**TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# PROTEİNLERİN MUTASYON HARİTALARININ ÇIKARILARAK EVRİMSEL DEĞİŞİMLERİNİN TAHMİN EDİLMESİ: ÖRNEK OLAY İNCELEMESİ OLARAK NÖRAMİNİDAZ PROTEİNİ (H1N1 VİRÜSÜ)

**YÜKSEK LİSANS TEZİ**

**Elif CANDAŞ**

**Biyomedikal Mühendisliği Anabilim Dalı**

**Tez Danışmanı: Dr. Öğr. Üyesi Ersin Emre ÖREN**

**AĞUSTOS 2019**

Fen Bilimleri Enstitüsü Onayı

…….
**Prof. Dr. Osman EROĞUL**
Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

…….
**Prof. Dr. Osman EROĞUL**
Anabilim Dalı Başkanı

TOBB ETÜ, Fen Bilimleri Enstitüsü'nün 161711030 numaralı Yüksek Lisans Öğrencisi **Elif CANDAŞ**'ın ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"PROTEİNLERİN MUTASYON HARİTALARININ ÇIKARILARAK EVRİMSEL DEĞİŞİMLERİNİN TAHMİN EDİLMESİ: ÖRNEK OLAY İNCELEMESİ OLARAK NÖRAMİNİDAZ PROTEİNİ (H1N1 VİRÜSÜ)"** başlıklı tezi **1 Ağustos 2019** tarihinde aşağıda imzaları olan jüri tarafından kabul edilmiştir.

**Tez Danışmanı :** **Dr. Öğr. Üyesi Ersin Emre ÖREN** .....
TOBB Ekonomi ve Teknoloji Üniversitesi

**Jüri Üyeleri :** **Dr. Öğr. Üyesi Mehmet TAN (Başkan)** .....
TOBB Ekonomi ve Teknoloji Üniversitesi

**Dr. Öğr. Üyesi Aytaç ÇELİK** .....
Sinop Üniversitesi

## 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.

.

Elif Candaş

# ÖZET


Yüksek Lisans

PROTEİNLERİN MUTASYON HARİTALARININ ÇIKARILARAK EVRİMSEL DEĞİŞİMLERİNİN TAHMİN EDİLMESİ: ÖRNEK OLAY İNCELEMESİ OLARAK NÖRAMİNİDAZ PROTEİNİ (H1N1 VİRÜSÜ)

Elif Candaş

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Biyomedikal Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Ersin Emre Ören

Tarih: Ağustos 2019

Tarih boyunca virüsler ve bakteriler dünyada birçok insanın yaşamını yitirmesine sebep olmuştur. Penisilinin bulunması ve sonrasında geliştirilen antibiyotik ve antiviral ilaçlar ile ölümler büyük oranda azaltılabilmiştir. Ancak, 1960'lı yıllardan itibaren bazı virüs ve bakterilerin ilaçlara karşı direnç gösterdikleri gözlenmektedir. Bu virüslere ve bakterilere karşı ilaçların etkinliği azalmakta hatta bazı ilaçlar hiç etki gösterememektedir. Bir ilaç tasarımı ve üretimi için uzun bir süreye ihtiyaç vardır ve bu süre içerisinde değişime uğramış virüslere karşı önlem alınması zor olmaktadır. Bu nedenle virüslerin evrimsel süreçlerinin anlaşılması, ileride karşılaşılabilecek tehlikeli (ilaçlardan etkilenmeyen) virüslere karşı önlem almak büyük bir önem taşımaktadır. Virüslerin evrimsel değişimi genetik materyallerinde gerçekleşen mutasyonlar yani değişimlerle meydana gelmektedir. Genetik materyalde (DNA ya da RNA) gerçekleşen mutasyonlar, proteinlerin yapılarında değişikliğe sebep olabilmektedir. Proteinlerin amino asit dizilimlerindeki değişiklikleri tanımlamak için oluşturulmuş skorlama fonksiyonları vardır. Bu tezde farklı skorlama fonksiyonları biyolojik ve

matematiksel özellikleri ile birlikte açıklanmakta ve birbirleri arasındaki ilişkiler yorumlanmaktadır. Bu bilgiden yararlanarak evrimsel süreçte oluşabilecek olası sekansların tahmin yöntemlerinden bahsedilmektedir. Örnek olay incelemesi olarak birçok ölüme sebep olan domuz gribi virüsü H1N1 kullanılmaktadır. İleride oluşabilecek protein sekanslarını/yapılarını tahmin etmek, antiviral ilaçlara karşı direnç mekanizmalarını anlamak ve yeni ilaç tasarlamak için yol gösterecektir.

