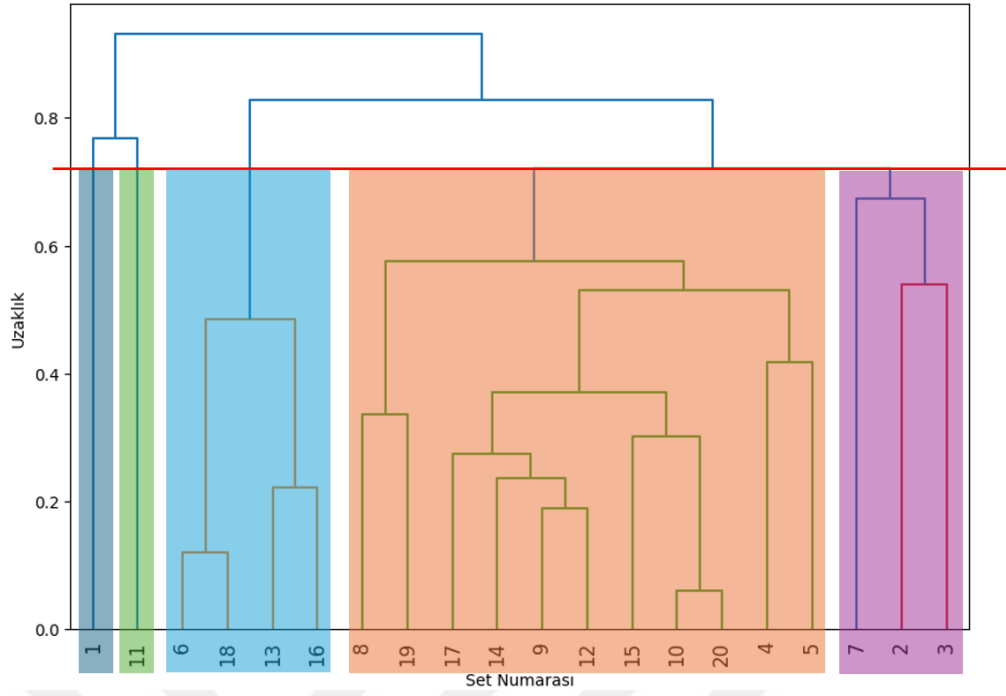
Örneğin, başlangıçta %90 korelasyona sahip olan setler, %70–%80 aralığında bir korelasyonla büyük ölçüde korunurken; başlangıçta %60 korelasyon değerine sahip setlerde ise korelasyon oranlarında belirgin bir değişim gözlenmemiştir. Bu durum,

veri setlerinin birbiriyle uyumlu olduğunu ve genel dağılımın baskın olabilecek mutasyonların etkisi olmadan da temsil edilebileceğini ortaya koymaktadır. Sonuç olarak, rastgele seçimlerin, ana veri setinin dağılımını çok farklı şekilde etkilemediği belirlenmiştir.



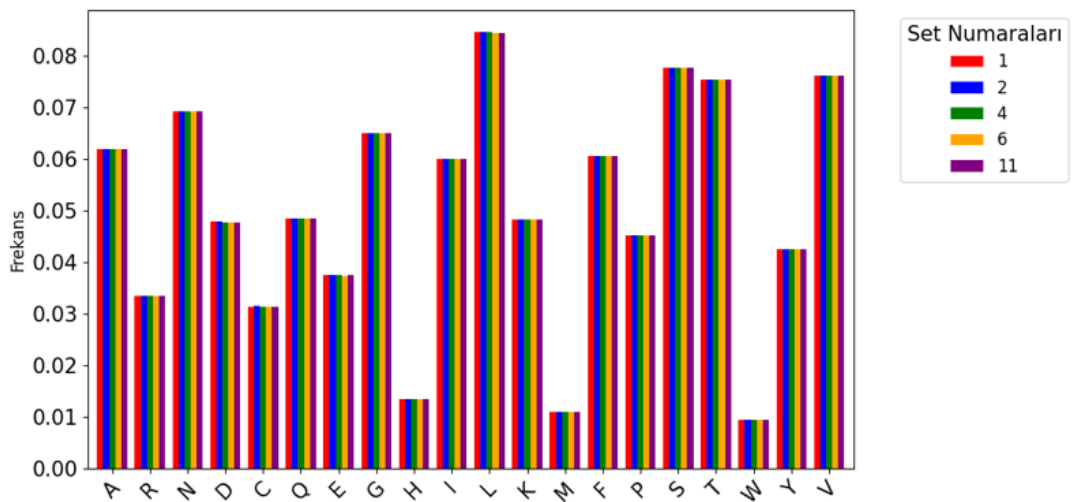
Şekil 3.7: Farklı setlerde gözlemlenen S proteini mutasyonlarının dağılımını gösteren grafik. Yatay eksen mutasyon pozisyonları, dikey eksen ise veri setleri yer almaktadır. Kırmızı noktalar, ilgili mutasyonun ilgili sette gözlemlendiğini ifade etmektedir. Örneğin T95I mutasyonu tüm setlerde görülmektedir.

Zaman kısıtlılığı nedeniyle tüm sekans setleri üzerinde çalışmak mümkün olmadığından, sekans setleri arasındaki benzerlikler analiz edilerek bir filogenik ağaç (dendrogram) oluşturulmuştur. Bu filogenik ağaç, protein sekanslarının birbirleriyle olan ilişkilerini görselleştirmekte ve gruplamaktadır. Grafik, sekansların ne kadar benzer olduğunu gösterirken, sekansları birbirinden ayırtmak için bir referans noktası seçebilmeyi mümkün kılmaktadır. Bu filogenik ağaç, hiyerarşik kümeleme yöntemi ile oluşturulmuştur. Rastgele seçilen setlerin mutasyon analizlerinden elde edilen bilgiler dikkate alınarak sekanslar arasındaki benzerliklere göre gruplama yapılmıştır (Şekil 3.8). Dalların birleşme noktaları, sekanslar arasındaki uzaklığı temsil etmektedir. Farklı renklerle belirtilen gruplar, sekansların benzerlik düzeylerine göre sınıflandırıldığını göstermektedir. Bu setleri ayırtmak için Şekil 3.8'de kırmızı çizgi ile gösterilen 0.7 eşik değeri seçilmiş ve bu değer referans alınarak benzerliklere göre gruplar oluşturulmuştur.



Şekil 3.8: 2019 ve 2021 yılına ait protein sekanslarından türetilen 20 setin, benzerlik ilişkilerine göre oluşturulan ve 5 gruba ayrılan filogenetik ağacı. Y ekseninde kümeler arası uzaklıklar, kırmızı çizgide ise ayırım eşiği gösterilmiştir.

Her bir renkle temsil edilen gruptan birer örnek seçilerek analizlerin daha etkili ve hızlı bir şekilde gerçekleştirilmesi hedeflenmiştir. Eşik değerine göre her gruptan seçilen alt veri setleri şu şekilde belirlenmiştir: 1 ve 11 numaralı gruplar bağımsız birer alt set olarak ele alınırken, diğer gruplar için belirlenen alt setler sırasıyla 6, 4 ve 2 olarak seçilmiştir (Şekil 3.9).



Şekil 3.9: Farklı veri setlerine ait gözlemlenmiş amino asit frekanslarının karşılaştırılması. Grafik, 1, 2, 4, 6 ve 11 numaralı setlerde yer alan amino asitlerin frekans dağılımını göstermektedir.

Skorlama fonksiyonlarının doğru işlemi yapabilmesi adına sekans setleri kendi aralarında Clustal Omega (Url-9) ile hizalanmıştır. Hizalanmış setler incelendiğinde, setler arasında farklı bölgelerde silinmeler (delesyon) olduğu gözlemlenmiştir (Çizelge 3.2). Bu silinmelerin veri setine işlenebilmesi için sekans hizalama (alignment) yöntemleri kullanılarak incelenmiş ve model üzerindeki etkileri değerlendirilmiştir. Silinmelerin matematiksel modelleme hesaplamalarında başarımı düşürmesi nedeniyle, bu veriler veri setlerinden çıkarılmıştır.

Çizelge 3.2: Sekanslardaki silinme miktarı.

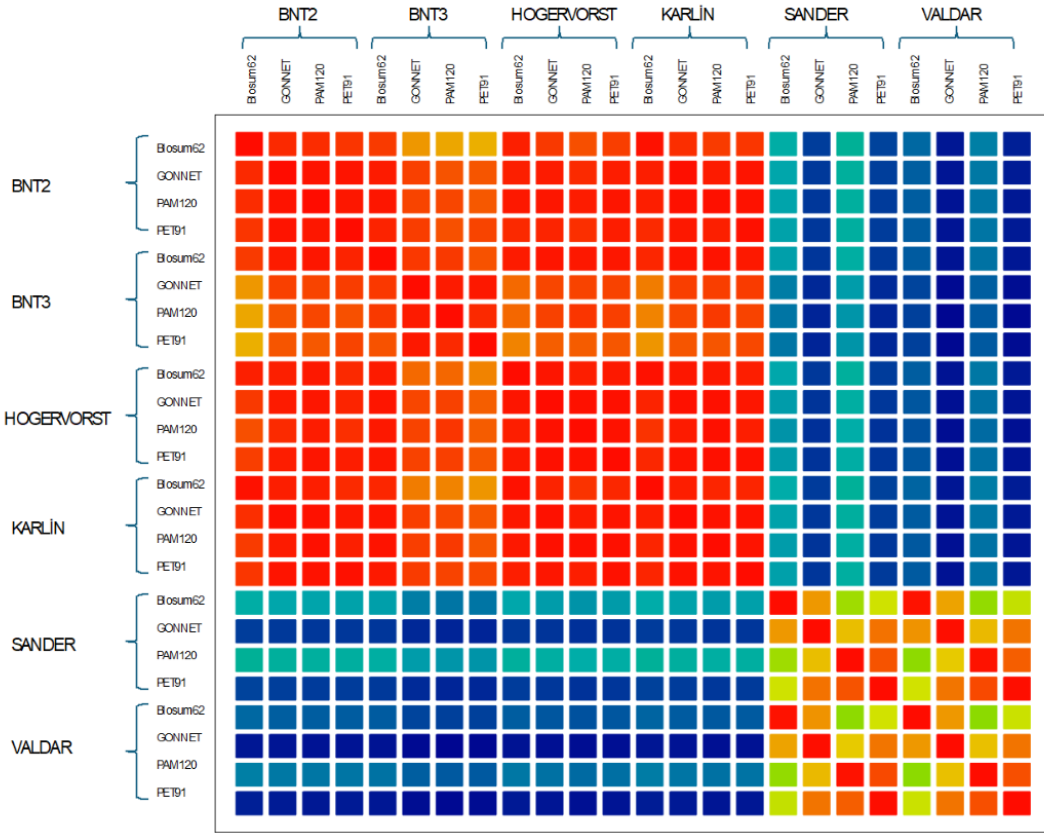
<i>Set Numarası</i>	<i>Silinme Sayısı</i>
1	25
2	16
4	27
6	25
11	34

Bu kaymalar, hesaplamalarda farklı anlamlara yol açabileceğinden, bu tez kapsamında silinmeler dikkate alınmamış ve tüm setlerin hizalanma sonucunda referans sekans ile aynı uzunlukta (1273 nükleotid) olacak şekilde düzenlenmiş ve kabul edilmiştir.

3.2 Metotlar Arası Korelasyon Değerlendirmesi

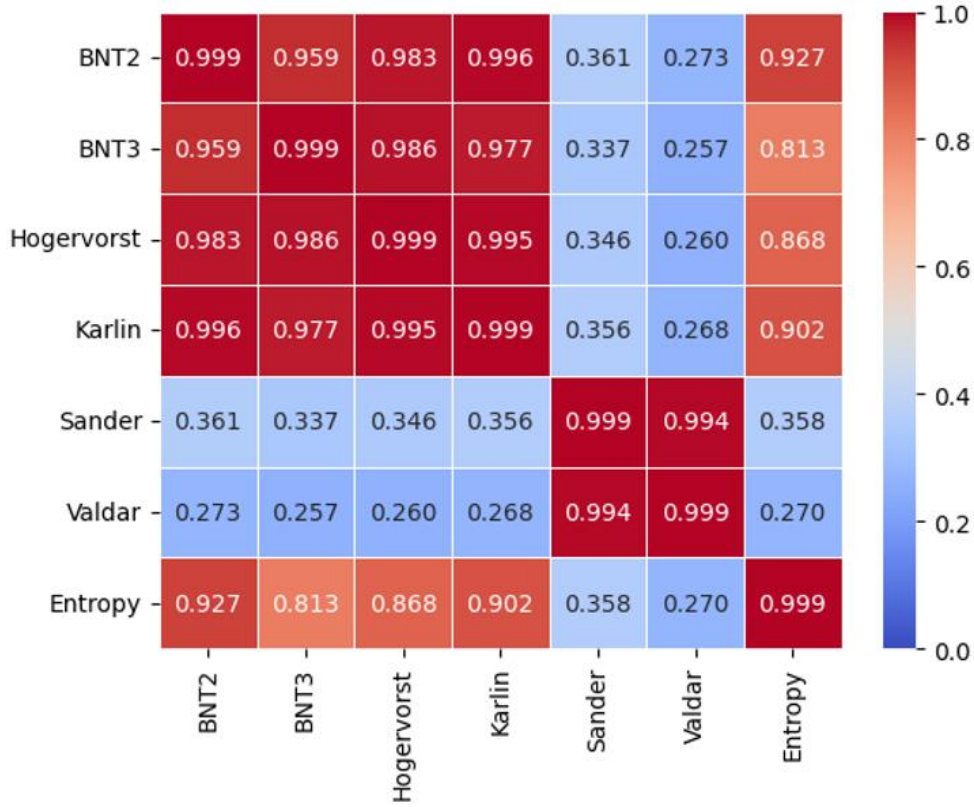
Bu tez kapsamında, kullanılan skor fonksiyonlarının ve veri setlerinin birbiriyle olan korelasyonu incelenmiş; setler ile skor fonksiyonları arasındaki benzerlikler ve farklılıklar detaylı bir şekilde ortaya konulmuştur. Toplam 5 farklı veri seti için 6 skor fonksiyonu ve 4 skor matrisi kullanılarak toplamda 120 farklı hesaplama yapılmıştır. Öncelikle, tüm skor fonksiyonlarının ve skor matrislerinin birbirleri ile olan korelasyon haritası Şekil 3.10'da gösterilmiştir. Bu şekilde renkler korelasyonu ifade etmektedir, kırmızı ile ifade edilen ikililer, birbirine çok yüksek oranda benzeyen çiftleri temsil ederken, mavinin hâkim olduğu bölgeler düşük korelasyonu göstermektedir. Verilen korelasyon haritasına bakıldığında sonuçlar arasında bir örüntü olduğu açıkça görülmektedir. Burada, BNT2, BNT3, Hogervorst ve Karlin kendi aralarında yüksek korelasyon gösterirken, Valdar ve Sander bu gruptan ayrılarak kendi aralarında korelasyon göstermiştir. Bunun sebebi, matematiksel modelleme bölümünde de anlatıldığı üzere, Sander ve Valdar'ın skorlama fonksiyonlarında

hesaplamalar sırasında belirli ağırlık değerleri kullanmaları sonucu diğer fonksiyonlardan farklı hesaplama yapılarıdır (Bölüm 3.2.2)



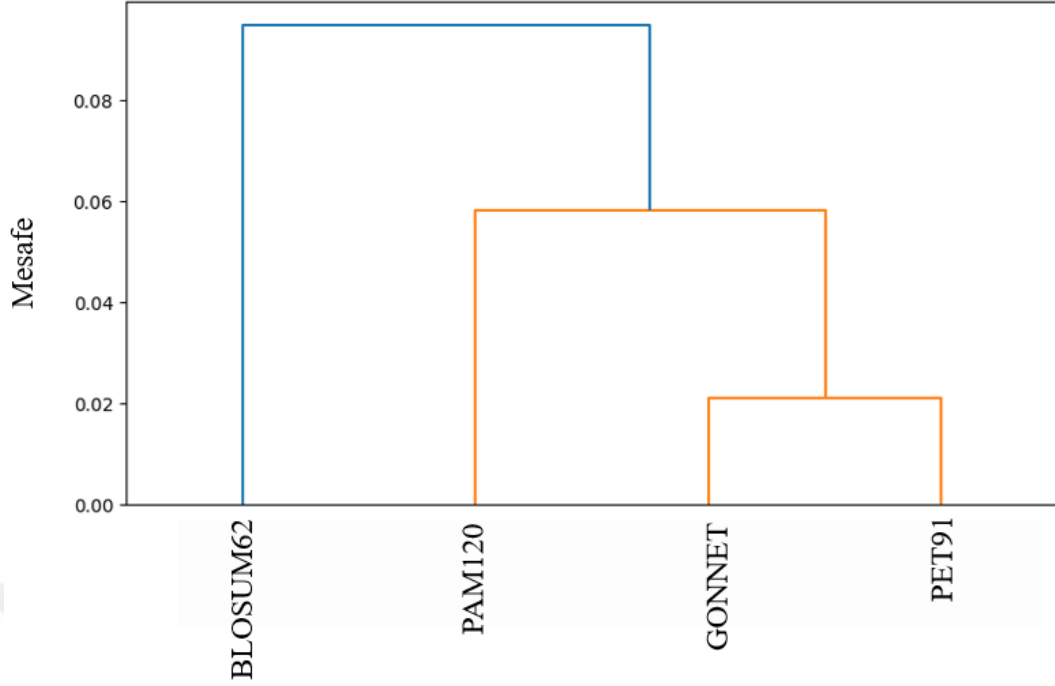
Şekil 3.10: Kullanılan skorlama fonksiyonları ve skorlama matrisleri (BLOSUM62, GONNET, PAM120, PET91) arasında hesaplanan korelasyonları gösteren ısı haritası. Grafik, her bir skorlama fonksiyonu-matris kombinasyonu ile olan ilişkisini görselleştirmektedir. Kırmızı tonlar yüksek korelasyonu, mavi tonlar ise düşük korelasyonu ifade etmektedir.

Skor fonksiyonlarına ek olarak kullanılan entropi değerleri de korelasyon açısından diğer fonksiyonlar ile incelenmiştir (Şekil 3.11). Entropi korelasyonları incelendiğinde, entropi değerlerinin BNT2, BNT3, Karlin ve Hogervorst skor fonksiyonları ile olan korelasyonlarının, Sander ve Valdar skor fonksiyonlarıyla olan korelasyonlarına kıyasla daha yüksek olduğu görülmektedir. Bu da entropi değerlerinin birinci grup (BNT2, BNT3, Karlin ve Hogervorst) skor fonksiyonlarına daha yakın bir analiz olduğunu göstermektedir.