**Anahtar Kelimeler:** Protein sekans analizi, Skorlama fonksiyonu, Mutasyon haritası, Nöraminidaz

## ABSTRACT

Master of Science

CALCULATION OF PROTEIN MUTABILITY LANDSCAPE AND THEREON FORECASTING EVOLUTIONARY PATHWAYS: NEURAMINIDASE OF H1N1 VIRUS AS A CASE STUDY

Elif Candaş

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

Supervisor: Assist. Prof. Dr. Ersin Emre Ören

Date: August 2019

Viruses and bacteria have been among the most harmful agents for human health in history. Many lives have been saved with the discovery of antibiotics and antiviral drugs. However, the rapid emergence of resistant strains became an ever-increasing health concern since the 1960s. These resistant strains are capable of inactivating the drug efficacy and survive in infected cells successfully.

Therefore, it is very important to analyze evolutionary pathways of viruses and understand their susceptibility and robustness to mutation with generating mutability landscape. Thus, predicting future mutant strains with the help of mutability probabilities is a potential to discover new drug candidates before emerging as a threat for humans. Despite decades of research, forecasting evolutionary pathways remains extremely challenging due to lack of both available data and appropriate methods. So far, amino acid frequency and substitution matrixes are the most widely used parameters in calculation of protein mutability. Here, we developed a model to predict the strains that may appraise in the future. The swine flu H1N1, caused many deaths,

is used as a case study. We generated the mutability landscape for neuraminidase protein in swine flu according to our mutability probabilities. Thus, we addressed the location of conserved and non-conserved residues in neuraminidase. With using amino asid frequencies, mutability landscape and mutation rate of neuraminidase, we have forecast the sequences. This prediction model may lead to obtain more accurate prediction in the future and allow us to design novel drugs in advance.

**Keywords:** Protein sequence analysis, Scoring function, Mutability landscape, Neuraminidase

## TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Dr. Öğr. Üyesi Ersin Emre Ören'e, yüksek lisans eğitimim boyunca sağladıkları burs imkanları için TOBB ETÜ'ye, kıymetli tecrübelerinden faydalandığım TOBB ETÜ Biyomedikal Mühendisliği Bölümü öğretim üyelerine, hep yanımda olan Mervenaz Şahin'e ve Pınar Alpaslan'a, hiçbir zaman yardımını esirgemeyen Büşra Demir'e, bu projeye başlarken birlikte çalıştığım Gizem Gökçe'ye ve değerli tüm BNT grup arkadaşlarıma ve destekleriyle her zaman yanımda olan aileme çok teşekkür ederim.

## İÇİNDEKİLER

**Sayfa**

## ŞEKİL LİSTESİ

**Sayfa**

# ÇİZELGE LİSTESİ

**Sayfa**

## KISALTMALAR

**BLOSUM** : Blok Değiştirme Matrisi (Block Substitution Matrix)
**CDC** : Hastalık Kontrol ve Önleme Merkezleri (Centers for Disease Control and Prevention)
**DNA** : Deoksiribonükleik asit
**ED** : Eşik Değeri
**FDA** : Gıda ve İlaç İdaresi (Food and Drug Administration)
**HA** : Hemaglutinin (Hemagglutinin)
**HI** : Hemaglutanin İnhibisyonu
**HIV/ AIDS** : İnsan Bağışıklık Yetmezliği Virüsü (Human Immunodeficiency Virus)
**$IC_{50}$** : % 50 İnhibitör Konsantrasyonu
**IRD/ Fludb** : Influenza Araştırma Veri Bankası (Influenza Research Database)
**IV** : İntravenöz
**mRNA** : Mesajcı Ribonükleik asit
**MSA** : Çoklu Sekans Hizalama (Multiple Sequence Alignment)
**NA** : Nöraminidaz (Neuraminidase)
**NAI** : Nöraminidaz İnhibitörü
**NEP** : Nükleer Çıkarma Proteini (Nuclear Export Protein)
**NMR** : Nükleer Manyetik Rezonans
**NP** : Nükleoprotein
**NS1** : Yapısal Olmayan Protein 1 (Non-Structural Protein 1 )
**NS2** : Yapısal Olmayan Protein 2 (Non-Structural Protein 2)
**PA** : Polimeraz Asidik Proteini (Polymerase Acidic Protein)
**PAM** : Noktasal Kabul Edilen Mutasyon (Point Accepted Mutation)
**PDB** : Protein Data Bankası
**PIR** : Protein Bilgi Kaynağı (Protein Information Resource)
**RMSD** : Kare Ortalamanın Karekökündeki Sapma (Root Mean Square Deviation)
**RNA** : Ribonükleik asit
**vRNA** : Viral RNA
**vRNP** : Viral Ribonükleoprotein
**XRD** : X Işını Difraktometresi
**SIFT** : Toleranslıdan Toleranssızların Sınıflandırılması (Sorting Intolerant From Tolerant)
**SNAP** : Kabul Edilmeyen Polimorfizmlerin Taranması (Screening for Non-Acceptable Polymorphisms)
**SNP** : Tek Nükleotit Poliformizmi (Single Nucleotide Polymorphism)
**kryo-EM** : kriyo Elektron Mikroskobu
**UV** : Ultraviyole
**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 |
| $C\,\mathrm{var}$ | Kovaryasyon |
| $d$ | Mesafe |
| $DA$ | Değer aralığı |
| $f_a^{norm}$ | 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 |
| $PDS$ | Mesafe matrisi |
| $PSS$ | Benzerlik matrisi |
| $Q$ | Doğal frekans |
| $s$ | Amino asit |
| $S$ | Göreceli olasılık oranı |
| $stdev$ | Standart sapma |
| $TH$ | Ortalama histogram değeri |
| $TSS$ | Toplam benzerlik skoru |
| $w$ | Ağırlık faktörü (Katsayı) |

# 1. GİRİŞ

Virüsler canlı hücrelerde çoğalabilen ve enfeksiyona neden olabilen ajanlardır. Salgınlara ve ölümlere sebebiyet veren virüslere karşı önlem alınması için birçok aşılama ve ilaç tedavi yöntemleri geliştirilmiştir. Virüsler de, diğer tüm canlılar gibi evrimsel süreçte değişimlere/mutasyonlara uğrayarak yapısal ve fonksiyonel olarak farklılıklar göstermektedirler. Bu mutasyonlar, virüsün yaşam döngüsü için zararlı, nötr veya faydalı olabilmektedir. Virüs için faydalı olan bazı mutasyonlar, virüse karşı geliştirilmiş ilaçların etki ettikleri bölgelerin yapısal değişimine neden olarak ilaçların etki mekanizmalarını azaltabilmekte, hatta tamamen yok edebilmektedir. Bu da virüslerin antiviral ilaçlara karşı direnç kazanmasına sebep olmaktadır. Var olan antiviral ilaçlara direnç kazanan bu virüsler ile mücadele edilebilmesi için yeni ilaçlara ve tedavi yöntemlerine ihtiyaç duyulmaktadır. Yeni ilaçların geliştirilebilmesi için ise, virüsün ilaç hedef bölgelerinin moleküler yapısının bilinmesi gerekmektedir. İlaç geliştirme ve üretim süreçlerinin çok uzun ve pahalı olması ise virüslere karşı genelde reaktif tepki vermemize neden olmaktadır. Reaktif tepkiden proaktif tepkiye geçebilmemiz virüslerin evrimsel değişim mekanizmalarının anlaşılabilmesine bağlıdır. Bu mekanizmaların anlaşılması ile gelecekte ne tür mutasyonlar ile karşılaşılabileceği, bu mutasyonların ne tür yapısal değişimlere neden olabileceği ve bu değişimlerin var olan ilaçlara karşı virüslere bir üstünlük sağlayıp sağlamayacağı önceden belirlenebilecektir. Antiviral dirence sahip virüsleri daha ortaya çıkmadan tahmin edebilmek ise proaktif bir şekilde yeni ilaç ve tedavi yöntemleri geliştirilebilmesi için çok değerli olan zamanı kazanmamızı sağlayacaktır.