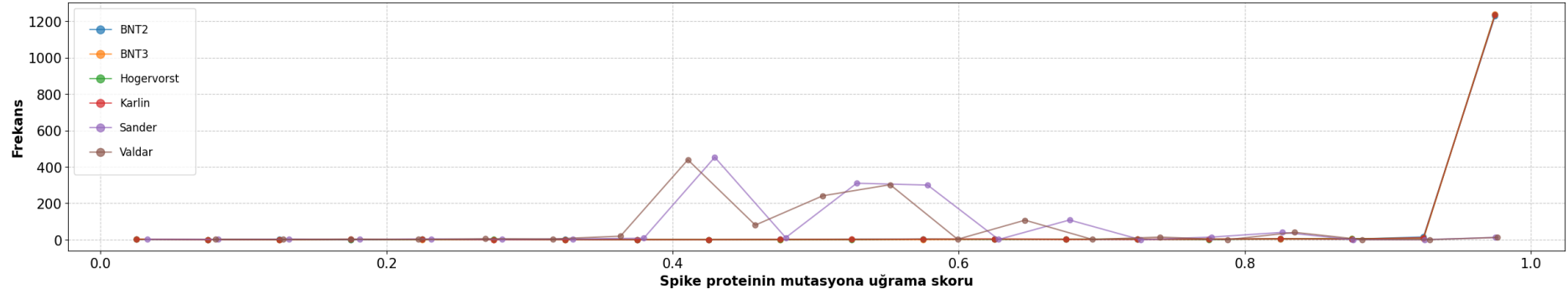
Şekil 3.11: Entropi değeri ile altı farklı mutasyon skorlama fonksiyonu (BNT2, BNT3, Hogervorst, Karlin, Sander, Valdar) arasında hesaplanan korelasyon matrisi. Renk skalası, korelasyon katsayısını ifade etmektedir; kırmızı tonlar yüksek pozitif korelasyonu, mavi tonlar ise düşük korelasyonu göstermektedir.

Yapılan bir diğer korelasyon analizi ise skor matrislerinin birbirileri ile olan benzerliklerini ortaya koymak için yapılan analizdir. Bu analiz filogenik ağaç hesaplaması ile ele alınmıştır (Şekil 3.12). BLOSUM62 skorlama matrisi, diğer matrislere kıyasla daha uzak bir mesafede konumlanmakta, dolayısıyla biyolojik prensipler açısından farklı bir temele dayandığını göstermektedir. Öte yandan, PAM120, GONNET ve PET91 matrisleri birbirine daha yakın bir mesafede gruplanmış ve benzer biyolojik varsayımlar doğrultusunda geliştirilmişlerdir. Bu durum, kullanılan veri kaynaklarının, evrimsel modellemelerin veya amino asit değişim frekansı yaklaşımlarının benzerliğinden kaynaklanabilir. Skor matrislerinin genel benzerliği ise, hepsinin ortak olarak amino asitler arasındaki değişim oranlarını temel olarak oluşturulmuş olmasından ileri gelmektedir. Böylece, farklı metodolojiler kullanılsa bile, doğadaki temel amino asit değişim dinamiklerinin korunarak modellendiği anlaşılmaktadır.

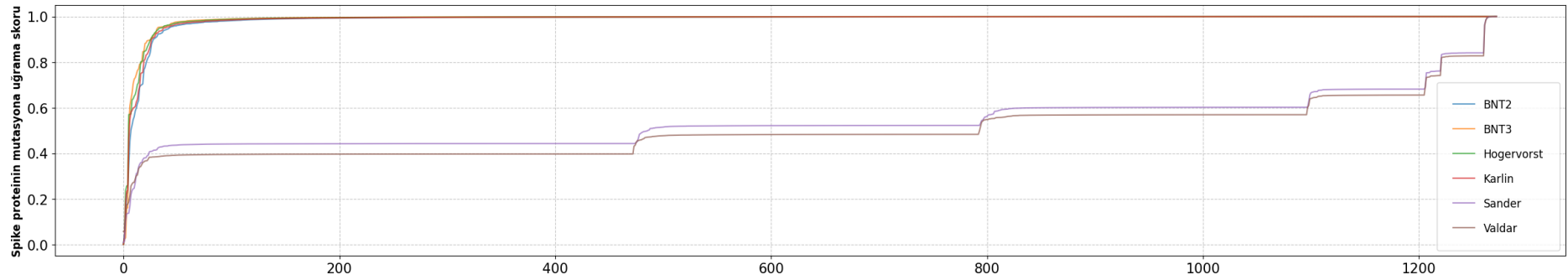


Şekil 3.12: Dört farklı skarlama matrisinin (BLOSUM62, PAM120, GONNET, PET91) benzerliklerine göre oluşturulmuş filogenik ağacı

Her set için elde edilen skorların dağılımları incelendiğinde ise korunan bölgelerin yani mutasyon skorları 1 olan değerlerin, değişen bölgelerden fazla olduğu görülmektedir (Şekil 3.13). Bu verileri iyi inceleyebilmek adına mutasyon skorları 1'den çıkarılarak mutasyona uğrama skorları (değişebilirlik değerleri) elde edilmiştir. Böylece mutasyona uğrayacak bölgelerin sayısal olarak daha fazla değer alması sağlanılmıştır. Mutasyona uğrama olasılığı yüksek olan bölgelere odaklanılması ve o bölge için tahminlerin incelenmesi S proteininde mutasyonların iyi şekilde analiz edilebilmesi için önemlidir. Analizler için seçilen pozisyonların doğru belirlenip incelenebilmesi amacıyla, mutasyona uğrama skorları sonuçları büyükten küçüğe sıralanarak ilerleyişleri karşılaştırılmıştır (Şekil 3.14). BNT2, BNT3, Karlin, Hogervorst skarlama fonksiyonlarında düzgün bir artış görülürken Sander ve Valdar fonksiyonlarından elde edilen skorlar bazı bölgelerde gruplaşmaktadır. Bu gruplaşma, tahmin algoritması için kullanılacak rastgele yürüyüş algoritması açısından, mutasyona uğrama olasılığı yüksek bölgelerin tahmin edilebilmesi için uygun çıktılar üretmemektedir. Bu yüzden ilerleyen adımlarda Sander ve Valdar skarlama fonksiyonlarının tahmin algoritmasında kullanılması uygun görülmemiştir. Buradan sonraki adımlardaki incelemelerde BNT2, BNT3, Karlin, Hogervorst ve entropi hesaplamaları ve incelemeleri dikkate alınmıştır.

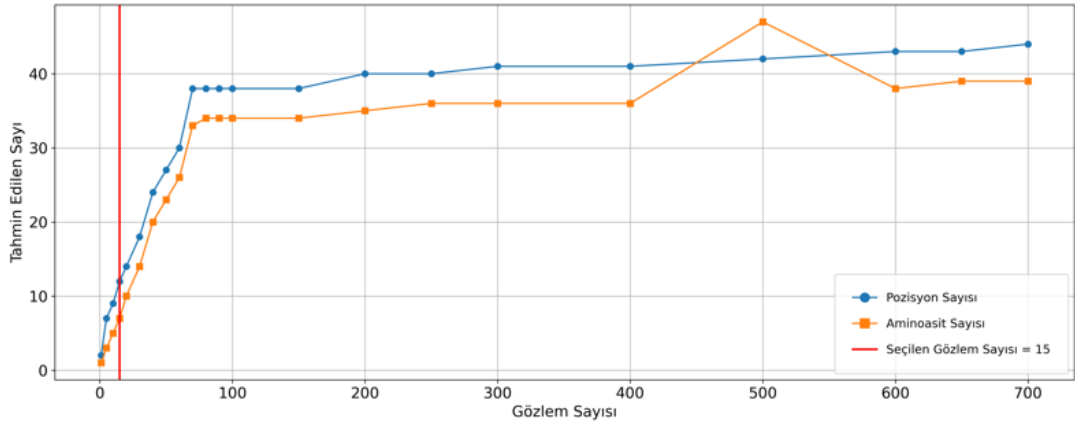


Şekil 3.13: S proteinine ait mutasyona uğrama skorlarının farklı skorlama fonksiyonlarına (BNT2, BNT3, Hogervorst, Karlin, Sander ve Valdar) göre dağılımı. Yatay eksen mutasyona uğrama skorları, dikey eksen ise bu skorların frekansı gösterilmektedir.



Şekil 3.14: S proteinine ait mutasyona uğrama skorlarının küçükten büyüğe sıralanarak çizildiği karşılaştırmalı grafik. Farklı skorlama fonksiyonları (BNT2, BNT3, Hogervorst, Karlin, Sander ve Valdar) kullanılarak elde edilen skorlar pozisyon bazında sıralanmıştır.

Yapılan sıralamadan sonra, gözlem yapılacak en değişken pozisyon sayısını belirlemek için farklı en değişken pozisyon sayıları ile yapılan tahminler incelenmiştir. Gözlem sayısı artıkça tahmin edilen mutasyon pozisyonu ve amino asit mutasyonu sayısı artmaktadır (Şekil 3.15). Tüm mutasyonları yakalayabilmek için en değişken ilk 700 bölgeyi incelemeye almak gerekmektedir. Bu, gözlem ve analiz yapmak için büyük bir sayıdır. Erken faz çalışmalar yaparken az sayıda çalışma ile en fazla ve anlamlı bilgiyi elde etmek önemlidir. En değişken 15 pozisyon gözlemi ile 12 pozisyon yakalanabilmektedir. Bu gözlem de erken aşamada gerekli bazı mutasyonların yakalanabileceğini ortaya koymaktadır. Bu durum da az sayıda gözlemler ile yeni bir varyantın tanımlanması ve değerlendirmeye alınması gibi durumlarda yeterli olabilmektedir. Bu sebeple bu çalışma kapsamında en değişken 15 pozisyon dikkate alınıp, inceleme ve analiz yapılmıştır.



Şekil 3.15: Gözlem sayısına göre tahmin edilen pozisyon ve amino asit sayılarının değişimi. Yatay eksen, kullanılan gözlem sayısı, dikey eksen, tahmin edilen toplam pozisyon ve amino asit sayıları gösterilmektedir.

Skorlama fonksiyonlarıyla yapılan hesaplamaların hepsi için, kullanılan dört farklı skorlama matrisi üzerinden en çok değişen 15 pozisyon belirlenmiştir. Bu doğrultuda, her set için toplamda değişen bölgeler tespit edilerek çizelge halinde sunulmuştur (Çizelge 3.3). Setler kendi aralarında incelendiğinde en değişken pozisyonlar belirlendikten sonra, set-1 için en benzersiz pozisyonlar 154, 155 ve 519 olarak tespit edilmiştir. Set-2 için bu pozisyon 501, set-3 için 110, 213 ve 214, set-6 için 142 ve 214, set-11 için ise 98 olarak belirlenmiştir.

Çizelge 3.3: Bütün setlerin tüm hesaplamaları için en değişken 15 bölge.

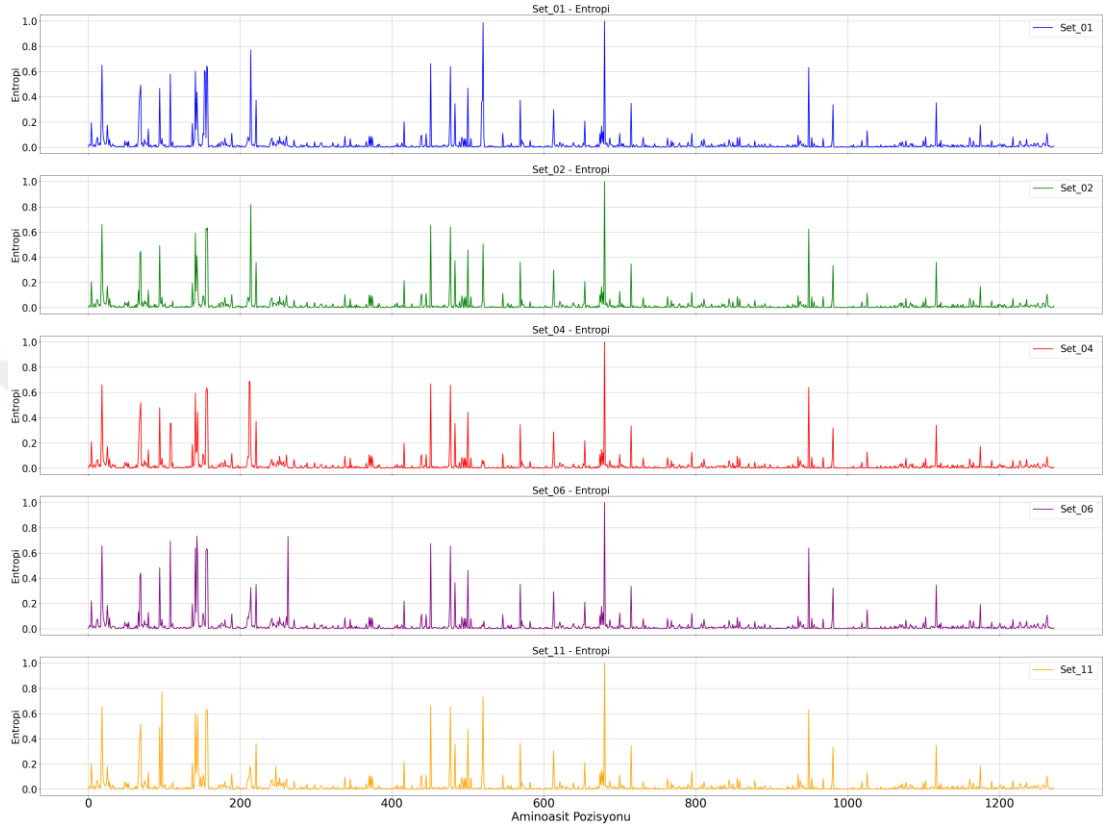
<i>En değişken pozisyonlar</i>	Set 01	Set 02	Set 04	Set 06	Set 11
	19	19	19	19	19
	68	69	68	69	68
	69	70	69	70	69
	70	95	70	95	70
	95	144	95	109	98
	109	145	109	142	98
	144	156	110	144	144
	154	157	145	145	145
	155	158	156	156	156
	157	215	157	157	157
	158	222	158	158	158
	215	452	213	215	452
	452	478	214	264	478
	478	501	222	452	520
	519	521	452	478	521
	520	570	478	570	570
	521	681	570	681	681
	570	716	681	982	716
	681	982	716		982

Rastgele yürüyüş algoritmasının doğru analiz yapabilmesi adına, hem korunan (Çizelge 3.4) hem de değişken bölgeler hesaplamaya dahil edilmiştir.

Çizelge 3.4: Korunan bölgelerin listesi.

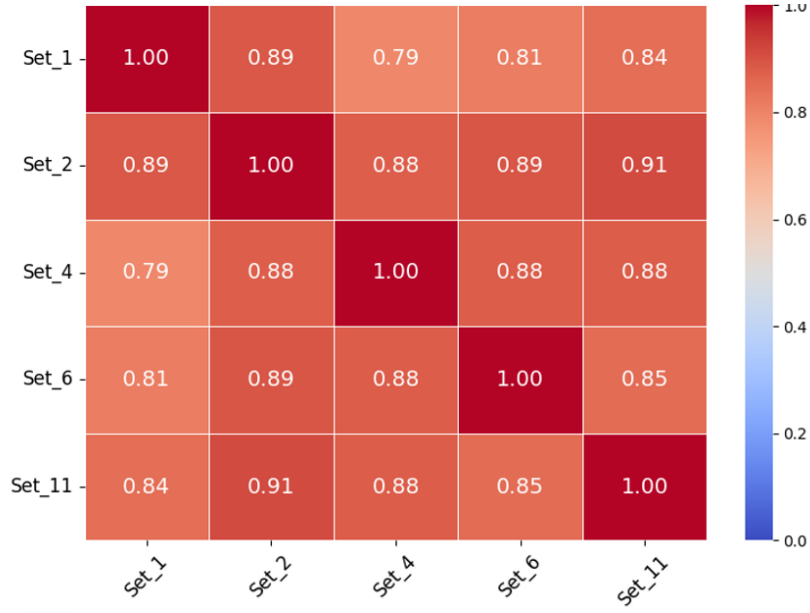
<i>Korunan Bölgeler</i>			
<i>Skorlama Fonksiyonları</i>			
897	1053	563	989
898	671	951	223
901	927	56	609
648	1059	953	994
650	548	1082	996
268	37	191	1126
526	166	959	999
910	291	578	104
398	1064	962	873
1044	295	965	746
277	422	971	749
1046	423	590	1013
663	425	718	1145
919	557	336	766
662	429	601	
665	816	985	

Bu çalışmalara ek olarak, her veri setinin entropisi de hesaplanmıştır. Tüm setler için elde edilen entropi değerleri grafik olarak Şekil 3.16'da verilmiş burada da görüldüğü üzere setlerin entropi değerlerinin benzer olduğu bölgeler olduğu gibi farklı bölgeleri de mevcuttur.



Şekil 3.16: Seçilen beş farklı veri setine (Set_01, Set_02, Set_04, Set_06, Set_11) ait amino asit pozisyonlarına göre hesaplanan entropi değerlerinin dağılım grafiği. Her bir grafik, ilgili setteki amino asit pozisyonlarının çeşitliliğini ve varyasyon seviyesini yansıtmaktadır. Yüksek entropi değerleri, ilgili pozisyonda daha fazla mutasyon çeşitliliği olduğunu göstermekte olup; bu bölgeler evrimsel baskı, fonksiyonel önem veya varyant oluşumu açısından dikkat çekici olabilir.

Setler arasındaki entropi değerlerinin daha iyi anlaşılması adına korelasyonların ısı grafiği elde edilmiştir (Şekil 3.17). Setler arasında entropi değerlerinde farklılıklar gözlemlense de genel özelliklerin korunduğu ve veri setlerinin birbirleriyle belirli bir tutarlılık gösterdiği görülmektedir.



Şekil 3.17: Seçilen beş farklı veri setine ait amino asit pozisyonlarına göre hesaplanan entropi değerlerinin birbiriyle olan korelasyon matrisi.

Entropi sonuçları da kullanılarak, bu sonuçlardan da en çok değişen 15 pozisyon belirlenmiştir. 15 pozisyon incelendiğinde, çoğu pozisyon setler arasında ortakken; bazı pozisyonlar yalnızca o sete özgü pozisyonlardır. Bu pozisyonlar şunlardır: set-1 için 154. ve 155. pozisyonlar, set-2 için 69. pozisyon, set-4 için 213. ve 214. pozisyonlar, set-6 için 144. ve 264. pozisyonlar, Set-11 için 98. pozisyon setlere özel pozisyonlardır.

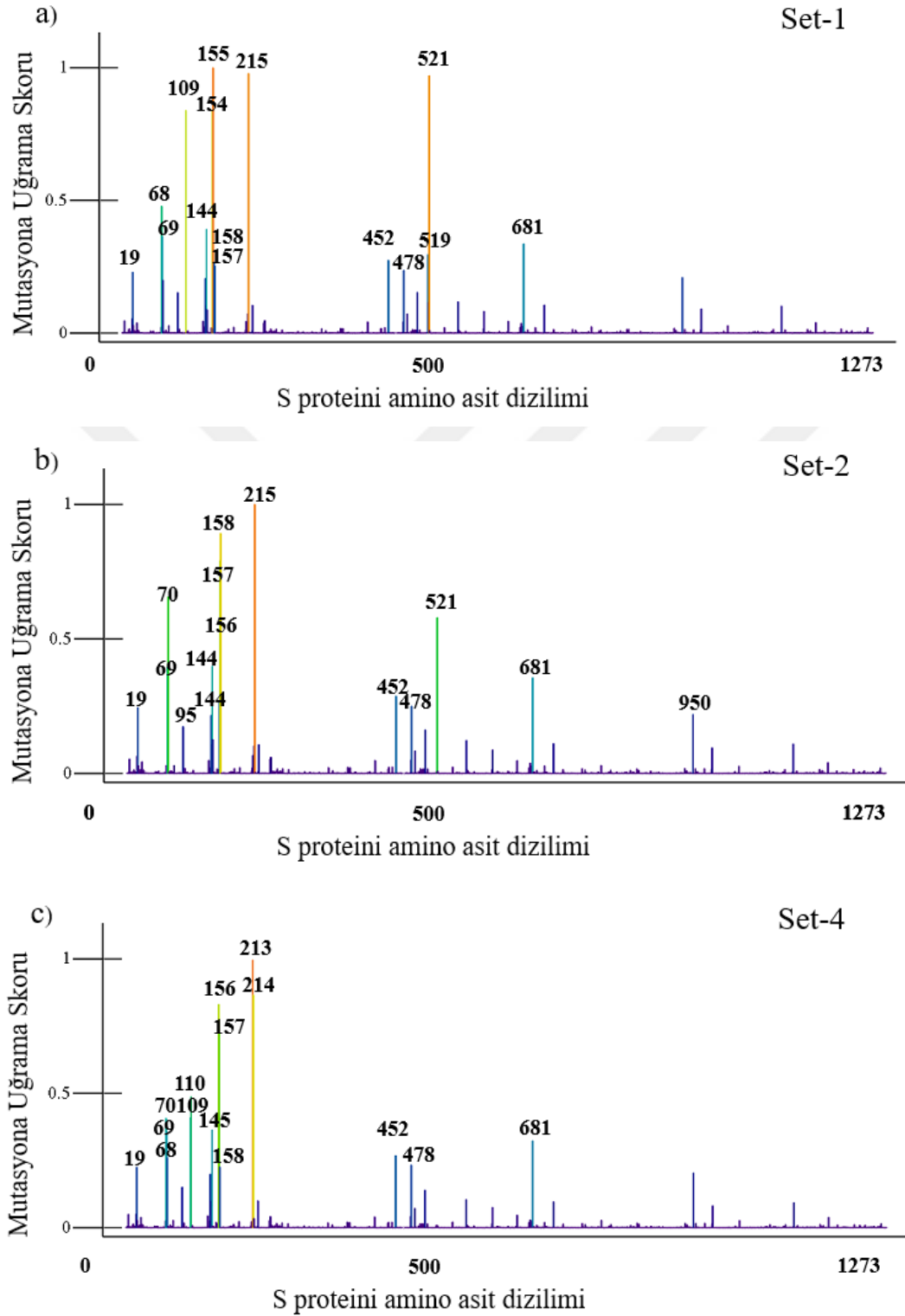
Çizelge 3.5: Bütün setler için entropi değerlerine göre en fazla değişen 15 bölge.

	Set 01	Set 02	Set 04	Set 06	Set 11
<i>En değişken pozisyonlar</i>	19	19	19	19	19
	70	69	70	95	70
	109	70	95	109	95
	142	95	142	142	98
	154	142	145	144	142
	155	156	156	145	145
	157	157	157	156	156
	158	158	158	157	157
	215	215	213	158	158
	452	452	214	264	452
	478	478	452	452	478
	501	501	478	478	501
	521	521	501	501	521
	681	681	681	681	681
	950	950	950	950	950

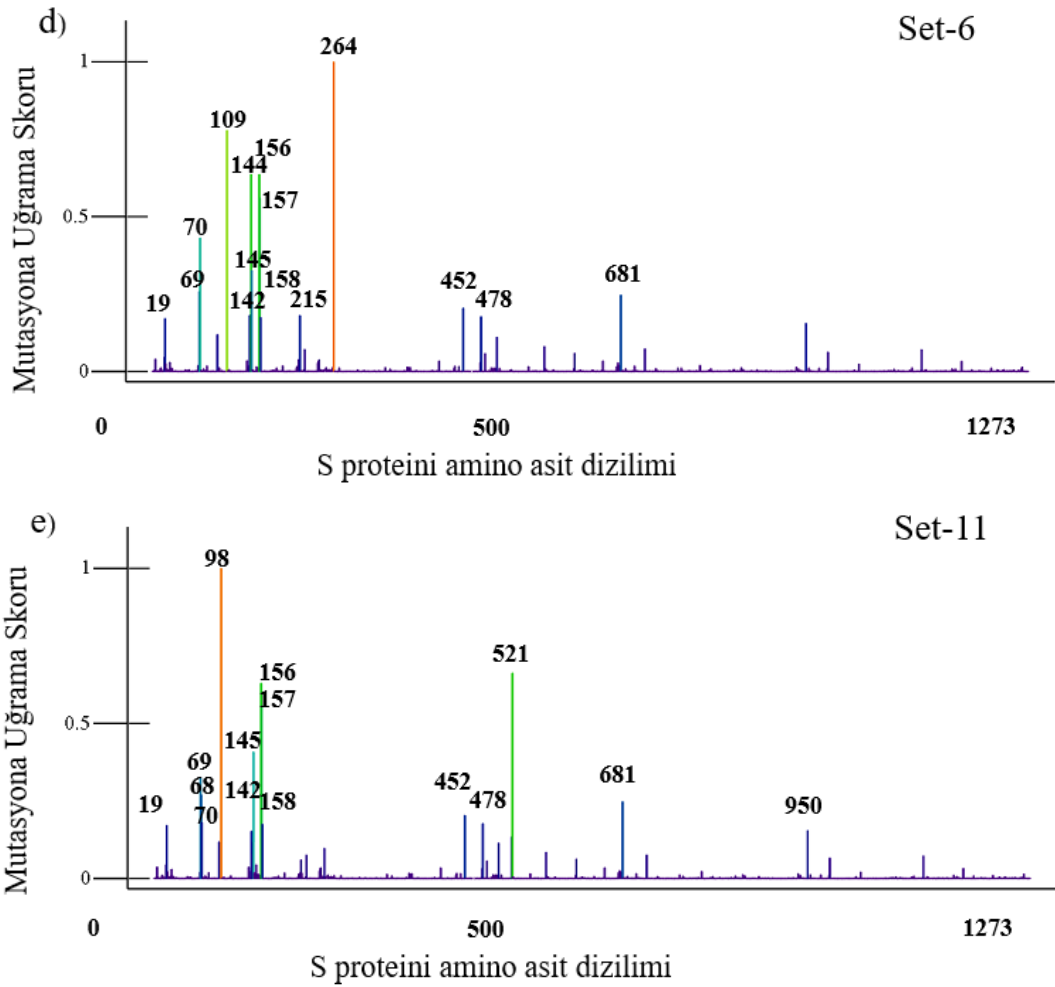
3.3 Mutasyon Haritalarının Oluşturulması

Rastgele yürüyüş algoritması ile yapılan tahminlerin ilk girdisi mutasyon skorlarıdır. 1. grup skorlama fonksiyonlarının kendi aralarında yüksek korelasyon gösterdiği daha önceki bölümlerde gösterilmiştir. Bu nedenle, bu bölümde her set için hesaplanan mutasyon skorları arasından, 1. grup skorlama fonksiyonunu temsilen BNT3 detaylı olarak incelenecektir. Rastgele yürüyüş algoritmasında yapılan tahminlerin pozisyon bilgisi, skor fonksiyonlarından elde edilen sonuçlara göre belirlenirken amino asit tahminleri ise matematiksel modelleme ve yöntemler bölümünde anlatılan amino asit frekanslarına göre hesaplanmaktadır. Öncelikle mutasyon skorları kullanılarak mutasyona uğrama skorları elde edilmiştir ve her veri seti için bu değerler grafik halinde verilmiştir. Setlerin BNT3 skor fonksiyonu BLOSUM62 skor matrisi sonucu Şekil 3.18’de gösterilmiştir. Grafiklerdeki a, b, c, d, e harfleri farklı veri setlerini temsil etmektedir. Ayrıca, mutasyona uğrama olasılığı en yüksek olan (en değişken) 15 pozisyon grafikler üzerinde gösterilmiştir. Burada görüldüğü gibi bazı pozisyonlar her set için ortak olsa da bazı pozisyonlarda farklılıklar ortaya çıkmaktadır. Bu grafiklerdeki minimum değerler ise korunan bölgeleri temsil etmektedir. Bu bilgilere ek olarak her set için mutasyona uğrama skorları grafikleri incelendiğinde mutasyona uğrama olasılığı en yüksek bölgeler 690. pozisyondan önceki bölgelerdir ve 2 grupta yoğunlaşmaktadır. Bu iki grup, S proteininin N-terminal bölgesi (NTD) ile Reseptör Bağlanma Bölgesi’ni (RBD) oluşturmaktadır. S1 alt birimi içerisinde yer alan N-terminal domain (NTD) ve Reseptör Bağlanma Bölgesi (Receptor Binding Domain (RBD)), virüsün canlı organizmaları enfekte etmesi ve bağışıklık sisteminden kaçınması açısından büyük öneme sahiptir (Zhou vd., 2021; Shang vd., 2020). Virüsün yaşam döngüsü için öneme sahip olan bu özellikler için belirtilen bu iki bölge mutasyona uğramaya daha yatkın bölgelerdir. Hesaplanan mutasyona uğrama skorlarının bu bölgede daha yüksek değerlere sahip olmasının biyolojik sebebi bu durum olabilir. Tahmin yapılırken kullanılan rastgele yürüyüş algoritmasındaki toplam vektörü, bu grafikteki sayısal değerlerin art arda toplanarak ilerlemesinden elde edilir. Bu toplama sonucunda, yüksek mutasyona uğrama skoruna sahip pozisyonların toplamı diğer düşük pozisyonlara kıyasla daha fazla artacaktır. Bundan dolayı, rastgele atılacak adımlar mutasyona uğrama skoru daha yüksek olan bölgelere daha fazla yakınsayacaktır. Sonuç olarak mutasyona uğrama skoru yüksek olan pozisyonların, mutasyon geçirme olasılığı diğer bölgelere kıyasla daha fazla olarak

hesaplanır ve rastgele yürüyüş algoritmasının tahminleri bu bölgelerde yoğunlaşmaktadır.



Şekil 3.18: Bütün setler için BNT3 fonksiyonu kullanılarak hesaplanan mutasyona uğrama skorlarının S proteini boyunca dağılımı. Grafik üzerinde, en yüksek varyasyon gösteren ilk 15 amino asit pozisyonu vurgulanmıştır.



Şekil 3.18(devamı): Bütün setler için BNT3 fonksiyonu kullanılarak hesaplanan mutasyona uğrama skorlarının S proteini boyunca dağılımı. Grafik üzerinde, en yüksek varyasyon gösteren ilk 15 amino asit pozisyonu vurgulanmıştır.

3.4 Tahmin Yöntemlerinin Performans Sonuçları

Kullanılan tahmin modelleri ile seçilen eğitim setleri içerisinde Omicron varyantını bulabilmek adına 1 yıllık tahminler yapılmıştır. Rastgele yürüyüş algoritmasıyla elde edilen sonuçlar bir yıl sonrası için tahmin mutasyonlarını ifade etmektedir. Bir önceki bölümde de bahsedildiği gibi çok değişken olan 15 bölge ele alınarak bu bölgeler için tahminler göz önüne alınmıştır. Yapılan tahminlerdeki öncelikli hedef Omicron varyantındaki mutasyonların bu yöntem kullanılarak tespit edilip edilemediğidir. Öncelikle, en fazla değişkenlik gösteren 15 pozisyon için yapılan tahminler her eğitim seti (Set 1, 2, 4, 6 ve 11) bazında incelenmiştir. 5 farklı set 4 farklı skor fonksiyonu, 4 farklı skor matrisi, 4 farklı amino asit frekansı için toplam 320 farklı kombinasyon kullanılarak tahminler yapılmıştır. Tahmin sonuçları değerlendirilirken, mutasyon

tahminlerinde Omicron varyantına en yüksek benzerliği gösteren setlerin belirlenmesi hedeflenmiştir.

Değerlendirmeler iki temel başlık altında ele alınmıştır:

- En iyi pozisyon tahminini yapan setler
- En iyi amino asit mutasyon tespitini yapan setler

Omicron varyantına ait gerçekleşen mutasyonların pozisyon listesi çıkarılmıştır ve hangi hesaplamalar için hangi mutasyon pozisyonları bulunduğu tespit edilmeye çalışılmıştır. Bu listede bulunan mutasyonlar belirlenmeye çalışılırken en değişken 15 pozisyon dikkate alınmıştır. Tahmin sonuçları değerlendirmeye alınırken tahmin dağılımlarını ve gerçek mutasyonu gösteren analizler yapılmıştır. Şekil 3.19, Şekil 3.20, Şekil 3.21 grafiklerinde görüldüğü gibi en değişken 15 pozisyon belirtilmiş ve bu 15 pozisyon için yapılan tahmin dağılımları verilmiştir. Tahmin dağılımlarının görsel gösterimleri, rastgele yürüyüş algoritması ile belirli bir pozisyonda en sık tahmin edilen amino asitleri de öne çıkarmaktadır. Büyük harflerle gösterilen amino asitler, ilgili pozisyonda rastgele yürüyüş algoritması tarafından çok sayıda tahmin edildiğini; küçük harflerle gösterilenler ise daha az sayıda tahmin edildiğini ifade etmektedir. Amino asitlerin renk kodlamaları ise her bir amino aside özgü olarak tanımlanmış ve görsel ayrımı desteklemek amacıyla kullanılmıştır. Son olarak o bölgede görülen gerçek mutasyon da belirtilerek tahmin edilen mutasyon verileri ile karşılaştırmalar yapılmıştır. 15 pozisyon için toplam 12 farklı pozisyon bölgesi tespit edilmiştir. Bu bölgeler, 60, 70, 95, 142, 143, 144, 145, 213, 214, 478, 501 ve 681 numaralı pozisyonlardır. Bazı eğitim veri setleri benzer tespit performansları sergilerken, diğer veri setleri ise grup halinde farklı performanslar göstererek farklı mutasyon pozisyonlarını tespit edebilmişlerdir. Tahminler açısından en başarılı setler şu şekilde belirlenmiştir: 2. setin BNT3 skor fonksiyonu ve PAM120 skor matrisi, gözlemlenmiş amino asit frekansı kombinasyonu 8 farklı mutasyon pozisyonu tespit edebilmiştir. Bu gruba ek olarak 4. set Hogervorst skor fonksiyonu BLOSUM62 skor matrisi ve referans amino asit frekansı kombinasyonu ise 7 farklı mutasyon bölgesi tespit etmiştir ve diğer grupların tespit edemediği 213. pozisyon yakalanmıştır. Son olarak 11.set Hogervorst skor fonksiyonu BLOSUM62 skor matrisi referans amino asit frekansı kombinasyonu da 7 farklı amino asit tespit ederek öne çıkmıştır. Bu bulgudan hareketle, her kombinasyonun farklı mutasyon bölgelerini ortaya çıkarabildiği anlaşılmış ve bu nedenle tüm tahmin kombinasyonları kapsamlı bir

şekilde analiz edilmiştir. Tahmin edilen amino asit mutasyonları incelendiğinde ise bu değerlendirme sonucunda 4. ve 11. setlerin Hogervorst skor fonksiyonu BLOSUM62 skor matrisi, referans frekans ile olan kombinasyonları en çok amino asit mutasyonu tahmin eden kombinasyonlar olmuştur. Bu çalışma kapsamında silinmeler (delesyonlar) dikkate alınmadığı için ne kadar pozisyon tahmini yapılmış olsa da o bölgedeki silinme tespit edilemediği için amino asit mutasyonu tahminlerinde düşüş yaşanmıştır. Örneğin set-2, BNT3, PAM120 skor matrisi gözlemlenmiş amino asit frekansı kombinasyonu tahminlerinde tespit edilen pozisyonların çoğu silinme içerdiği için amino asit mutasyonu tespit edilememiş gibi görünmektedir.

REFERANS		TAHMİN DAĞIMLARI																				GERÇEK MUTASYON
19	T			D	E		G		I	K	L		N	P	Q	R	S		V		Y	
69	H	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V	W	Y	DEL
70	V	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V	W	Y	DEL
142	G	A		D	E	F					L		N		Q		S	T	V		Y	D
143	V	A	C	D		F		H	I	K		M	N	P	Q	R	S		V		Y	DEL
144	Y	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V		Y	DEL
145	Y	A	C	D		F	G	H		K	L		N				S	T	V		Y	D
156	E	A	C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V			
157	F	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
158	R	A	C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V	W	Y	
214	R	A		D		F	G		I		L			P		R	S	T	V		Y	EPE
215	D	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
452	L	A	C	D	E	F	G				L			P	Q		S	T	V			
521	P	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
681	R		C	D	E	F				K	L		N			R	S	T	V	W	Y	H

Şekil 3.19: Set-2 için BNT3 skarlama fonksiyonu, PAM120 skarlama matrisi ve gözlemlenmiş amino asit frekanslarının kombinasyonu kullanılarak yapılan tahminlerin sonuçları.

Set-4, Hogervorst, BLOSUM62, referans frekans kombinasyonu tahminleri incelendiğinde 7 farklı pozisyon tahmini için farklı amino asitler tahmin edilmiştir ve 145D, 213P, 478K, 681H mutasyonları tespit edilmiştir. Bu bilgilere ek olarak, 213. pozisyonda yapılan Prolin (P) tahmini, tahmin edilen diğer amino asit olan Glisin (G)' e göre daha düşük olasılığa sahip olmasına rağmen, doğru tahmin yapıldığı için değerlendirmeye dahil edilmiştir (Şekil 3.20).