## 1.1 Tezin Amacı

Virüslere ait genetik materyaller (DNA ya da RNA) ve taşıdıkları protein bilgileri gelişen sekanslama (Sanger Yöntemi, Yüksek Verimli Sıralama (HTS)) ve yapı analizi (NMR, XRD ve kriyo elektron mikroskopisi) teknikleri sayesinde elde edilebilmekte ve bu veriler biyolojik veri bankalarında depolanarak araştırmacıların erişimine

sunulmaktadır. Bu tez kapsamında, H1N1 influenza A virüsünün bir yüzey proteini olan Nöraminidaz (NA) proteinine ait veri bankalarında bulunan protein sekans bilgileri kullanılarak öncelikle aminoasitlerin mutasyona uğrama olasılıklarının hesaplanması ve daha sonra da gelecekte ne tür mutasyonlar ile karşılaşılacağının tahmin edilmesi amaçlanmıştır. Bunun için pozisyona bağlı skorlama fonksiyonları geliştirilerek mevcut metotlarla karşılaştırılmış ve aralarındaki ilişkiler çıkarılmış, daha sonra da bu fonksiyon sonuçları ve deneysel olarak hesaplanmış mutasyon hızları kullanılarak ileride oluşabilecek evrimsel değişimler sonucu karşılaşabileceğimiz NA proteinleri tahmin edilmiştir.

## 1.2 Literatür Araştırması

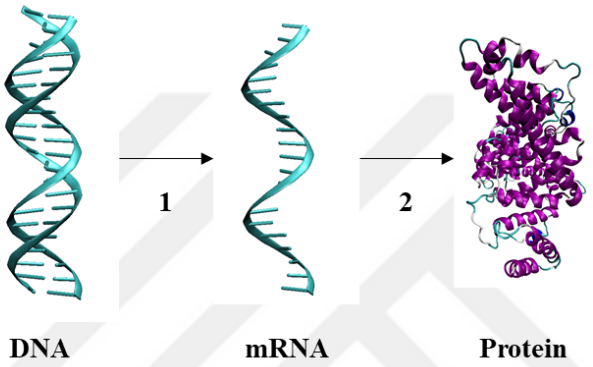
### 1.2.1 Biyoenformatik çalışma alanı – Biyolojik veri inceleme

Bilgi teknolojilerinin hızlı gelişimi ile birlikte birçok farklı alanda multidisipliner araştırmalar ve çalışmalar yapılmaktadır. Biyoenformatik alanı da bilgisayar bilimi, istatistik ve matematik ana bilim dallarından yararlanarak biyolojik verilerin incelemesini ve araştırmasını yapan bilim dalıdır (Keith, 2008). Biyolojik alandaki çalışmaların artması ile deneyler sonucu elde edilen verilerin boyutu üstel olarak artmaktadır. Özellikle genomiks ve proteomiks alanında yapılan çalışmaların incelenmesi için bilgisayar tabanlı metotlara ve algoritmalara ihtiyaç vardır (Mathura & Kangueane, 2009).

Genomiks, organizmaların tüm genom bilgilerinin incelendiği; proteomiks ise proteinlerin geniş çaplı incelendiği bir çalışma alanıdır. Bir proteom, bir organizmada veya biyolojik sistemde üretilen bir dizi proteindir. 1970-1980'lerde Fred Sanger'in grubu dizileme, genom haritalama, veri depolama ve biyoenformatik analiz tekniklerini geliştirmişlerdir. Bu çalışmalar, 1990'lardaki insan genom projesinin yolunu açmıştır. 2003 yılında tüm insan genom dizisinin yayınlanmasıyla birlikte yeni nesil sekans teknolojilerinin gelişmesine olanak sağlamıştır. Dahası, biyoenformatik alanındaki gelişmeler, yüzlerce yaşam bilimi veri tabanını ve bilimsel araştırmaya destek sağlayan projeleri hayata geçirmiştir (Url-1). Bu veri tabanlarında saklanan ve organize edilen bilgiler yardımıyla hastalıklara karşı kişisel tedavi ve gelişmiş ilaç hedeflerinin keşfi gibi önemli konular araştırılabilmekte, var olan başka sistemlerle karşılaştırılabilmekte ve analiz edilebilmektedir.

### 1.2.2 Genetik kod (DNA ve RNA)

Deoksiribonükleik asit (DNA), ribonükleik asit (RNA), protein, karbonhidrat ve lipitler hücrenin temel yapı taşlarını oluşturur. Hücresel seviyede birçok prosesin gerçekleşmesinde rol alırlar. DNA ve bazı virüslerde RNA hücrenin genetik bilgisini taşır. DNA ve RNA'lar nükleotit olarak adlandırılan blok yapılardan oluşur. Belli grup nükleik asitler genleri ifade eder. Genetik bilgi, merkezi dogma adı verilen bir süreçle proteinlere aktarılır. İlk olarak transkripsiyon evresiyle DNA, mesajcı RNA'yı (mRNA) oluşturur. Daha sonra mRNA'dan amino asit ve protein sentezi gerçekleşir, bu evre translasyon olarak adlandırılır (Şekil 1.1).