REFERANS		TAHMİN DAĞIMLARI																						GERÇEK MUTASYON
EN DEĞİŞKEN 15 POZİSYON	19	T	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V				
	68	I	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V				
	69	H	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		DEL
	109	T	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V				
	110	L	A	C	D	E	F	G		I	K	L	M	N		Q	R	S	T	V		Y		
	142	G	A		D	E	F	G		I	K	L		N	P		R	S	T	V	W	Y		DEL
	145	Y	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V		Y		D
	156	E	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		
	157	F	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		
	158	R	A	C	D		F	G		I		L		N	P	Q	R	S	T	V				
	213	V	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		P
	214	R	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V		Y		EPE
	452	L	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V		Y		
	478	T	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V		Y		K
	681	R	A	C	D	E	F	G	H	I	K	L	M	N	P			S		V	W	Y		H

Şekil 3.20: Set-4 için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahminlerin sonuçları.

REFERANS		TAHMİN DAĞIMLARI																						GERÇEK MUTASYON
EN DEĞİŞKEN 15 POZİSYON	19	T	A	C		E	F	G		I	K	L	M	N	P	Q	R	S	T	V				
	68	I	A	C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V		Y		
	69	H	A	C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V	W	Y		DEL
	95	T	A	C		E		G		I		L			P	Q	R	S	T			Y		I
	98	S	A	C	D	E	F	G	H	I	K	L		N	P	Q		S	T	V		Y		
	142	G	A	C		E	F	G			K	L	M	N	P	Q	R	S	T	V		Y		D
	145	Y	A	C	D	E	F	G		I	K	L		N		Q	R	S	T	V	W	Y		D
	156	E	A		D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		
	157	F	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		
	158	R	A		D	E	F	G	H	I	K	L		N	P		R	S	T	V		Y		
	452	L	A		D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V	W	Y		
	478	T	A	C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V	W	Y		K
	501	N	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V	W	Y		Y
	521	P	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V	W	Y		
	681	R	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		H

Şekil 3.21: Set-11 için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.

Set-11, Hogervorst, BLOSUM62, referans frekans kombinasyonu tahminleri incelendiğinde ise 7 farklı pozisyon bölgesinden 5 tanesi için doğru amino asit mutasyonu tahmini yapılmıştır. Bunlar 95I, 145D, 478K, 501Y, 681H mutasyonlarıdır. Bu bilgilere ek olarak 681.pozisyon için diğer amino asit tahminleri H (Histidin) tahmini sayısından fazladır fakat yine de doğru tahmin yapıldığı için sayılmıştır (Şekil 3.21). Sonuç olarak Omicron varyantına ait 44 mutasyon pozisyonun 12 tanesi, 48

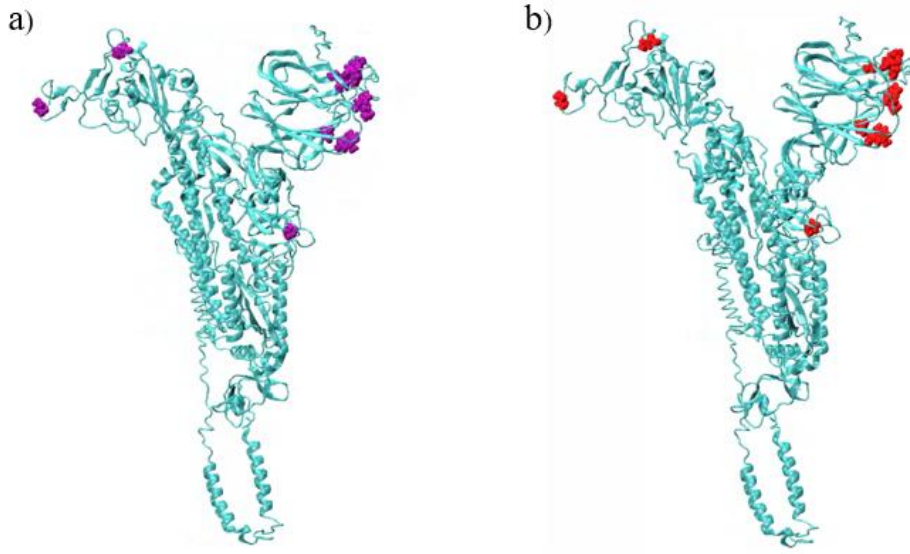
Omicron amino asit mutasyonundan 7 tanesi bu hesaplamalar ile tespit edilmiştir. Bu çalışma kapsamında silinmeler hesaplamaya alınmadığı için yapılan tahminlerin içinde silinme bulunmamaktadır. Bu sebeple tespit edilen mutasyon sayısı silinme olmadan da değerlendirmeye alınmıştır. Silinme olmayan mutasyonların 41 tanesinin 7 tanesi tespit edilmiştir (Çizelge 3.6). Tahmin sonuçlarının 3 boyutlu yapıda incelenmesi sonucunda, yapılan tahminlerin büyük bir kısmının S proteininin N-terminal domaini (NTD) ile Reseptör Bağlanma Bölgesi'nde (Receptor Binding Domain, RBD) yoğunlaştığı görülmüştür (Şekil 3.22). Daha önce de belirtildiği gibi, bu bölgeler virüsün konak hücreyi enfekte etmesinde ve bağışıklık sisteminden kaçışında kritik rol oynamaktadır. Bu nedenle, bu bölgelerde meydana gelebilecek mutasyonlar, viral adaptasyon ve bağışıklık sistemiyle etkileşim açısından belirleyici olup, tahminlerin bu bölgelerde yoğunlaşması, beklenen bir durumdur.

Çizelge 3.6: Omicron varyantı için yapılan tahminlerin performans karşılaştırması.

	Pozisyon Mutasyon Sayısı	Amino asit Mutasyon Sayısı	Silinme Olmadan Amino asit Mutasyon Sayısı
Gerçek	44	48	41
Tahmin edilen	12	7	7

Yapılan bu tahminlerin aynısı, seçilen 5 farklı set için hesaplanan entropi değerleri için de yapılmıştır. 5 set ve 4 farklı amino asit frekansı kullanılarak yapılan hesaplamalarda toplam 20 farklı kombinasyon için tahmin sonucu elde edilmiştir (Çizelge 3.7). Analizler sonucunda, tüm veri setleri dört farklı amino asit frekansı açısından tutarlı sonuçlar vermiştir. 4. set ile 70, 95, 142, 144, 145, 213, 214, 478, 501 ve 681 numaralı pozisyonlar tahmin edilerek toplamda 9 farklı amino asit mutasyonu tespit edilmiştir. Set 2 ile 7 farklı amino asit pozisyonu tespit edilmiştir ve 4. set tahminlerinden farklı olarak 69. pozisyon bu set ile tespit edilebilmiştir. 6. set kombinasyonları ile de toplam 7 farklı amino asit pozisyonu tespit edilmiştir ve diğer setlerden farklı olarak 144. pozisyon tespit edilmiştir. Set 11 kombinasyonları ile toplam 7 farklı pozisyon, Set 1 kombinasyonları ile ise 5 farklı pozisyon tespit edilmiştir; ancak, diğer setlerden farklı

herhangi bir pozisyon tespit edilememiştir. 11 farklı mutasyon pozisyonu set 4, set 2 ve set 6 ile başarılı bir şekilde saptanmıştır.



Şekil 3.22: SARS-CoV-2 S proteininin 3B yapısı üzerinde Omicron varyantı için tespit edilen bölgelerin görselleştirilmesi. (a) Mor renk ile gösterilen bölgeler, skor fonksiyonlarına dayalı olarak yapılan tahminler ile tespit edilen amino asit pozisyonlarını temsil etmektedir. (b) Kırmızı renk ile gösterilen alanlar ise bu pozisyonlarda tespit edilen spesifik amino asit değişimlerini göstermektedir.

Çizelge 3.7: Entropi kullanılarak yapılan mutasyon pozisyonu tahminleri.

	69	70	95	142	144	145	213	214	478	501	681
Set Numarası: 04		✓	✓	✓		✓	✓	✓	✓	✓	✓
Set Numarası: 02	✓	✓	✓	✓					✓	✓	✓
Set Numarası: 06			✓	✓	✓	✓			✓	✓	✓
Set Numarası: 11		✓	✓	✓		✓			✓	✓	✓
Set Numarası: 01		✓		✓					✓	✓	✓

Amino asit mutasyonları için yapılan tahminler incelendiğinde, setlerin farklı kombinasyonları ile farklı amino asit mutasyonları tespit edilmiştir. Tüm tahminler

göz önünde bulundurulduğunda, toplamda 7 farklı amino asit mutasyonu belirlenmiştir. Tespit edilen mutasyonlar Çizelge 3.8’de sunulmuştur.

Çizelge 3.8: Entropi kullanılarak yapılan amino asit mutasyonu tahminleri.

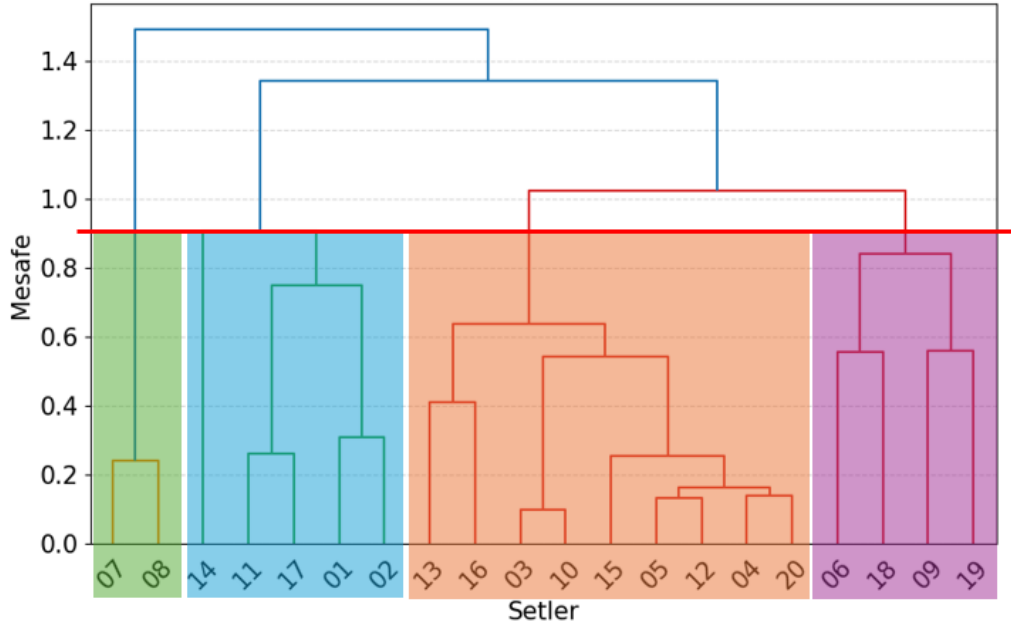
	95I	142D	478K	501Y	681H	145D	213P
Set Numarası:04 Frekans: Referans	✓		✓	✓	✓	✓	✓
Set Numarası: 06 Frekans: Referans	✓	✓	✓	✓	✓	✓	
Set Numarası: 02 Frekans: Doğal	✓	✓	✓	✓	✓		
Set Numarası:04 Frekans: Gözlemlenmiş	✓		✓	✓		✓	✓
Set Numarası:06 Frekans: Dönüşüm	✓	✓	✓		✓	✓	
Set Numarası:06 Frekans: Gözlemlenmiş	✓	✓	✓	✓		✓	
Set Numarası:11 Frekans: Gözlemlenmiş	✓	✓	✓	✓		✓	
Set Numarası:01 Frekans: Dönüşüm		✓	✓	✓	✓		
Set Numarası:01 Frekans: Referans		✓	✓	✓	✓		
Set Numarası:02 Frekans: Gözlemlenmiş	✓	✓	✓		✓		
Set Numarası:02 Frekans: Referans	✓	✓	✓	✓			
Set Numarası:04 Frekans: Dönüşüm	✓			✓	✓		✓
Set Numarası:04 Frekans: Doğal		✓	✓		✓		✓
Set Numarası:06 Frekans: Doğal	✓			✓	✓	✓	
Set Numarası:11 Frekans: Dönüşüm	✓	✓		✓		✓	
Set Numarası:11 Frekans: Doğal	✓			✓	✓	✓	
Set Numarası:01 Frekans: Gözlemlenmiş		✓		✓	✓		
Set Numarası:11 Frekans: Referans	✓	✓	✓				
Set Numarası:01 Frekans: Doğal			✓		✓		
Set Numarası:02 Frekans: Dönüşüm		✓		✓			

Entropi deęerleri kullanılarak yapılan tahminler, skor fonksiyonları ile yapılan tahminlerle karřılařtırıldığında, entropi tabanlı tahminlerin skor fonksiyonlarıyla elde edilen tahminlerden farklı bir mutasyon pozisyonu tespit edememiřtir. Skorlama fonksiyonları kullanılarak yapılan pozisyon tespitlerinde, entropi temelli tespitlere kıyasla bir pozisyon daha fazla tespit yapılmıřtır. Tespit edilen bu pozisyon 143. pozisyonudur ve entropi skorları kullanılarak bu pozisyon tespit edilememiřtir. Amino asit mutasyonu aısından ise aynı sonuları vermiřlerdir. Sonu olarak, en deęiřken 15 pozisyon incelenerek 44 mutasyon pozisyonunun 11 tanesi 48 amino asit mutasyonunun 7 tanesi tespit edilmiřtir. alıřmalar sırasında, Blm 1.2.7.1’de de belirtildięi zere Omicron varyantında trler arası geiřin gerekleřtięinin gzlemlenmesi ve farklı pozisyonlarda mutasyonların ortaya ıkması, kullanılan yntemlerin doęruluęunu deęerlendirmek amacıyla yeni bir deney seti oluřturulmasını gerekli kılmıřtır. Bu nedenle, bu ařamaya kadar uygulanan iřlemler, Delta varyantı iin yapılan tahminler aısından da tekrarlanmıřtır.

3.5 Sonuların Doęrulanması

Omicron varyantının genetik yapısına bařka trlerin dahil olması tespit edilemeyen farklı mutasyonların ortaya ıkmasına neden olmuř olabilir. Genetik bilginin iine giren farklı trlerin genetik bilgileri sadece insanlardan alınan sekans bilgileri ile tespit edilemez, bu sebeple bir doęrulama seti incelemesi yapılmıřtır. Burada Alpha, Beta, Gamma varyantları kullanılarak, Delta varyantının mutasyonlarının tahmini iin alıřılmıřtır. Alpha, Beta, Gamma ve Delta varyantları, dięer trlerin genetik bilgisinden etkilenmedięi iin, mutasyonların da benzer řekilde trler arası etkiden etkilenmemesi beklenir. Bu nedenle yapılan tahminlerin de bu doęrultuda gerekleřmesi gerekmektedir. Bu sebeple kullanılan yntemlerin doęruluęunu tespit etmek iin, dięer trlerden etkilenmemiř bir varyantın mutasyon tahminleri iin sonular analiz edilmiřtir. ncelikle, Alpha, Beta ve Gamma varyantlarını ieren yani 2021 ncesi protein sekans verileri kullanılarak Delta varyantının mutasyonları tahmin edilmeye alıřılmıřtır. Daha nce bahsedilen yntemlerin aynısı 2021 ncesi protein sekans verisi ile yapılmıřtır. 20 defa ve her periyottan 600 sekans olacak řekilde setler oluřturulmuřtur. Elde edilen 20 protein sekans seti iin filogenik aęa oluřturulmuř seilen mesafe eřik deęeri ise bu rastgele setler iin 0,9 olarak belirlenmiřtir ve

içlerinden birbirlerine uzak olacak şekilde 4 set seçilmiştir. Seçilen set numaraları 03, 06, 07 ve 11 şeklindedir (Şekil 3.23).



Şekil 3.23: Delta varyantına ait analizlerde kullanılan veri setlerinin benzerliğine göre oluşturulan filogenetik ağaç. Dikey eksende kümeler arası mesafe ölçütü yer almakta olup, kırmızı çizgi belirli bir eşik değeri üzerinde kümelerin ayrımını göstermektedir ve toplam dört ana kümeye ayrılmıştır.

4 set için skortlama fonksiyonları hesaplanmış ve elde edilen skorlar kullanılarak mutasyona uğrama skorları (değişebilirlik) elde edilmiştir. Rastgele yürüyüş algoritmasının tahminleri, Omicron varyantında olduğu gibi en yüksek değişebilirlik skoruna sahip 15 bölge üzerinde odaklanarak incelenmiştir. Her bir veri seti için, uygulanan tüm hesaplama yöntemleri (BNT2, BNT3, Karlin ve Hogervorst) sonucunda elde edilen en değişken 15 pozisyon belirlenmiş ve analiz kapsamında sunulmuştur (Çizelge 3.9).

Çizelge 3.9: Delta varyantı için yapılan hesaplamada en değişken bölgeler.

Değişken Bölgeler	
Set 03	5, 12, 13, 18, 19, 24, 27, 50, 69, 70, 98, 142, 143, 145, 157, 158, 213, 214, 215, 221, 222, 242, 243, 254, 255, 262, 264, 371, 373, 375, 452, 477, 478, 501, 570, 614, 640, 675, 676, 677, 678, 681, 704, 716, 936, 939, 982, 1176
Set 06	5, 12, 13, 18, 19, 24, 50, 69, 70, 98, 109, 142, 143, 144, 145, 156, 157, 213, 215, 222, 243, 254, 255, 262, 264, 371, 373, 375, 452, 477, 478, 501, 570, 614, 640, 681, 716, 936, 939, 982, 1176, 1252
Set 07	5, 12, 13, 18, 19, 24, 69, 70, 98, 109, 142, 143, 144, 147, 148, 213, 214, 215, 222, 243, 254, 255, 259, 260, 261, 262, 263, 371, 373, 375, 452, 477, 478, 501, 547, 570, 614, 640, 681, 716, 939, 940, 982, 1176, 1252
Set 11	5, 12, 18, 19, 50, 67, 68, 69, 70, 98, 138, 143, 144, 145, 156, 157, 213, 214, 222, 243, 254, 255, 371, 373, 375, 452, 477, 478, 501, 520, 521, 570, 614, 640, 681, 688, 716, 813, 936, 939, 940, 982, 1176

İncelenen verilerde tespit edilen bir diğer önemli bulgu, setlerin kendi aralarında benzer sonuçlar etrafında yoğunlaşmasıdır. Pozisyon tahminleri değerlendirildiğinde; set-2'e ait BNT3, PAM120 ve dönüşüm amino asit frekansı kombinasyonu, yine set-3'ün Hogervorst BLOSUM62 referans amino asit frekansı kombinasyonu ve buna ek olarak set-6'nın Hogervorst PET91 referans amino asit frekansı sonuçları toplamda 7 farklı mutasyon pozisyonu ve 6 farklı amino asit mutasyonu tespit etmiştir. Bu pozisyonlar 142, 156, 157, 158, 452, 614 ve 681. pozisyonlarıdır. Tespit edilen amino asit mutasyonları ise 142D, 156G, 158G 452R, 614G, 681R mutasyonlarıdır.