Şekil 1.1: Merkezi dogma. 1: Transkripsiyon, 2: Translasyon.

Protein sentezi prosesinde, mRNA'nın kodon adı verilen nükleotit üçlüleri, proteinlerin polipeptit zincirini oluşturan 20-sembol amino asit koduna dönüştürür. Bir amino asit birden fazla kodon tarafından sentezlenebilir. Örneğin, Glutamik asit amino asidi, GAA ya da GAG kodonları tarafından oluşturulabilir. Bu süreçte (merkezi dogma), çevresel faktörlerin etkisiyle hücre mutasyonlara ya da modifikasyonlara uğrayabilir. Bu durumda hücre çevresine ya adapte olur ya da doğal seleksiyon ile elenir. Kısacası organizmalar bu şekilde evrimleşirler (Mathura & Kangueane, 2009).

### 1.2.3 Protein yapı ve fonksiyonu

Proteinler, hücre içerisinde en çok çeşitliliğe sahip makro moleküllerdir. Bir hücrenin çalışma mekanizmasının anlaşılması için proteinlerin gerçekleştirdikleri fonksiyonların anlaşılması gerekir. Kataliz, taşıma, molekül depolama, mekanik destek, hücre bölünmesi ve bağışıklık gibi önemli hücre olaylarını gerçekleştirirler (Berg, Tymoczko, & Stryer, 2002).

Protein yapıları üç deneysel teknik ile elde edilir: X-ışını kristalografisi, nükleer manyetik rezonans (NMR) ve kriyo elektron mikroskobu (kryo-EM). X-ışını kristalografisinde, X ışını protein yapısına göre farklı yönlerde kırınıma uğrar ve bu şekilde protein molekülünün kristal yapısı çıkarılır. NMR tekniği, çözelti içindeki molekülün manyetik alan içerisinde titreşim hareketlerini ölçmeye dayanır. Kriyo elektron mikroskobu ise diğer yöntemlere göre daha detaylı sonuçlar verir. Küçük dalga boyuna sahip elektronlar ile görüntüleme yapan bu mikroskop kompleks biyolojik yapıları sıvı içerisinde görüntüleyebilmektedir.

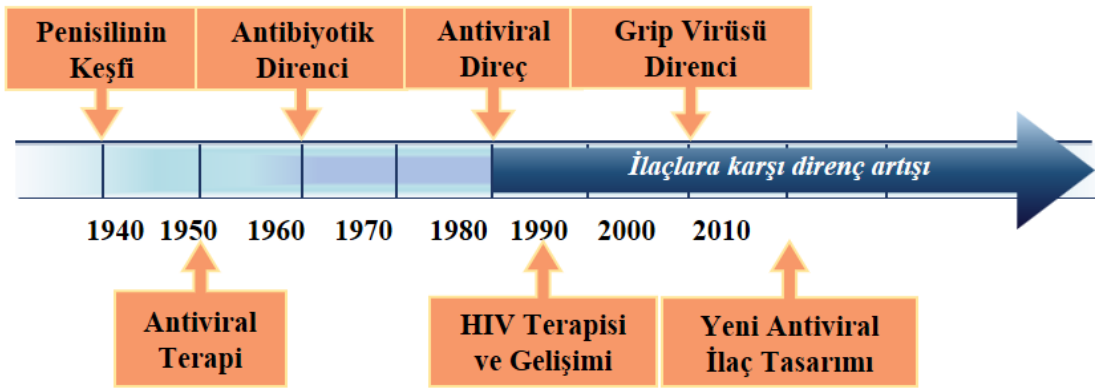
Primer protein yapısı, aminoasit dizilimi ya da protein sekansı olarak tanımlanır. Bir protein sekansı en fazla 20 farklı aminoasit içerebilir. Protein sekans uzunluğu proteinlere göre farklılık göstermektedir. Örneğin; aktif proteinlerin uzunluğu 50 amino asit uzunluğundan fazla olmaktadır. Primer protein yapısı üç boyutlu protein yapısı hakkında tek başına bir bilgi verememektedir. Protein katlanmaları, üç boyutlu protein yapıları hakkında bilgi veren bir yaklaşımdır. Bir proteinin primer sekansında gerçekleşen değişiklikler proteinin üç boyutlu yapısını etkilemeyebilir çünkü aminoasitler fizikokimyasal özelliklerine göre gruplara ayrılırlar ve birbirlerinin yerine geçmeleri, bazı durumlarda üç boyutlu yapı üzerinde bir değişikliğe sebep olmaz. Ancak protein yapısı bir değişikliğe uğrarsa, protein gerçekleştirmesi gereken fonksiyonu yerine getiremeyebilir. Böyle durumlarda hücre içerisinde protein agregasyonları ya da yanlış protein katlanmaları görülebilir. Bunlara örnek olarak Alzheimer hastalığı, Creutzfeldt-Jakob hastalığı, Huntington hastalığı, Tip II diyabet ve Parkinson hastalığı vs. verilebilir. Bu nedenle protein yapı ve fonksiyonu arasındaki ilişkiyi anlamak çok büyük bir önem taşımaktadır. Yukarıda verilen hastalıklara çözüm bulabilmek için bilgisayar tabanlı çalışmalara ihtiyaç vardır. Genomik dizi analizine benzer olarak, protein yapılarının biyoenformatik çalışmaları da, proteinlerin katlanması, evrimi ve işlevi, protein-ligand ve protein-protein etkileşimlerinin doğası ve mekanizmaları gibi konuların anlaşılması için bir yön göstermektedir. Bu tür çalışmaların başarısı sadece bilimde değil tüm toplumda hastalıkların moleküler seviyedeki etkileşimleri hakkında bilgi sağlaması, yeni, etkili terapötik ajanlar ve tedavi rejimleri geliştirilmesi adına çok büyük bir öneme sahiptir (Y. Xu, Xu, & Liang, 2007).