Set-2, BNT3, PAM120, dönüşüm frekansı kombinasyon tahmini incelendiğinde 614G, 158G, 142D mutasyonları tespit edilmiştir. Burada tahminler genel olarak pozisyonlara dengeli şekilde dağılmış olsa da 145. pozisyon için diğer pozisyonlara kıyasla daha fazla sayıda tahmin gerçekleştirilmiştir ancak Delta varyantı için burada herhangi bir mutasyon olmadığı için, bu bölge değerlendirmeye alınmamıştır. (Şekil 3.24).

EN DEĞİŞKEN 15 POZİSYON	REFERANS		TAHMİN DAĞIMLARI																				GERÇEK MUTASYON
	69	H	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	70	V	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V		Y	
	142	G	A		D		F	G		I	K	L	M	N	P	Q		S	T	V		Y	D
	143	V	A	C	D			G		I	K	L	M	N	P		R		T	V			
	145	Y	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	157	F	A		D	E	F	G			K	L	M	N	P	Q	R	S	T	V		Y	DEL
	158	R	A		D	E		G		I		L		N		Q		S		V			G
	213	V	A		D		F	G			K			N	P	Q	R		T		W	Y	
	214	R	A		D	E	F	G		I	K	L		N		Q	R	S	T	V		Y	
	215	D	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V			
	222	A		C		E	F	G		I	K	L		N	P			S				Y	
	242	L	A	C	D	E	F	G		I				N	P	Q	R	S	T	V		Y	
	243	A	A	C	D		F	G		I	K	L		N	P	Q	R	S	T	V	W	Y	
	264	A	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V		Y	
614	V	A	C	D	E	F	G	H	I	K	L	M	N	P			S	T	V		Y	G	

Şekil 3.24: Set-3 için BNT3 skarlama fonksiyonu, PAM120 skarlama matrisi ve dönüşüm frekansı kombinasyonu kullanılarak yapılan tahminlerin sonuçları.

Set-3, Hogervorst, BLOSUM62, referans frekansı kombinasyonu tahmini incelendiğinde ise 614G, 681R mutasyonu tespit edilirken 452.pozisyon için tahmin yapılsa bile 452R amino asit mutasyonu yakalanamamıştır. Bu sebeple pozisyon sayımına dahil edilmiş olsa da amino asit mutasyonu sayımında değerlendirmeye alınmamıştır (Şekil 3.25).

	REFERANS	TAHMİN DAĞIMLARI																				GERÇEK MUTASYON
		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
EN DEĞİŞKEN 15 POZİSYON	5	L	A			F	G			K	L		N	P	Q		S	T				
	18	L	A		D	E	F	G	H	I	L	M	N		Q			T	V			
	69	H	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V	W	Y
	70	V	A		D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V		Y
	142	G			D		F	G		I	K	L		N	P	Q	R	S	T	V		Y
	145	Y	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	214	R	A	C		E		G	H			L	M	N	P				T	V		Y
	215	D	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V		Y
	222	A	A	C	D	E	F	G		I	K	L	M	N	P		R	S	T	V	W	Y
	264	A	A	C		E	F	G		I	K	L		N	P	Q		S	T	V	W	
	452	L	A		D		F	G			K			N	P	Q			V		Y	R
	501	N	A		D	E	F		H		K		M		Q		S	T				
	570	A	A			E			H	I	K	L		N		Q		S		V		Y
	614	V	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V		Y
	681	R	A		D	E	F	G		I	K	L		N	P	Q	R	S	T		W	Y

Şekil 3.25: Set-3 için Hogervorst skarlama fonksiyonu, BLOSUM62 skarlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.

Set-6, Hogervorst, PET91, referans frekansı kombinasyonu tahmini incelendiğinde ise 142D, 452R, 614G 681R mutasyonları yakalanmıştır. Şekil 3.26'da görüldüğü üzere,

452R mutasyonu tespit edilmiş olsa da bu amino asit değişimi için diğer 452 pozisyonundaki mutasyonlara kıyasla daha az sayıda tahmin yapılmıştır.

REFERANS		TAHMİN DAĞILIMLARI																							GERÇEK MUTASYON
69	H	A	C		E	F	G	H	I	K	L		N	P	Q	R	S	T	V		Y				
70	V	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y				
109	T	A	C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V	W	Y				
142	G		C	D		F	G	H	I	K	L	M	N	P		R	S	T		W	Y	D			
143	V		C		E	F	G				L		N	P	Q		S		V		Y				
144	Y	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V	W	Y				
145	Y	A	C		E	F	G	H	I	K	L		N		Q	R	S	T	V	W	Y				
156	E			D	E		G		I	K			N	P	Q		S	T	V			G			
215	D			D	E	F	G		I	K	L	M	N	P	Q		S	T	V		Y				
222	A	A	C			F	G	H		K	L		N		Q	R	S	T	V						
264	A	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y				
452	L	A		D							L		N		Q	R	S	T	V			R			
501	N	A	C	D		F	G				L			P			S	T			Y				
614	V	A	C	D		F	G	H	I	K	L		N	P	Q		S	T	V		Y	G			
681	R	A	C	D	E	F	G	H	I					P		R	S	T	V		Y	R			

Şekil 3.26: Set-6 için Hogervorst skorumu fonksiyonu, PET91 skorumu matrisi ve referans amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.

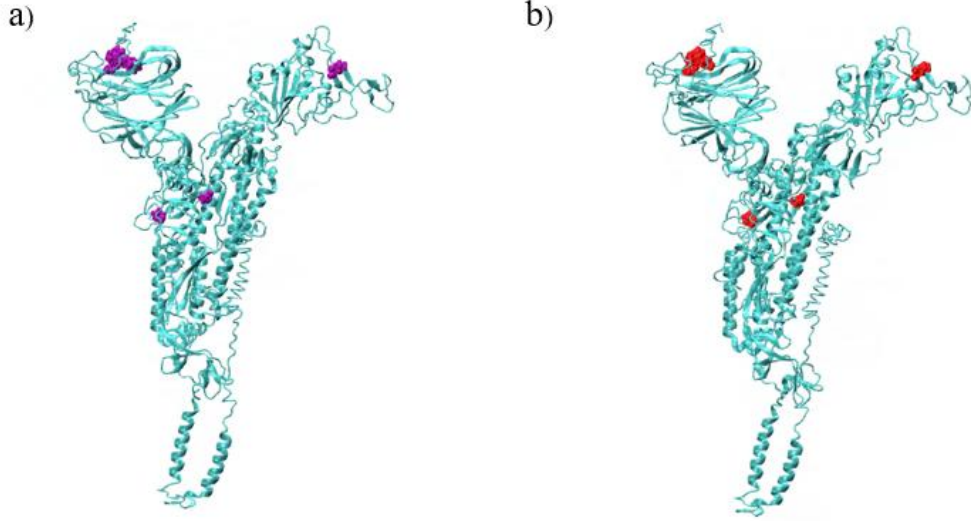
Buradan da çıkarılabileceği gibi farklı setler ve kombinasyonların kullanılması farklı mutasyonların yakalanması için gereklidir. Yapılan tahminlerin performansını değerlendirmek için tahmin edilen ve gerçek değerler sayısal olarak incelenmiştir. 16 pozisyon mutasyonundan 15 gözlemle 7 adet doğru tahmin edilmiştir. 18 adet amino asit mutasyon sayısından ise 15 gözlem sayısı ile 6 tanesi doğru tahmin edilmiştir.

Çizelge 3.10: Delta varyantı için yapılan tahminlerin performans karşılaştırması.

	Pozisyon Mutasyon Sayısı	Amino asit Mutasyon Sayısı	Silinme Olmadan Amino asit Mutasyon Sayısı
Gerçek	16	18	15
Tahmin edilen	7	6	6

Yapılan tahminler, S proteininin üç boyutlu yapısı üzerinde de incelenmiştir (Şekil 3.27). Elde edilen sonuçlar, mutasyonların hem belirli pozisyonlarda hem de amino asit düzeyinde, özellikle N-terminal domain (NTD) ve Reseptör Bağlanma Bölgesi (RBD) gibi kritik bölgelerde yoğunlaştığını ortaya koymaktadır. Buna ek olarak, S2

alt birimi ile trimerik yapının merkezini oluşturan ve füzyon ile zar birleşimi süreçlerinde önemli rol oynayan gövde kısmında da mutasyon tahminlerinin yapıldığı gözlemlenmiştir.



Şekil 3.27: SARS-CoV-2 S proteininin 3B yapısı üzerinde Delta varyantı için tespit edilen bölgelerin görselleştirilmesi. (a) Mor renk ile gösterilen bölgeler, skor fonksiyonlarına dayalı olarak yapılan tahminler ile tespit edilen amino asit pozisyonlarını temsil etmektedir. (b) Kırmızı renk ile gösterilen alanlar ise bu pozisyonlarda tespit edilen spesifik amino asit değişimlerini göstermektedir.

Delta varyantı için yapılan tahmin analizleri, seçilen setlerin entropi değerleri için de yapılmıştır. Entropi değerlerinin tahmin sonuçlarında 4 farklı mutasyon pozisyonu tespit edilmiştir. Bu pozisyonlar 142, 452, 614 ve 681. pozisyonlardır. Bu pozisyonlara ek olarak yapılan amino asit tahminleri incelendiğinde 4 farklı amino asit mutasyonu tespit edilmiştir. Bunlar 142D, 452R, 614G ve 681R mutasyonlarıdır. Delta varyantı için, entropi değerleri kullanılarak yapılan tahminler ile, skor fonksiyonları ile yapılan tahminler karşılaştırıldığında, entropi tabanlı tahminlerin skor fonksiyonlarıyla elde edilen tahminlerden farklı bir mutasyonu tespit edemediği ve skor fonksiyonlarından daha az sayıda mutasyon tespit ettiği gözlemlenmiştir. Bu durumun temel nedeni, entropi hesabının amino asitler arasındaki dönüşüm ilişkilerini dikkate almamasıdır. Entropi yalnızca değişim sayısını hesaba katar, bu değişimlerin hangi amino asitler arasında gerçekleştiğine dair bilgiyi hesaplamasında kullanmaz. Dolayısıyla, değişimin yönü ve biyolojik anlamı göz ardı edildiğinden, yapılan tahminlerde yetersiz kalmıştır. Bu nedenle, ilerleyen adımlarda yapılacak gelecek tahminlerinde entropi değerleri dikkate alınmamıştır.

3.6 Gelecek Mutasyonların Tahmini

Gelecek mutasyonların tahmini için, 2025 Şubat ayına kadar olan protein sekans verileri GISAID veri tabanından elde edilmiştir. Tez boyunca daha önce kullanılan tüm yöntemler uygulanmıştır. Öncelikle protein sekans verileri tarihlerine göre yıllara bölünmüştür (Çizelge 3.11). 2022, 2023, 2024, 2025 yıllarına ait sekans verileri incelenmiş ve her yıla ait verilerin dağılımları incelendikten sonra her yıldan 1500 protein sekansı olacak şekilde rastgele seçimler yapılmıştır. Elde edilen rastgele sekanslar birleştirilmiş ve toplam 5814 adet protein sekansı içeren set elde edilmiştir. Gelecek mutasyonların tahmini için Omicron sonrası verilerin alınmasına özen gösterilmiştir.

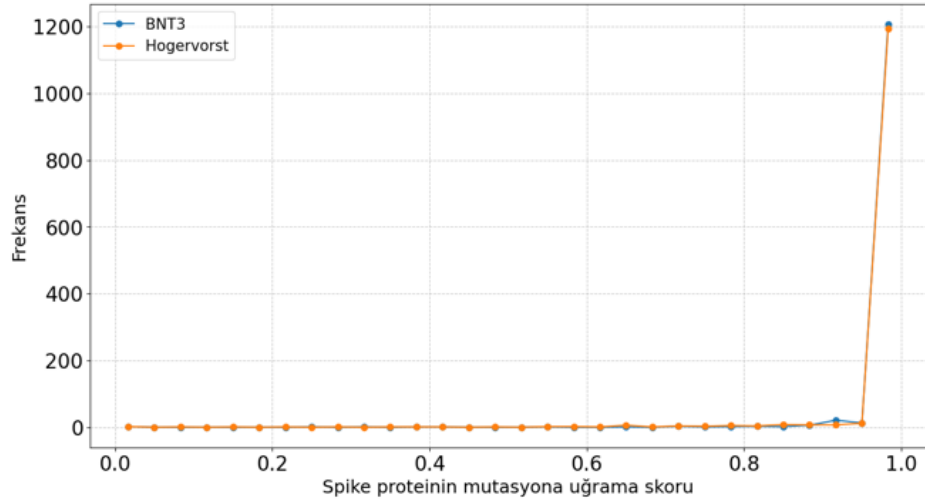
Çizelge 3.11: Gelecek için kullanılacak verinin sekans dağılımı.

Yıl	2022	2023	2024	2025
Sekans Sayısı	131495	83138	34162	1314

Oluşturulan setlere ait protein sekansları hizalandıktan sonra, Omicron ve Delta varyantların tahminlerinde iyi sonuçları gösteren BNT3 ve Hogervorst skorlama fonksiyonları ile mutasyona uğrama skorları elde edilmiştir. Elde edilen mutasyon skorlarının dağılımları incelenmiştir (Şekil 3.28). Dağılımlara bakılınca, S proteinin farklı değerlerde mutasyon skoruna sahip olduğu ve korunan bölgelerin sayısının (1 değerine sahip pozisyonlar) diğer bölgelerden fazla olduğu görülmüştür.

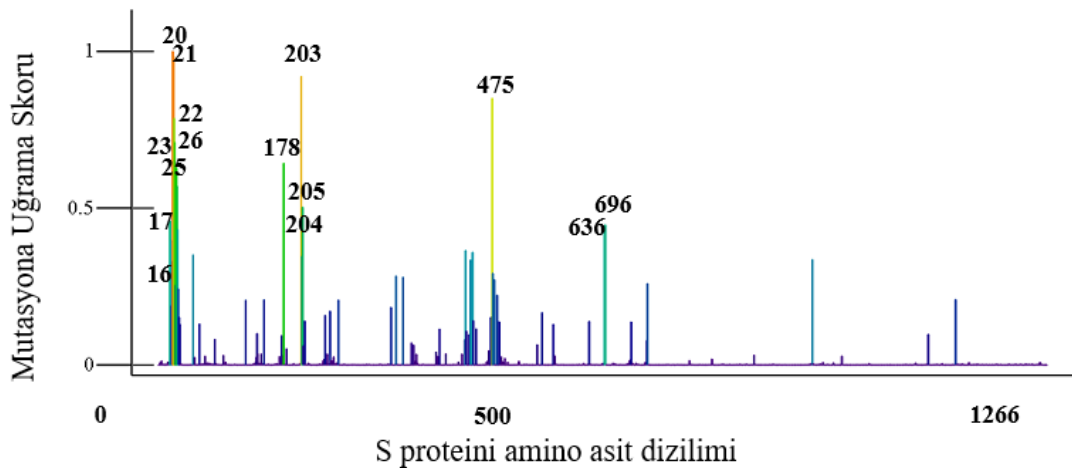
Çizelge 3.12: Gelecek tahminler için yapılan çalışmadaki en değişken 15 bölge.

EN DEĞİŞKEN 15 BÖLGE						
BNT3	17	18	21	22	23	24
	26	27	179	204	205	206
	476	637	697			
Hogervorst	17	18	21	22	23	24
	26	27	28	179	204	206
	476	636	637			



Şekil 3.28: Gelecek tahminler için yapılan çalışmadaki mutasyon skorlarının dağılımı.

Buna ek olarak, her skorlama fonksiyonunun çıktıları kullanılarak, en yüksek değişim gösteren 15 bölge tespit edilmiştir (Çizelge 3.12). Seçilen bütün gruplar için ortak değişken amino asit pozisyonları; 17, 21, 22, 23, 24, 26, 27, 204, 476 pozisyonlarıdır. Bu pozisyonlar tüm tahmin grupları için ortak olduğu için bu bölgelerde mutasyon olma olasılığının daha fazla olduğu söylenebilir. En değişken bölgelerin S proteini üzerindeki dağılımını görsel olarak sunabilmek amacıyla, mutasyona uğrama skorlarının yer aldığı grafik elde edilmiştir. Şekil 3.29’te bu grafik gösterilmiştir. En değişken pozisyonlar incelendiğinde, özellikle 15. ile 30. pozisyonlar arasında belirgin bir yığılma gözlemlenmiştir (Şekil 3.29).



Şekil 3.29: Gelecek tahminler için yapılan çalışmadaki mutasyona uğrama skorların grafiği.

Bu durumu daha iyi anlayabilmek adına ilgili bölgeye yönelik yapılan incelemede, bu bölgede bazı sekanslar varlığını korurken bazı sekansların ise bu bölgede amino asit içermediği tespit edilmiştir ve bu nedenle hizalamada oluşan boşluklardan dolayı bu görüntü çıkmıştır. Bu durum, bu pozisyonlarda yoğun bir değişim varmış izlenimi oluştursa da gerçekte iki farklı grup sekansın varlığına işaret etmektedir. Seçilen setteki sekanslar incelenmiş ve belirli bir bölgeye göre iki ayrı sekans grubu oluşturulmuştur. Her grup için yeniden hizalama yapılmış ve daha önceki tahmin sonuçları göz önünde bulundurularak, en iyi performansı gösteren BNT3, PAM120 ve Hogervorst, BLOSUM62 kombinasyonları kullanılarak analizler gerçekleştirilmiştir. Her grup için yapılan kombinasyonlarda, diğer analizlerde yapıldığı gibi 15 en değişken bölge belirlenmiştir ve bu bölgeler için yapılan tüm tahminler detaylı olarak incelenmiştir. En değişken 15 pozisyon incelendiğinde ise 1.grup hesaplamaların sonucu kendi aralarında, 2.grup sekans grubu ise kendi aralarında tutarlıdır.

Çizelge 3.13: Gelecek varyantlar için en değişken 15 pozisyon.

Set	1	1	2	2
Numarası				
Fonksiyon	BNT3	Hogervorst	BNT3	Hogervorst
Matris	PAM120	Blosum62	PAM120	Blosum62
En Değişken Bölgeler	66	66	18	17
	67	67	19	18
	141	141	22	19
	142	143	24	21
	183	183	25	24
	208	210	26	25
	209	249	31	26
	210	336	143	31
	249	343	144	144
	343	442	145	145
	442	443	209	209
	449	449	210	210
	483	483	481	481
	487	487	641	641
	701	701	642	642

BNT3, PAM120 kombinasyonu, gözlemlenen amino asit frekanslarına dayalı analizlerde başarılı sonuçlar ortaya koymuştur. Öte yandan, Hogervorst, BLOSUM62 kombinasyonu özellikle referans amino asit frekanslarını temel alan hesaplamalarda etkili performans sergilemiştir. Bu doğrultuda, geleceğe yönelik yapılan tahminleme çalışmalarında söz konusu iki kombinasyon tercih edilmiştir.

BNT3, PAM120 ve gözlemlenmiş amino asit frekansı kombinasyonuna ait sonuçlar incelendiğinde, 67. ve 210. pozisyonların en fazla sayıda tahmin edilen konumlar olduğu görülmektedir (Şekil 3.30). 67. pozisyon için en çok tahmin Serin (S) üzerine yapılmıştır. Bu tahmini sırasıyla Treonin (T), Lösin (L) ve Alanin (A) takip etmektedir. 210. pozisyon için ise en çok tahmin Lösin (L) üzerine yapılmış olup, bu tahminleri Serin (S), Treonin (T), Alanin (A), Fenilalanin (F) ve Valin (V) takip etmektedir. 67. pozisyon, referans sekansta Valin (V) amino asidini içermektedir. Daha önceki bölümlerde de belirtildiği üzere, elde edilen dönüşüm matrisi incelendiğinde (Bölüm 2.2.2), Valin (V) amino asidinin Lösin (L) ve Serin (S) amino asitlerine toplamda 24 farklı dönüşüm yolu olduğu gözlemlenmiştir. Bu durum, 67. pozisyonda yapılan tahminlerde L ve S amino asitlerinin daha yüksek sıklıkta gözlemlenme olasılığını desteklemektedir. Dolayısıyla, ilgili pozisyonda bu iki amino asidin öne çıkması, dönüşüm olasılıklarıyla uyumlu bir sonuç olarak değerlendirilebilir (Şekil 3.30).

EN DEĞİŞKEN 15 POZİSYON	REFERANS	TAHMİN DAĞIMLARI																				
	66	H	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	67	V	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	141	Y	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	142	Y	A	C	D		F	G		I	K	L	M	N	P	Q		S	T	V		Y
	183	F	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	208	N		C	D	E	F			I	K	L	M	N	P	Q	R	S	T	V		Y
	209	L	A	C	D	E	F	G		I		L					R	S	T			
	210	G	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	249	G	A		D	E	F	G	H		K	L		N		Q		S		V		Y
	343	R		C	D						K	L						S	T	V		
	442	V	A	C	D	E	F	G		I	K			N				S	T	V		
	449	L	A	C	D	E			H	I		L		N	P	Q		S	T	V		
	483	F	A	C	D	E	F	G	H	I				N	P	Q		S	T	V		Y
	487	F	A	C	D	E	F	G		I		L	M	N	P	Q	R	S	T	V	W	Y
	701	S	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y

Şekil 3.30: Birinci grup sekanslar için BNT3 skarlama fonksiyonu, PAM120 skarlama matrisi ve gözlemlenmiş amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.

Hogervorst – BLOSUM62 – Referans frekansı kombinasyonu incelendiğinde, belirgin bir tahmin yığılması görülmemektedir. 66. pozisyon için Lösin (L) ve 67. pozisyon için Alanin (A) 210.pozisyon içinse N (Asparjin) amino asitlerinin tahmini ağır basmaktadır. Diğer pozisyonlarda amino asit tahminlerinin eşite yakın bir şekilde

dağılması, ilgili bölgelerde bir mutasyon olasılığının bulunduğunu; ancak hangi amino aside dönüşeceği konusunda belirgin bir yönelimin olmadığını, yani tahminlerin farklı amino asitler arasında dengeli şekilde dağıldığını göstermektedir (Şekil 3.31). Bu sonuçlar bağlamında dönüşüm matrisi incelendiğinde, 66. pozisyon, referans sekansta Histidin (H) amino asidini içermekte olup, dönüşüm matrisi doğrultusunda Lösin (L) amino asidine 12 farklı şekilde dönüşebilmektedir. 67. pozisyona ait dönüşüm analizine bakıldığında ise, Valin (V) amino asidinin Alanin (A) amino asidine dönüşme sayısı, Arginin (R), Lösin (L) ve Serin (S) amino asitlerine olan dönüşüm sıklıklarına kıyasla daha düşük sayıda kalmaktadır. Benzer şekilde, 210. pozisyon için referans amino asit olan Glisin (G)'in Asparajin (N) amino asidine dönüşme olasılığı da Arginin (R), Lösin (L) ve Serin (S) dönüşümleriyle karşılaştırıldığında daha sınırlı düzeydedir.

EN DEĞİŞKEN 15 POZİSYON	REFERANS		TAHMİN DAĞILIMLARI																									
	66	H	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y						
	67	V	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y						
	141	Y	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y						
	143	H	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V	W	Y						
	183	F	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V	W	Y						
	210	G	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y						
	249	G	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V	W	Y						
	336	D	A	C	D	E	F	G		I		L		N	P	Q	R	S	T	V								
	343	R		C	D	E	F	G	H	I	K	L	M	N	P	Q		S	T	V	W	Y						
	442	V	A	C	D	E	F	G		I	K	L	M	N	P		R	S	T	V	W	Y						
	443	G	A		D	E	F	G	H	I	K	L		N			R	S	T	V	W	Y						
	449	L	A		D	E	F	G		I	K	L		N	P	Q	R	S	T	V	W	Y						
	483	F	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V	W	Y						
	487	F	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V								
	701	S	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T		W	Y						

Şekil 3.31: Birinci grup sekanslar için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans amino asit frekansı kombinasyonu kullanılarak elde edilen tahmin sonuçları.

Birinci grup sekanslardan elde edilen sonuçlara dayanarak, bu gruba yakın özellikler taşıyan sekansların, ilerleyen dönemlerde tahmin edilen bölgelere ve amino asit mutasyonlarına benzer mutasyon profilleri sergilemesi olasıdır. Özellikle, 67. ve 210. pozisyonlarda mutasyon oluşma ihtimali öne çıkmaktadır. Bu bağlamda gelecek için, 67. pozisyonda Lösin (L) veya Serin (S) amino asitlerinin, 210. pozisyonda ise yine Lösin (L) veya Serin (S) amino asitlerinin görülme olasılığı yüksek görülmektedir.

REFERANS		TAHMİN DAĞIMLARI																				
EN DEĞİŞKEN 15 POZİSYON	18	L		C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V	W	Y
	19	R	A	C	D	E	F	G	H	I	K	L	M	N	P	Q		S	T	V		Y
	22	T	A				F	G	H	I	K	L	M	N	P	Q		S	T	V		Y
	24	L	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	25	P	A	C	D	E	F	G	H	I	K	L		N	P	Q	R	S	T	V		Y
	26	P	A	C	D	E	F	G	H	I	K	L	M	N		Q	R	S		V		Y
	31	S	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	143	V	A	C	D		F	G	H	I	K	L		N	P	Q	R	S	T	V		Y
	144	Y	A	C	D	E	F	G		I	K	L	M	N	P	Q	R	S	T	V		Y
	145	Y	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	209	N	A	C	D	E	F	G		I	K	L		N	P	Q	R	S	T	V		Y
	210	L	A			E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V		Y
	481	V	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	641	F	A	C	D		F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	642	Q	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V		Y

Şekil 3.32: İkinci grup sekanslar için BNT3 skorlama fonksiyonu, PAM120 skorlama matrisi ve gözlemlenmiş amino asit frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.

İkinci grup sekanslar incelendiğinde, BNT3-PAM120-gözlemlenmiş frekans kombinasyonu ile yapılan analizlerde, en fazla tahminin 31. pozisyon için yapıldığı görülmektedir. 31. pozisyona ait tahminlerde, bu bölgede Lösin (L), Treonin (T) ve Alanin (A) tahminlerinin diğer tahminlere göre baskın olduğu görülürken, aynı zamanda 31. pozisyonun orijinal amino asidi olan Serin'e (S) dönüşme eğilimi de mevcuttur. Diğer dağılımlarda ise genellikle tahminlerde baskın bir dağılım görülmemekle beraber bazı bölgeler için de eşit dağılım görülmektedir. Bu sonuçlar bağlamında dönüşüm matrisi incelendiğinde, Serin (S) amino asidinin Lösin (L) amino asidine 36 farklı dönüşüm yoluyla dönüşebilmektedir. Bu sayı, Serinin diğer amino asitlere dönüşüm yollarına kıyasla daha fazladır. Dolayısıyla, Serin (S)'den Lösin (L)'ne olan dönüşüm, Serin için en yüksek dönüşüm potansiyelini yansıtmaktadır ve bu durum 31. pozisyonda Lösin (L) amino asidinin görülme olasılığını artırmaktadır (Şekil 3.32).

EN DEĞİŞKEN 15 POZİSYON	REFERANS		TAHMİN DAĞIMLARI																					
	17	N	A	C	D	E	F	G			I	K	L		N	P	Q		S	T	V		Y	
	18	L	A	C	D	E	F	G	H		I	K	L	M	N	P	Q		S	T	V		Y	
	19	R	A	C	D	E		G	H		I	K	L		N	P	Q	R	S	T	V	W	Y	
	21	R		C	D	E	F	G			I	K	L	M	N	P		R	S	T	V		Y	
	24	L		C	D	E	F	G	H		I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	25	P	A		D	E	F	G	H		I	K	L		N	P	Q	R	S	T	V		Y	
	26	P	A	C	D	E	F	G	H		I	K	L		N	P		R	S	T	V	W	Y	
	31	S	A	C	D	E	F	G	H		I	K	L		N	P	Q	R	S	T	V		Y	
	144	Y	A	C	D	E	F	G	H		I	K	L	M	N	P	Q	R	S	T	V		Y	
	145	Y	A	C	D	E	F	G			I	K	L		N	P	Q	R	S	T	V		Y	
	209	N	A	C	D	E	F		H		I	K	L		N	P	Q	R	S	T	V		Y	
	210	L	A	C	D	E	F	G			I	K	L	M	N	P	Q	R			V		Y	
	481	V	A		D	E	F	G	H		I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	641	F	A	C	D	E	F	G			I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	642	Q	A	C	D	E	F	G	H		I	K	L		N	P	Q	R	S	T	V		Y	

Şekil 3.33: İkinci grup sekanslar için Hogervorst skorlama fonksiyonu, BLOSUM62 skorlama matrisi ve referans frekansı kombinasyonu kullanılarak yapılan tahmin sonuçları.

Hogervorst fonksiyonu ve BLOSUM62 matrisine dayalı referans frekansı analizine göre, bazı pozisyonlarda belirli amino asit değişim oranının yüksek olduğu gözlemlenmektedir. Özellikle 31. pozisyonda yer alan amino asit Lösin (L) olup, bu pozisyon için yapılan tahminlerde de Lösin öne çıkmaktadır. Bununla birlikte, 642. pozisyon için yapılan tahminlerde ise Fenilalanin (F), Lösin (L) ve Serin (S) amino asitlerinin öne çıktığı görülmektedir. Bu sonuçlar doğrultusunda dönüşüm matrisi incelendiğinde, 31. pozisyon için referans amino asit olan serin (S)'in, Lösin (L) amino asidine 36 farklı yolla dönüşebildiği belirlenmiştir. Bu yüksek dönüşüm sayısı, ilgili pozisyonda $S \rightarrow L$ mutasyonunun görülme olasılığını artırmakta ve bu bölgenin potansiyel bir mutasyon noktası olabileceğini düşündürmektedir. Benzer şekilde, 642. pozisyon özelinde yapılan analizlerde, referans amino asit olan Glutamin (Q)'un, özellikle L ve S amino asitlerine dönüşme olasılığının, diğer amino asitlere kıyasla daha yüksek olduğu gözlemlenmiştir. Bu durum, 642. pozisyonda da $Q \rightarrow L$ ve $Q \rightarrow S$ mutasyonlarının daha muhtemel olduğunu göstermektedir (Şekil 3.33).

İkinci grup sekanslara ait tahminler ve dönüşüm matrisi analizleri temel alındığında, bu gruba yakınsayan sekansların, ilerleyen zamanlarda 31. pozisyonda bir mutasyon geçirme olasılığı öne çıkmaktadır. Yapılan değerlendirmeler, bu mutasyonun büyük olasılıkla Lösin (L) amino asidine doğru gerçekleşeceğini göstermektedir. Bu bulgu, dönüşüm matrisinde serin (S) amino asidinin L amino asidine yüksek dönüşüm olasılığı ile de desteklenmektedir.

4. SONUÇ VE ÖNERİLER

Biyoinformatik çalışmaları, gen dizilimleri sayesinde birçok hastalığın kaynağının anlaşılmasına katkı sağlamıştır. Hastalık yapıcı patojenlerdeki mutasyon istatistiklerinin hesaplanabilir olması ise, birçok hastalığa karşı önceden önlem alınmasına imkân tanımaktadır. Bu nedenle, biyoinformatik tabanlı çalışmalar, genetik hastalıkların ve anomalilerin incelenmesinin yanı sıra olası genetik değişimlerinin belirlenip, bunlara karşı önlem alınmasında kritik bir rol oynamaktadır. Bu çalışma kapsamında SARS-COV-2'ye ait yapısal proteinlerinden birisi olan S proteininde sonradan ortaya çıkabilecek mutasyonların önceden belirlenebilmesi açısından bir model geliştirilmiştir. Tahmin yapılacak modeller için öncelikle, veri bankalarından S proteinine ait protein sekansları elde edilmiştir. Elde edilen sekansların hesaplama açısından fazla sekans içermesinden dolayı rastgele seçime dayanan bir yaklaşım ile farklı zaman periyotları için örneklem hazırlanmıştır. Seçilen setler ile skoreleme fonksiyonları ile mutasyon skorları elde edilmiştir. Skor değerleri 2 grup olarak: 1. grup Karlin, BNT2, BNT3 ve Hogervorst olarak ele alınırken 2. Grup Sander ve Valdar ele alınmıştır. Birinci grup skor fonksiyonları, tahmin modeli için daha uygun girdiler verdiği için hesaplamalara 1. grup skor fonksiyonları ile devam edilmiştir. Bu skora ek olarak literatürde değişim miktarını belirlemek için kullanılan entropi değerleri de her set için ve her pozisyon için hesaplanmıştır. Elde edilen skor ve entropi değerleri ile en değişken 15 pozisyon tespit edilmiş ve tahmin algoritması olan rastgele yürüyüş algoritmasında kullanılmıştır. Rastgele yürüyüş algoritması kullanılırken pozisyonların mutasyona uğrama skorları ve amino asit frekansları kullanılmıştır. Amino asit frekansı için 4 farklı yöntem kullanılmıştır. İlk olarak kodon tablosundan elde edilen doğal frekans, ikinci olarak her set için hesaplanan gözlemlenen frekanslardır. Bu frekanslara ek olarak kodon tablosundaki dönüşümlerin mutasyon geçirme olasılıklarına göre dönüşümleri hesaplanmış, dönüşüm frekansları elde edilmiştir ve son olarak referans frekansı olarak da referans olarak belirlenen sekanstan referans amino asit frekansı elde edilmiştir. Rastgele yürüyüş (random walk) algoritmasıyla elde edilen tahminlerin, gerçekte ne kadarlık

bir zaman aralığını temsil ettiğini belirlemek amacıyla SARS-CoV-2'nin mutasyon hızı dikkate alınmıştır. Mutasyon hızı, literatürde yaygın olarak kabul gören S genine ait nükleotid düzeyindeki değişim oranı, yani her bir nükleotid pozisyonunda meydana gelebilecek değişikliklerin sıklığı, 8.066×10^{-4} olarak hesaplanmıştır. Bu oran kullanılarak yapılan olasılık temelli hesaplamalarda, bir yıl içinde 1273 amino asit içeren S proteininde rastgele seçilen bir pozisyonda mutasyon oluşma olasılığı %95 olarak hesaplanmıştır. Yapılan tahminlerin 1 yıl sonrasına ait olduğu kabul edilmiştir. Rastgele yürüyüş algoritması ile elde edilen tahminler 2 farklı şekilde incelenmiştir. Pozisyon bazlı tahminler ve amino asit bazlı tahminler olarak incelenmiştir. Bu hesaplamalara ek olarak, belirlenen setler kullanılarak entropi değerleri hesaplanmış ve bu değerler üzerinden de mutasyon tahminleri yapılmıştır. Entropi, pozisyonlardaki çeşitlilik miktarını ifade ettiği için, potansiyel olarak mutasyon geçirme olasılığı yüksek bölgelerin belirlenmesinde kullanılmıştır ancak, entropi kullanılarak yapılan tahmin sonuçları skorumla fonksiyonlarıyla yapılan tahminlere kıyasla daha düşük doğruluk göstermiştir. Bu nedenle, gelecek tahmini için yapılan çalışmalarda entropi temelli yöntemler yerine skorumla fonksiyonları kullanılarak hesaplamalar yapılmıştır. Omicron varyantını tespit etmek için yapılan tahminler sırasında, Omicron varyantına ait genetik bilgide farklı türlerden gelen genetik materyallerin etkisi olabileceği varsayılarak yöntemin doğrulamasını yapmak adına aynı işlemler Delta varyantı için de yapılmıştır. Bu doğrulama çalışması için Alfa, Beta, Gamma varyantları kullanılarak Delta varyantı tespit edilmeye çalışılmıştır. Delta varyantı üzerinde yapılan tahmin sonuçları hem mutasyon bölgesi hem de mutasyon sonucu oluşan amino asitler açısından incelenmiş ve yöntemi doğrular nitelikte olduğu gözlenmiştir. Bu tez kapsamında seçilen veri setleri ile yapılan tahmin ve doğrulama analizlerinde, çeşitli skorumla fonksiyonları ve matrislerinin performansları karşılaştırılmıştır. Hogervorst skorumla fonksiyonu, BLOSUM62 skorumla matrisi ile referans frekans kullanılarak yapılan hesaplamalar başarılı sonuçlar vermiştir. Benzer şekilde, BNT3 skorumla fonksiyonları, PAM120 skorumla matrisi ve gözlemlenen amino asit frekansı ile kullanıldığında iyi performans göstermiştir. Bu tez kapsamında, zaman kısıtı nedeniyle yalnızca belirli sayıda sekans üzerinde analiz yapılmıştır. Ancak ilerleyen çalışmalar kapsamında, sekans sayısı artırılarak daha kapsamlı bir inceleme gerçekleştirilerek farklı pozisyonlar tespit edilebilir. Bu yaklaşım, mutasyon eğilimlerini daha iyi anlamak ve gelecekte ortaya çıkabilecek olası mutasyonları öngörebilmek açısından kapsamlı analiz sunmaktadır. Özellikle bazı pozisyonlarda

tekrarlayan mutasyon örüntüleri, bu bölgelerin evrimsel baskı altında olabileceğini ve gelecekteki mutasyon tahminlerinde kritik rol oynayabileceğini göstermektedir ancak bu yaklaşım, setlerin içerisinde benzer veya tekrarlayan sekanslar içermesi nedeniyle sınırlı kalmıştır. İlerleyen çalışmalarda, benzersiz sekanslardan oluşan setler oluşturularak farklı sekansların tahminlere etkilerinin daha ayrıntılı şekilde incelenmesi önerilmektedir. Buna ek olarak, her bir sekans farklı sayıda amino asit içerdiğinden, BLOSUM62 ve PAM120 gibi genel matrisler yerine, SARS-CoV-2 virüsünün evrimsel sürecine özgü amino asit dönüşüm matrisleri gibi vektör-spesifik matrisler de hesaplanarak oluşturulabilir, bu özel matrisler üzerinden tahminler yapılarak sonuçların analiz edilmesi daha güvenilir sonuçlar elde edilmesini sağlayacaktır. Bu tez kapsamında rastgele yürüyüş algoritması için 2000 tahmin sayısı belirlenmiştir ancak ilerleyen çalışmalarda bu sayı artırılarak gözlemler yapılması planlanmaktadır. Gelecekteki çalışmalarda analiz edilen pozisyon sayısının 15'in üstüne çıkarılması ve amino asit tahminlerinde değişimine bakılması doğru olacaktır.