### 1.2.4 Antibiyotik ve antiviral ilaçların keşfi ve tasarımı

Mikroorganizmalar konak hücreyi enfekte ederek hastalıklara, hatta ölümlere sebep olurlar. Bulaşıcı hastalıklardan sıtma ve tüberküloz, insanlık tarihinde tüm savaşlardan daha fazla ölüme sebep olmuştur. 50-100 milyon insan 1918 influenza (ispanyol gribi) salgını yüzünden hayatını kaybetmiştir (Knobler, Mack, Mahmoud, & Lemon, 2005);(Klebe, 2013).

Tedavi amaçlı olarak 1940'lı yıllarda Alexander Fleming tarafından keşfedilen penisilin ilk antibiyotik olarak kullanılmıştır (Davies & Davies, 2010). 1960'lı yıllarda ise herpes simpleks (uçuk hastalığı) virüsüne karşı üretilen idoxuridine ilacının kullanımı ile antiviral terapi başlamıştır (de Clercq, 2012).

İlaçlar hastaya ulaşmadan önce etkin ve güvenilir olup olmadıkları denetlenmelidir. Bu denetimler belli kurumlarca yapılmaktadır. Amerika'da ABD Gıda ve İlaç İdaresi (The U.S. Food and Drug Administration FDA) (Url-4), ülkemizde ise Türkiye İlaç ve Tıbbi Cihaz Kurumu (Url-5) tarafından ilaçlar denetlenmektedir. 1940 ile 1960'ların ortalarına kadar yeni antibiyotikler ve diğer antibakteriyel ajanlar keşfinde hızlı ilerlemeler kaydedilmiştir. Bununla birlikte, antibiyotik çağının en erken döneminden bile, ilaca dirençli bakteriler ortaya çıkmaya başlamıştır. Antiviral ilaçlara karşı direnç ise 1980'lerde rapor edilmeye başlamıştır (Bolken & Hruby, 2008). Antibiyotiklerin ve antiviral ilaçların keşfi ve tarihi gelişim akışı Şekil 1.2'de görülmektedir. Bir ilaç, ilaç tasarımı ve gelişimi, klinik öncesi çalışmalar, klinik çalışmalar ve tedavi onayı gibi safhalardan geçerek kullanıcıya ulaşmaktadır. Bir ilaç üretim süreci yaklaşık 10 yılı bulmaktadır (Url-6). Bu nedenle antibiyotik ve antiviral ilaçların etkisiz hale gelmesiyle, direnç mekanizmalarının anlaşılması insan sağlığı için büyük bir önem taşımaktadır.




Şekil 1.2: Antibiyotiklerin ve antiviral ilaçların keşfi ve tarihi gelişim akışı.