KAYNAKLAR

- Aboul-Fotouh, S., Mahmoud, A.N., Elnahas, E.M. vd.** What are the current anti-COVID-19 drugs? From traditional to smart molecular mechanisms. *Virology* 20, 241 (2023). <https://doi.org/10.1186/s12985-023-02210-z>
- Adedeji, W. A. (2016).** The treasure called antibiotics. *Annals of Ibadan Postgraduate Medicine*, 14(2), 56-57.
- Aiello TF, García-Vidal C, Soriano A.** Antiviral drugs against SARS-CoV-2. *Rev Esp Quimioter.* 2022 Oct;35 Suppl 3(Suppl 3):10-15. doi: 10.37201/req/s03.03.2022. Epub 2022 Oct 24. PMID: 36285850; PMCID: PMC9717461.
- Akkız, H. (2022).** The biological functions and clinical significance of SARS-CoV-2 variants of concern. *Frontiers in Medicine*, 9, 849217. <https://doi.org/10.3389/fmed.2022.849217>
- Aminov, R. I. (2010).** A brief history of the antibiotic era: Lessons learned and challenges for the future. *Frontiers in Microbiology*, 1, 134. <https://doi.org/10.3389/fmicb.2010.00134>
- Apweiler, R., Bairoch, A., & Wu, C. H. (2004).** Protein sequence databases. *Current Opinion in Chemical Biology*, 8(1), 76–80. <https://doi.org/10.1016/j.cbpa.2003.12.004>
- Arora, P., Pöhlmann, S. & Hoffmann, M.** Mutation D614G increases SARS-CoV-2 transmission. *Sig Transduct Target Ther* 6, 101 (2021). <https://doi.org/10.1038/s41392-021-00502-w>
- Atkins, P., & de Paula, J. (2010).** *Physical Chemistry* (9th ed.). Oxford University Press
- Awadasseid A, Wu Y, Tanaka Y, Zhang W. (2021).** Effective drugs used to combat SARS-CoV-2 infection and the current status of vaccines. *Biomed Pharmacother.* 2021 May; 137:111330. doi: 10.1016/j.biopha.2021.111330. Epub 2021 Jan 28. PMID: 33550043; PMCID: PMC7843108.
- Batool, S., Chokkakula, S., Jeong, J. H., Baek, Y. H., & Song, M.-S. (2025).** SARS-CoV-2 drug resistance and therapeutic approaches. *Heliyon*, 11(2), e41980. <https://doi.org/10.1016/j.heliyon.2025.e41980>
- Bayat, A. (2002).** *Bioinformatics*. *BMJ*, 324(7344), 1018–1022. doi: 10.1136/bmj.324.7344.1018
- Berg, J. M., Tymoczko, J. L., & Stryer, L. (2002).** *Biochemistry* (5th ed.). W. H. Freeman.
- Blundell, S. J., & Blundell, K. M. (2016).** *Concepts in Thermal Physics* (2nd ed.). Oxford University Press.
- Bule, M., Khan, F., & Niaz, K. (2019).** Antivirals: Past, present and future. In *Recent Advances in Animal Virology* (pp. 425–446). Springer. https://doi.org/10.1007/978-981-13-9073-9_22
- Cann, A. J. (2011).** *Principles of molecular virology* (5th ed.). Academic Press.

- Cao, B., Wang, Y., Wen, D., Liu, W., Wang, J., Fan, G., ... & Wang, C. (2020).** A trial of lopinavir–ritonavir in adults hospitalized with severe Covid-19. *New England Journal of Medicine*, 382(19), 1787–1799. <https://doi.org/10.1056/NEJMoa2001282>
- Casella, G., Fienberg, S., & Olkin, I. (2011).** *Probability and statistics for machine learning*. Springer.
- Choi JY, Smith DM. (2021)** SARS-CoV-2 Variants of Concern. *Yonsei Med J.* 2021 Nov;62(11):961-968. doi: 10.3349/ymj.2021.62.11.961. PMID: 34672129; PMCID: PMC8542474.
- Clardy, J., Fischbach, M. A., & Walsh, C. T. (2006).** New antibiotics from bacterial natural products. *Nature biotechnology*, 24(12), 1541–1550. <https://doi.org/10.1038/nbt1266>
- Clausius, R. (1865).** *The Mechanical Theory of Heat*. London: John van Voorst.
- Codling, E. A., Plank, M. J., & Benhamou, S. (2008).** Random walk models in biology. *Journal of the Royal Society Interface*, 5(25), 813–834. <https://doi.org/10.1098/rsif.2008.0014>
- Cover, T. M., & Thomas, J. A. (2006).** *Elements of Information Theory* (2nd ed.). Wiley.
- Crits-Christoph, A., Keim, P., Li, D., Lu, J., Zhen, Q., Tabing, D., ... & Andersen, K. G. (2024).** Genetic tracing of market wildlife and viruses at the epicenter of the COVID-19 pandemic. *Cell*, 187(19), 5468–5482.e11. <https://doi.org/10.1016/j.cell.2024.08.026>
- Davies, J., & Davies, D. (2010).** Origins and Evolution of Antibiotic Resistance. *Microbiology and Molecular Biology Reviews*, 74(3), 417–433.
- Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978).** A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (Vol. 5, pp. 345–352). National Biomedical Research Foundation.
- De Clercq, E. (2002).** Strategies in the design of antiviral drugs. *Nature Reviews Drug Discovery*, 1(1), 13–25. <https://doi.org/10.1038/nrd703>
- Deng X, Garcia-Knight MA, Khalid MM, Servellita V, Wang C, Morris MK, Sotomayor-González A, Glasner DR, Reyes KR, Gliwa AS, Reddy NP, Sanchez San Martin C, Federman S, Cheng J, Balcerek J, Taylor J, Streithorst JA, Miller S, Sreekumar B, Chen PY, Schulze-Gahmen U, Taha TY, Hayashi JM, Simoneau CR, Kumar GR, McMahon S, Lidsky PV, Xiao Y, Hemarajata P, Green NM, Espinosa A, Kath C, Haw M, Bell J, Hacker JK, Hanson C, Wadford DA, Anaya C, Ferguson D, Frankino PA, Shivram H, Lareau LF, Wyman SK, Ott M, Andino R, Chiu CY. (2021).** Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell*. 2021 Jun 24;184(13):3426-3437.e8. doi: 10.1016/j.cell.2021.04.025. Epub 2021 Apr 20. PMID: 33991487; PMCID: PMC8057738.
- Flores-Vega, V. R., Monroy-Molina, J. V., Jiménez-Hernández, L. E., Torres, A. G., Santos-Preciado, J. I., & Rosales-Reyes, R. (2022).** SARS-CoV-2: Evolution and Emergence of New Viral Variants. *Viruses*, 14(4), 653. <https://doi.org/10.3390/v14040653>
- Gonnet, G. H., Cohen, M. A., & Benner, S. A. (1992).** Exhaustive matching of the entire protein sequence database. *Science*, 256(5062), 1443–1445.
- Grimmett, G., & Stirzaker, D. (2001).** *Probability and Random Processes* (3rd ed.). Oxford University Press.

- Guruprasad, L., Naresh, G. K. R. S., & Boggarapu, G. (2023).** Taking stock of the mutations in human SARS-CoV-2 spike proteins: From early days to nearly the end of COVID-19 pandemic. *Current Research in Structural Biology*, 6, 100107. <https://doi.org/10.1016/j.crstbi.2023.100107>
- Halawa EM, Fadel M, Al-Rabia MW, Behairy A, Nouh NA, Abdo M, Olga R, Fericean L, Atwa AM, El-Nablaway M, Abdeen A. (2024).** Antibiotic action and resistance: updated review of mechanisms, spread, influencing factors, and alternative approaches for combating resistance. *Front Pharmacol.* 2024 Jan 12; 14:1305294. doi: 10.3389/fphar.2023.1305294. PMID: 38283841; PMCID: PMC10820715.
- Halawa, E. M., Fadel, M., Al-Rabia, M. W., Behairy, A., Nouh, N. A., Abdo, M., Olga, R., Fericean, L., Atwa, A. M., El-Nablaway, M., & Abdeen, A. (2024).** Antibiotic action and resistance: Updated review of mechanisms, spread, influencing factors, and alternative approaches for combating resistance. *Frontiers in Pharmacology*, 14, 1305294. <https://doi.org/10.3389/fphar.2023.1305294>
- Haritha, P., Shanmugavadivu, P., & Dhamodharan, S. (2018).** A comprehensive review on protein sequence analysis techniques. *International Journal of Computer Sciences and Engineering*, 6(7), 1–10. <https://www.ijcseonline.org>
- Harvey, W. T., Carabelli, A. M., Jackson, B., vd. (2021).** SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19, 409–424. <https://doi.org/10.1038/s41579-021-00573-0>
- Harvey, W.T., Carabelli, A.M., Jackson, B. vd. (2021).** SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 19, 409–424. <https://doi.org/10.1038/s41579-021-00573-0>
- Henikoff, S., & Henikoff, J. G. (1992).** Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22), 10915–10919. <https://doi.org/10.1073/pnas.89.22.10915>
- Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., ... & Pöhlmann, S. (2020).** SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181(2), 271–280.e8. <https://doi.org/10.1016/j.cell.2020.02.052>
- Ibrahim, R. (2021).** Mutation: A review. *ACE Research Journal of Microbiology and Biotechnology*, 1(3), 66–70. Retrieved from <https://www.researchgate.net/publication/353830498>
- Jangra S, Ye C, Rathnasinghe R, Stadlbauer D; (2021).** Personalized Virology Initiative study group; Krammer F, Simon V, Martinez-Sobrido L, García-Sastre A, Schotsaert M. SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *Lancet Microbe*. 2021 Jul;2(7):e283-e284. doi: 10.1016/S2666-5247(21)00068-9. Epub 2021 Apr 7. PMID: 33846703; PMCID: PMC8026167.
- Jangra, M., Travin, D.Y., Aleksandrova, E.V. vd. A broad-spectrum lasso peptide antibiotic targeting the bacterial ribosome. Nature (2025).** <https://doi.org/10.1038/s41586-025-08723-7>
- Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992).** The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3), 275–282.

- Kausar, S., Khan, F. S., Mujeeb Ur Rehman, M. I., Akram, M., Riaz, M., Rasool, G., Khan, A. H., Saleem, I., Shamim, S., & Malik, A. (2021).** A review: Mechanism of action of antiviral drugs. *International Journal of Immunopathology and Pharmacology*, 35, 1–12. <https://doi.org/10.1177/20587384211002621>
- Klassen, S. A., Senefeld, J. W., Johnson, P. W., Carter, R. E., Wiggins, C. C., Shoham, S., Grossman, B. J., Henderson, J. P., Musser, J., Salazar, E., Hartman, W. R., Bouvier, N. M., Liu, S. T. H., Pirofski, L., Baker, S. E., van Helmond, N., Wright, R. S., Fairweather, D., Bruno, K. A., ... Joyner, M. J. (2021).** The effect of convalescent plasma therapy on mortality among patients with COVID-19: Systematic review and meta-analysis. *Mayo Clinic Proceedings*, 96(5), 1262–1275. <https://doi.org/10.1016/j.mayocp.2021.02.008>
- Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., Angyal, A., Brown, R. L., Carrilero, L., Green, L. R., ... Montefiori, D. C. (2020).** Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, 182(4), 812–827.e19. <https://doi.org/10.1016/j.cell.2020.06.043>
- LaTourrette, K.; Garcia-Ruiz, (2022).** H. Determinants of Virus Variation, Evolution, and Host Adaptation. *Pathogens* 2022, 11, 1039. <https://doi.org/10.3390/pathogens11091039>
- Lauring, A. S., Frydman, J., & Andino, R. (2013).** The role of mutational robustness in RNA virus evolution. *Nature Reviews Microbiology*, 11(5), 327–336.
- Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., ... Tan, W. (2020).** Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *The Lancet*, 395(10224), 565–574. [https://doi.org/10.1016/S0140-6736\(20\)30251-8](https://doi.org/10.1016/S0140-6736(20)30251-8)
- Machová Klimešová, M. (2021).** Phylogenetic trees and other evolutionary diagrams in biology textbooks and their importance in secondary science education. *Scientia in Education*, 12(1), 16–36. <https://doi.org/10.14712/18047106.1923>
- MacKay, D. J. C. (2003).** *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.
- Markov, P.V., Ghafari, M., Beer, M. vd. The evolution of SARS-CoV-2. Nat Rev Microbiol 21, 361–379 (2023).** <https://doi.org/10.1038/s41579-023-00878-2>
- Mathura, V. S., & Kanguane, P. (2009).** *Bioinformatics: A concept-based introduction*. Springer. doi:10.1007/978-0-387-84870-9
- Mingaleeva, R.N., Nigmatulina, N.A., Sharafetdinova, L.M. vd. (2022).** Biology of the SARS-CoV-2 Coronavirus. *Biochemistry Moscow* 87, 1662–1678. <https://doi.org/10.1134/S0006297922120215>
- Minkoff, J.M., tenOever, B. Innate immune evasion strategies of SARS-CoV-2. Nat Rev Microbiol 21, 178–194 (2023).** <https://doi.org/10.1038/s41579-022-00839-1>
- Mushebenge, A. G.-A., Ugbaja, S. C., Mbatha, N. A., Khan, R. B., & Kumalo, H. M. (2023).** A Comprehensive Analysis of Structural and Functional

- Changes Induced by SARS-CoV-2 Spike Protein Mutations. *COVID*, 3(9), 1454-1472. <https://doi.org/10.3390/covid3090100>
- Naqvi, A. A. T., Fatima, K., Mohammad, T., Fatima, U., Singh, I. K., Singh, A., Atif, S. M., Hariprasad, G., Hasan, G. M., & Hassan, M. I. (2020).** Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1866(10), 165878. <https://doi.org/10.1016/j.bbadis.2020.165878>
- Niknam, Z., Jafari, A., Golchin, A., Danesh Pouya, F., Nemati, M., Rezaei-Tavirani, M., & Rasmi, Y. (2022).** Potential therapeutic options for COVID-19: An update on current evidence. *European Journal of Medical Research*, 27(1), 6. <https://doi.org/10.1186/s40001-021-00626-3>
- Norris, J. R. (1998).** Markov Chains. Cambridge University Press.
- Okada, P., Buathong, R., Phuygun, S., Thanadachakul, T., Parnmen, S., Wongboot, W., Waicharoen, S., Wacharapluesadee, S., Uttayamakul, S., Vachiraphan, A., Chittaganpitch, M., Mekha, N., Janejai, N., Iamsirithaworn, S., Lee, R. T. C., & Maurer-Stroh, S. (2020).** Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. *Eurosurveillance*, 25(8), 2000097. <https://doi.org/10.2807/1560-7917.ES.2020.25.8.2000097>
- Payne, S. (2017).** Virus evolution and genetics. *Viruses*, 81–86. <https://doi.org/10.1016/B978-0-12-803109-4.00008-8>
- Phan, T., Zitzmann, C., Chew, K. W., Smith, D. M., Daar, E. S., Wohl, D. A., Eron, J. J., Currier, J. S., Hughes, M. D., Choudhary, M. C., Deo, R., Li, J. Z., Ribeiro, R. M., Ke, R., Perelson, A. S., & ACTIV-2/A5401 Study Team. (2024).** Modeling the emergence of viral resistance for SARS-CoV-2 during treatment with an anti-spike monoclonal antibody. *PLoS Pathogens*, 20(4), e1011680. <https://doi.org/10.1371/journal.ppat.1011680>
- Saberian, M., Karimi, E., Khademi, Z., vd. (2022).** SARS-CoV-2: Phenotype, genotype, and characterization of different variants. *Cellular and Molecular Biology Letters*, 27, 50. <https://doi.org/10.1186/s11658-022-00352-6>
- Saparov, A., & Zech, M. (2025).** Big data and transformative bioinformatics in genomic diagnostics and beyond. *Parkinsonism & Related Disorders*, 134, 107311. <https://doi.org/10.1016/j.parkreldis.2025.107311>
- Shah, M., & Woo, H. G. (2022).** Omicron: A heavily mutated SARS-CoV-2 variant exhibits stronger binding to ACE2 and potentially escapes approved COVID-19 therapeutic antibodies. *Frontiers in Immunology*, 12, 830527. <https://doi.org/10.3389/fimmu.2021.830527>
- Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., ... & Li, F. (2020).** Structural basis of receptor recognition by SARS-CoV-2. *Nature*, 581(7807), 221–224. <https://doi.org/10.1038/s41586-020-2179-y>
- Shannon, C. E., & Weaver, W. (1949).** The mathematical theory of communication. University of Illinois Press.
- Sievers, F., Wilm, A., Dineen, D. G., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011).** Fast, scalable generation of high-quality protein

- multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, 539. <https://doi.org/10.1038/msb.2011.75>
- Smaoui, M.R., Yahyaoui, H. (2021).** Unraveling the stability landscape of mutations in the SARS-CoV-2 receptor-binding domain. *Sci Rep* 11, 9166 <https://doi.org/10.1038/s41598-021-88696-5>
- Smith, J., & Doe, J. (2021).** Mutation: A review. ResearchGate. Retrieved from https://www.researchgate.net/publication/353830498_Mutation_A_Review
- Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Diggins, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veasley, D., & Bloom, J. D. (2020).** Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, 182(5), 1295–1310.e20. <https://doi.org/10.1016/j.cell.2020.08.012>
- Strasfeld, L., & Chou, S. (2010).** Antiviral drug resistance: Mechanisms and clinical implications. *Infectious Disease Clinics of North America*, 24(2), 413–437. <https://doi.org/10.1016/j.idc.2010.01.001>
- Tao, K., Tzou, P. L., Nouhin, J., vd. (2021).** The biological and clinical significance of emerging SARS-CoV-2 variants. *Nature Reviews Genetics*, 22, 757–773. <https://doi.org/10.1038/s41576-021-00408-x>
- Valdar, W. S. J. (2002).** Scoring residue conservation. *Proteins: Structure, Function, and Genetics*, 48(2), 227–241.
- Wang, L., Li, X., Zhang, H., Wang, J., Jiang, D., Xue, Z., & Wang, Y. (2025).** A comprehensive review of protein language models. *arXiv*. <https://doi.org/10.48550/arXiv.2502.06881>
- Wang, S., Xu, X., Wei, C., Li, S., Zhao, J., Zheng, Y., Liu, X., Zeng, X., Yuan, W., & Peng, S. (2021).** Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. *Journal of Medical Virology*, 93(12), 6736–6748. <https://doi.org/10.1002/jmv.27331>
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009).** Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9), 1189–1191. <https://doi.org/10.1093/bioinformatics/btp033>
- Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009)** "Jalview Version 2 - a multiple sequence alignment editor and analysis workbench"
- Wei, C., Shan, K. J., Wang, W., Zhang, S., Huan, Q., & Qian, W. (2021).** Evidence for a mouse origin of the SARS-CoV-2 Omicron variant. *Journal of Genetics and Genomics*, 48(12), 1111–1121. <https://doi.org/10.1016/j.jgg.2021.12.003>
- Weiss, G. H. (1994).** Aspects and Applications of the Random Walk. North-Holland.
- Xu, Y., Xu, D., & Liang, J. (Ed.). (2007).** Computational Methods for Protein Structure Prediction and Modeling Volume 1: Basic Characterization. Springer.
- Zhang, J., Cai, Y., Lavine, C. L., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Mayer, M. L., Rits-Volloch, S., Wang, S., Sliz, P., Wesemann, D. R., Yang, W., Seaman, M. S., Lu, J., Xiao, T., & Chen, B. (2022).** Structural and functional impact by SARS-CoV-2 Omicron spike mutations. *Cell Reports*, 39(4), 110729. <https://doi.org/10.1016/j.celrep.2022.110729>

Zhao Z, Xie Y, Bai B, Luo C, Zhou J, Li W, Meng Y, Li L, Li D, Li X, Li X, Wang X, Sun J, Xu Z, Sun Y, Zhang W, Fan Z, Zhao X, Wu L, Ma J, Li OY, Shang G, Chai Y, Liu K, Wang P, Gao GF, Qi J. (2023) Structural basis for receptor binding and broader interspecies receptor recognition of currently circulating Omicron sub-variants. *Nat Commun.* 2023 Jul 21;14(1):4405. doi: 10.1038/s41467-023-39942-z. PMID: 37479708; PMCID: PMC10362042.

Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A. J., Ginn, H. M., ... & Screaton, G. R. (2021). Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell*, 184(9), 2348–2361.e6. <https://doi.org/10.1016/j.cell.2021.02.037>

Url-1 [genome.gov](https://www.genome.gov) alındığı tarih: 07.03.2025

Url-2 <https://www.ncbi.nlm.nih.gov/> alındığı tarih: 01.01.2025

Url-3 <http://www.kbdna.com/publishinglab/struc-bio> alındığı tarih: 07.03.2023

Url-4 <https://www.who.int/> alındığı tarih: 01.01.2025

Url-5 <http://www.sinobiological.com/research/virus/sars-cov-2-spike-mutation#SARS-CoV-2> alındığı tarih: 07.03.2025)

Url-6 <http://www.antibodies-online.com/areas/infectious-disease/covid-19/sars-cov-2-mutations-tracker/> alındığı tarih: 07.05.2024

Url-7 <https://microbenotes.com/protein-databases-types-and-importance/> alındığı tarih: 01.01.2025

Url-8 <http://www.gisaid.org/> alındığı tarih: 01.01.2023

Url-9 <http://www.clustal.org/omega/> alındığı tarih: 29.03.2025



EKLER

EK 1: Skarlama Matrisleri

EK 2: Amino Asit Tablosu



EK 1

Çizelge Ek 1: PAM120 matrisi.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	-
A	3	-3	-1	0	-3	-1	0	1	-3	-1	-3	-2	-2	-4	1	1	1	-7	-4	0	0	-1	-1	-8
R	-3	6	-1	-3	-4	1	-3	-4	1	-2	-4	2	-1	-5	-1	-1	-2	1	-5	-3	-2	-1	-2	-8
N	-1	-1	4	2	-5	0	1	0	2	-2	-4	1	-3	-4	-2	1	0	-4	-2	-3	3	0	-1	-8
D	0	-3	2	5	-7	1	3	0	0	-3	-5	-1	-4	-7	-3	0	-1	-8	-5	-3	4	3	-2	-8
C	-3	-4	-5	-7	9	-7	-7	-4	-4	-3	-7	-7	-6	-6	-4	0	-3	-8	-1	-3	-6	-7	-4	-8
Q	-1	1	0	1	-7	6	2	-3	3	-3	-2	0	-1	-6	0	-2	-2	-6	-5	-3	0	4	-1	-8
E	0	-3	1	3	-7	2	5	-1	-1	-3	-4	-1	-3	-7	-2	-1	-2	-8	-5	-3	3	4	-1	-8
G	1	-4	0	0	-4	-3	-1	5	-4	-4	-5	-3	-4	-5	-2	1	-1	-8	-6	-2	0	-2	-2	-8
H	-3	1	2	0	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-3	-1	-3	1	1	-2	-8
I	-1	-2	-2	-3	-3	-3	-3	-4	-4	6	1	-3	1	0	-3	-2	0	-6	-2	3	-3	-3	-1	-8
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	5	-4	3	0	-3	-4	-3	-3	-2	1	-4	-3	-2	-8
K	-2	2	1	-1	-7	0	-1	-3	-2	-3	-4	5	0	-7	-2	-1	-1	-5	-5	-4	0	-1	-2	-8
M	-2	-1	-3	-4	-6	-1	-3	-4	-4	1	3	0	8	-1	-3	-2	-1	-6	-4	1	-4	-2	-2	-8
F	-4	-5	-4	-7	-6	-6	-7	-5	-3	0	0	-7	-1	8	-5	-3	-4	-1	4	-3	-5	-6	-3	-8
P	1	-1	-2	-3	-4	0	-2	-2	-1	-3	-3	-2	-3	-5	6	1	-1	-7	-6	-2	-2	-1	-2	-8
S	1	-1	1	0	0	-2	-1	1	-2	-2	-4	-1	-2	-3	1	3	2	-2	-3	-2	0	-1	-1	-8
T	1	-2	0	-1	-3	-2	-2	-1	-3	0	-3	-1	-1	-4	-1	2	4	-6	-3	0	0	-2	-1	-8
W	-7	1	-4	-8	-8	-6	-8	-8	-3	-6	-3	-5	-6	-1	-7	-2	-6	12	-2	-8	-6	-7	-5	-8
Y	-4	-5	-2	-5	-1	-5	-5	-6	-1	-2	-2	-5	-4	4	-6	-3	-3	-2	8	-3	-3	-5	-3	-8
V	0	-3	-3	-3	-3	-3	-3	-2	-3	3	1	-4	1	-3	-2	-2	0	-8	-3	5	-3	-3	-1	-8
B	0	-2	3	4	-6	0	3	0	1	-3	-4	0	-4	-5	-2	0	0	-6	-3	-3	4	2	-1	-8
Z	-1	-1	0	3	-7	4	4	-2	1	-3	-3	-1	-2	-6	-1	-1	-2	-7	-5	-3	2	4	-1	-8
X	-1	-2	-1	-2	-4	-1	-1	-2	-2	-1	-2	-2	-2	-3	-2	-1	-1	-5	-3	-1	-1	-1	-2	-8
-	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

Çizelge Ek 2: BLOSUM62 matrisi.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	-
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
-	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Çizelge Ek 3: GONNET matrisi.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	-
A	2.4	-0.6	-0.3	-0.3	0.5	-0.2	0	0.5	-0.8	-0.8	-1.2	-0.4	-0.7	-2.3	0.3	1.1	0.6	-3.6	-2.2	0.1	-0.3	-0.1	0	-5
R	-0.6	4.7	0.3	-0.3	-2.2	1.5	0.4	-1	0.6	-2.4	-2.2	2.7	-1.7	-3.2	-0.9	-0.2	-0.2	-1.6	-1.8	-2	0	0.95	-1	-5
N	-0.3	0.3	3.8	2.2	-1.8	0.7	0.9	0.4	1.2	-2.8	-3	0.8	-2.2	-3.1	-0.9	0.9	0.5	-3.6	-1.4	-2.2	3	0.8	-1	-5
D	-0.3	-0.3	2.2	4.7	-3.2	0.9	2.7	0.1	0.4	-3.8	-4	0.5	-3	-4.5	-0.7	0.5	0	-5.2	-2.8	-2.9	3.45	1.8	-1	-5
C	0.5	-2.2	-1.8	-3.2	11.5	-2.4	-3	-2	-1.3	-1.1	-1.5	-2.8	-0.9	-0.8	-3.1	0.1	-0.5	-1	-0.5	0	-2.5	-2.7	-2	-5
Q	-0.2	1.5	0.7	0.9	-2.4	2.7	1.7	-1	1.2	-1.9	-1.6	1.5	-1	-2.6	-0.2	0.2	0	-2.7	-1.7	-1.5	0.8	2.2	-1	-5
E	0	0.4	0.9	2.7	-3	1.7	3.6	-0.8	0.4	-2.7	-2.8	1.2	-2	-3.9	-0.5	0.2	-0.1	-4.3	-2.7	-1.9	1.8	2.65	-1	-5
G	0.5	-1	0.4	0.1	-2	-1	-0.8	6.6	-1.4	-4.5	-4.4	-1.1	-3.5	-5.2	-1.6	0.4	-1.1	-4	-4	-3.3	0.25	-0.9	-1	-5
H	-0.8	0.6	1.2	0.4	-1.3	1.2	0.4	-1.4	6	-2.2	-1.9	0.6	-1.3	-0.1	-1.1	-0.2	-0.3	-0.8	2.2	-2	0.8	0.8	-1	-5
I	-0.8	-2.4	-2.8	-3.8	-1.1	-1.9	-2.7	-4.5	-2.2	4	2.8	-2.1	2.5	1	-2.6	-1.8	-0.6	-1.8	-0.7	3.1	-3.3	-2.3	-1	-5
L	-1.2	-2.2	-3	-4	-1.5	-1.6	-2.8	-4.4	-1.9	2.8	4	-2.1	2.8	2	-2.3	-2.1	-1.3	-0.7	0	1.8	-3.5	-2.2	-1	-5
K	-0.4	2.7	0.8	0.5	-2.8	1.5	1.2	-1.1	0.6	-2.1	-2.1	3.2	-1.4	-3.3	-0.6	0.1	0.1	-3.5	-2.1	-1.7	0.65	1.35	-1	-5
M	-0.7	-1.7	-2.2	-3	-0.9	-1	-2	-3.5	-1.3	2.5	2.8	-1.4	4.3	1.6	-2.4	-1.4	-0.6	-1	-0.2	1.6	-2.6	-1.5	-1	-5
F	-2.3	-3.2	-3.1	-4.5	-0.8	-2.6	-3.9	-5.2	-0.1	1	2	-3.3	1.6	7	-3.8	-2.8	-2.2	3.6	5.1	0.1	-3.8	-3.25	-1	-5
P	0.3	-0.9	-0.9	-0.7	-3.1	-0.2	-0.5	-1.6	-1.1	-2.6	-2.3	-0.6	-2.4	-3.8	7.6	0.4	0.1	-5	-3.1	-1.8	-0.8	-0.35	-2	-5
S	1.1	-0.2	0.9	0.5	0.1	0.2	0.2	0.4	-0.2	-1.8	-2.1	0.1	-1.4	-2.8	0.4	2.2	1.5	-3.3	-1.9	-1	0.7	0.2	0	-5
T	0.6	-0.2	0.5	0	-0.5	0	-0.1	-1.1	-0.3	-0.6	-1.3	0.1	-0.6	-2.2	0.1	1.5	2.5	-3.5	-1.9	0	0.25	-0.05	0	-5
W	-3.6	-1.6	-3.6	-5.2	-1	-2.7	-4.3	-4	-0.8	-1.8	-0.7	-3.5	-1	3.6	-5	-3.3	-3.5	14.2	4.1	-2.6	-4.4	-3.5	-2	-5
Y	-2.2	-1.8	-1.4	-2.8	-0.5	-1.7	-2.7	-4	2.2	-0.7	0	-2.1	-0.2	5.1	-3.1	-1.9	-1.9	4.1	7.8	-1.1	-2.1	-2.2	-1	-5
V	0.1	-2	-2.2	-2.9	0	-1.5	-1.9	-3.3	-2	3.1	1.8	-1.7	1.6	0.1	-1.8	-1	0	-2.6	-1.1	3.4	-2.55	-1.7	-1	-5
B	-0.3	0	3	3.45	-2.5	0.8	1.8	0.25	0.8	-3.3	-3.5	0.65	-2.6	-3.8	-0.8	0.7	0.25	-4.4	-2.1	-2.55	3.45	1.3	-1	-5
Z	-0.1	0.95	0.8	1.8	-2.7	2.2	2.65	-0.9	0.8	-2.3	-2.2	1.35	-1.5	-3.25	-0.35	0.2	-0.05	-3.5	-2.2	-1.7	1.3	2.65	-1	-5
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-5
-	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

Çizelge Ek 4: PET91 matrisi.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	-
A	2	-1	0	0	-1	-1	-1	1	-2	0	-1	-1	-1	-3	1	1	2	-4	-3	1	0	-1	0	-5
R	-1	5	0	-1	-1	2	0	0	2	-3	-3	4	-2	-4	-1	-1	-1	0	-2	-3	-0.5	1	-1	-5
N	0	0	3	2	-1	0	1	0	1	-2	-3	1	-2	-3	-1	1	1	-5	-1	-2	2.5	0.5	-1	-5
D	0	-1	2	5	-3	1	4	1	0	-3	-4	0	-3	-5	-2	0	-1	-5	-2	-2	3.5	2.5	-1	-5
C	-1	-1	-1	-3	11	-3	-4	-1	0	-2	-3	-3	-2	0	-2	1	-1	1	2	-2	-2	-3.5	-2	-5
Q	-1	2	0	1	-3	5	2	-1	2	-3	-2	2	-2	-4	0	-1	-1	-3	-2	-3	0.5	3.5	-1	-5
E	-1	0	1	4	-4	2	5	0	0	-3	-4	1	-3	-5	-2	-1	-1	-5	-4	-2	2.5	3.5	-1	-5
G	1	0	0	1	-1	-1	0	5	-2	-3	-4	-1	-3	-5	-1	1	-1	-2	-4	-2	0.5	-0.5	-1	-5
H	-2	2	1	0	0	2	0	-2	6	-3	-2	1	-2	0	0	-1	-1	-3	4	-3	0.5	1	-1	-5
I	0	-3	-2	-3	-2	-3	-3	-3	4	2	-3	3	0	-2	-1	1	-4	-2	4	-2.5	-3	-1	-5	
L	-1	-3	-3	-4	-3	-2	-4	-4	-2	2	5	-3	3	2	0	-2	-1	-2	-1	2	-3.5	-3	-1	-5
K	-1	4	1	0	-3	2	1	-1	1	-3	-3	5	-2	-5	-2	-1	-1	-3	-3	-3	0.5	1.5	-1	-5
M	-1	-2	-2	-3	-2	-2	-3	-3	-2	3	3	-2	6	0	-2	-1	0	-3	-2	2	-2.5	-2.5	-1	-5
F	-3	-4	-3	-5	0	-4	-5	-5	0	0	2	-5	0	8	-3	-2	-2	-1	5	0	-4	-4.5	-1	-5
P	1	-1	-1	-2	-2	0	-2	-1	0	-2	0	-2	-2	-3	6	1	1	-4	-3	-1	-1.5	-1	-2	-5
S	1	-1	1	0	1	-1	-1	1	-1	-1	-2	-1	-1	-2	1	2	1	-3	-1	-1	0.5	-1	0	-5
T	2	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	0	-2	1	1	2	-4	-3	0	0	-1	0	-5
W	-4	0	-5	-5	1	-3	-5	-2	-3	-4	-2	-3	-3	-1	-4	-3	-4	15	0	-3	-5	-4	-2	-5
Y	-3	-2	-1	-2	2	-2	-4	-4	4	-2	-1	-3	-2	5	-3	-1	-3	0	9	-3	-1.5	-3	-1	-5
V	1	-3	-2	-2	-2	-3	-2	-2	-3	4	2	-3	2	0	-1	-1	0	-3	-3	4	-2	-2.5	-1	-5
B	0	-0.5	2.5	3.5	-2	0.5	2.5	0.5	0.5	-2.5	-3.5	0.5	-2.5	-4	-1.5	0.5	0	-5	-1.5	-2	3.5	1.5	-1	-5
Z	-1	1	0.5	2.5	-3.5	3.5	3.5	-0.5	1	-3	-3	1.5	-2.5	-4.5	-1	-1	-1	-4	-3	-2.5	1.5	3.5	-1	-5
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-5
-	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

Çizelge Ek 5: Amino asit tablosu.

A	Alanin	Ala	M	Metiyonin	Met
R	Arjinin	Arg	F	Fenilalanin	Phe
N	Asparajin	Asn	P	Prolin	Pro
D	Aspartik Asit	Asp	S	Serin	Ser
C	Sistein	Cys	T	Treonin	Thr
Q	Glutamin	Gln	W	Triptofan	Trp
E	Glutamik Asit	Glu	Y	Tirozin	Tyr
G	Glisin	Gly	V	Valin	Val
H	Histidin	His	B	Asparajin ya da Aspartik Asit	
I	İzolösin	Ile	Z	Glutamin ya da Glutamik Asit	
L	Lösin	Leu	X	Tespit Edilemeyen	
K	Lizin	Lys	-	Boşluk	