### 1.2.5 Virüs evrimi

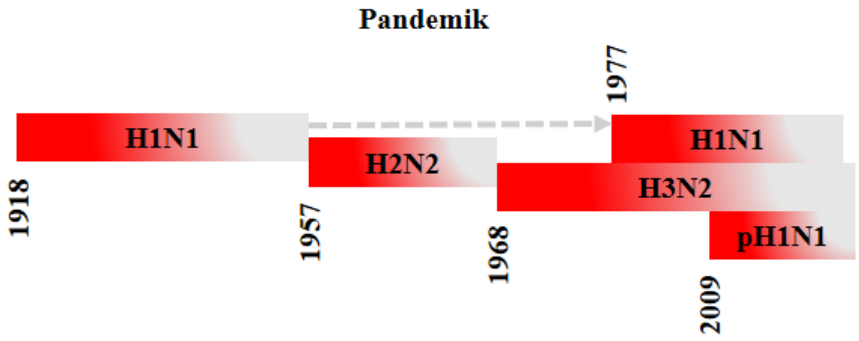
1970'lerden beri gen sıralama ve protein yapılarının belirlenmesi için geliştirilen yöntemler, evrim çalışmaları özellikle de virüs evriminin incelenmesi için gerekli metotların geliştirilebilmesini sağlamıştır. Virüslerin evrimi, günlük hayatımızın ve tüm diğer canlı organizmaların yaşamları için önemli sonuçlara sahiptir. İnfluenza, uçuk, AIDS, hemorajik ateş ve diğer birçok viral hastalıkları anlamak ve kontrol etmek, özellikle moleküler düzeyde viral evrimi anlamamıza bağlıdır (Gibbs, Calisher, & Garcia-Arenal, 1996).

Viral gen sekansları, virüsler arasındaki yakınlık ilişkisi hakkında bilgi verir. Virüslerin taşıdıkları gen materyali DNA ya da RNA olabilir ve viral genom üzerinde mutasyonlar rastsal şekilde gerçekleşir. Mutasyon hızı, viral evrimi anlamak için kritik bir parametredir. Çift iplikli DNA'ya sahip virüsler bir bp (base pair) üretimi başına $10^{-6}$-$10^{-8}$ mutasyon hızına sahipken, RNA taşıyan virüslerde bu değer $10^{-4}$-$10^{-6}$'dır. Bunun en temel sebeplerinden biri RNA virüsleri replike olurken RNA polimeraz enzimi düzeltme okuması (proofreading) özelliğine sahip değildir (Lauring, Frydman, & Andino, 2013). Sonuç olarak, RNA virüs popülasyonları son derece yüksek genetik değişkenliğe sahiptir. Bu mutasyonların bir kısmı virüsün kendisi için zararlı etkilere sahip olabilir ve virüsün elenmesine sebep olabilir. Bu nedenle, bu zararlı mutasyonlara sahip virüsler zamanla ortadan kalkar ve bu mutasyonlar, kabul edilmeyen mutasyonlar olarak adlandırılır. Kabul edilen mutasyonlar ise virüsün kendisi için faydalı olanları ifade etmektedir. Başka bir deyişle, virüsler kabul görmüş mutasyonlarla çevrelerine daha dayanıklı ve dirençli olabilecek bazı özellikler kazanabilir. Genel olarak bu virüsler iki şekilde değişime uğrarlar. Genetik sürüklenme (genetic drift), virüs çoğaldıkça zaman içinde sürekli olarak meydana gelen virüslerin genlerindeki küçük değişikliklerdir. Bu küçük genetik değişiklikler genellikle birbiriyle yakından ilişkili olan virüsler üretir. Genetik kayma (genetic shift), virüslerde ani, büyük bir değişiklik olup, virüsün yeni bir alt tipine sahip bir virüsün veya yeni bir virüsün oluşmasıyla sonuçlanır (Petrova & Russell, 2018). Virüslerin mutasyona uğrama kapasitesindeki görece yükseklik, değişen ortamlara hızla adapte olmalarına ve böylece, virüslerin ilaç direnci oluşturmasına neden olur. Konak aralığı, bulaşma veya patojenite gibi parametreler değişikliğe uğrar (Plant & Ye, 2013). Bu durum RNA virüsleri için daha hızlı bir şekilde gerçekleşir. Bu

mutasyonların bir sonucu olarak halihazırda geliştirilmiş olan antiviral ilaçlar hastalıkların tedavisi için yetersiz kalmaktadır (Combe & Sanjuán, 2014).

### 1.2.6 İnfluenza (Grip) virüsü

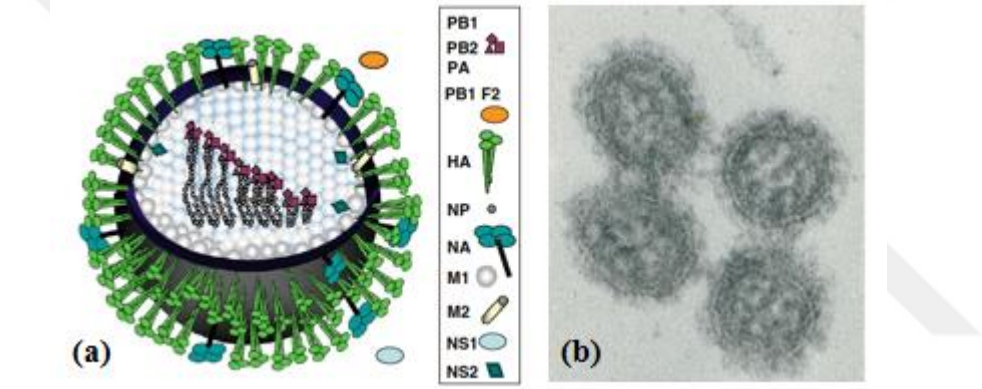
Grip virüsü olarak bildiğimiz influenza virüsü yüksek bulaşıcılık ve ölüm oranlarıyla sonuçlanabilecek salgınlara neden olmaktadır. Hastalık Kontrol ve Önleme Merkezleri'ne (Centers for Disease Control and Prevention (CDC)) göre her yıl dünya çapında 291,000-646,000 kişi gripten dolayı hayatını kaybetmektedir (Url-8). İnfluenza virüsü, 1918, 1957, 1968 ve 2009 yıllarında pandemilere sebep olarak dünya çapında büyük kayıplara sebep olmuştur. Şekil 1.3'de pandemik süreçler görülmektedir. 1918-1919 arasında görülen pandemik, influenza A/H1N1 virüsünden kaynaklanmış ve ABD'de 500.000'den fazla kişi hayatını kaybetmiştir ve dünya çapında bu kayıp tahmini 50-100 milyondur. 1957-1958 yıllarında Asya grip salgını sırasında, influenza A/H2N2 virüsü, ABD'de yaklaşık 70.000 ölüme ve dünya çapında yaklaşık 2 milyon kişinin ölümüne sebep olmuştur. 1968-1969 yıllarında, influenza A/H3N2 virüsünün neden olduğu Hong Kong grip salgını, ABD'de yaklaşık 30.000 kişinin ölümüyle, 1 milyon dünya çapında ölümle, en yakın zamanda görülen pandemik, 2009'da, H1N1 virüsü ABD'de yaklaşık 12.000 kişinin ölümüyle, dünyada yaklaşık 280.000 ölümle sonuçlandığı tahmin edilmektedir. Şu anda bir sonraki influenza pandemik olayının zamanlamasını veya buna sebep olan virüsü tahmin etmek büyük bir önem taşımaktadır. Yeni bir pandemiyle karşılaşma durumuna karşı çalışmalar ve yatırımlar, uluslararası düzeyde zamanla artmıştır (Schuchat, Tappero, & Blandford, 2014);(Compans & Oldstone, 2014).




Şekil 1.3: Pandemik influenza virüsü zaman çizelgesi (Compans & Oldstone, 2014).

### 1.2.6.1 İnfluenza virüsünün çeşitleri, yapısı ve evrimi

İnfluenza virüsleri 1931'de domuzlarda ve kısa bir süre sonra insanda keşfedilmiştir. İnsanda rastlanan influenza virüslerinin hayvan virüsü rezervuarından ortaya çıktığı kabul edilmektedir. İnfluenza virüsleri Orthomyxoviridae ailesinin üyeleridir. Nükleokapsid ve matriks proteinlerinin antijenik farklılıklarına dayanarak, üç farklı tip ayırt edilmiştir: influenza A, B ve C virüsleri. İnfluenza A virüsleri ayrıca 16 farklı hemaglutinin (H1-H16) ve dokuz farklı nöraminidaz (N1-N9) ile karakterize edilen alt tiplere ayrılır. İnfluenza virüslerinin çapları yaklaşık olarak 100 nm'dir. Viral glikoproteinler, konaktan türetilen lipit zarfına gömülür ve partiküller elektron mikroskobu altında görüntülendiğinde virüsün dış yüzeyinden yayılan sivri uçlar olarak görünür. (Şekil 1.4)

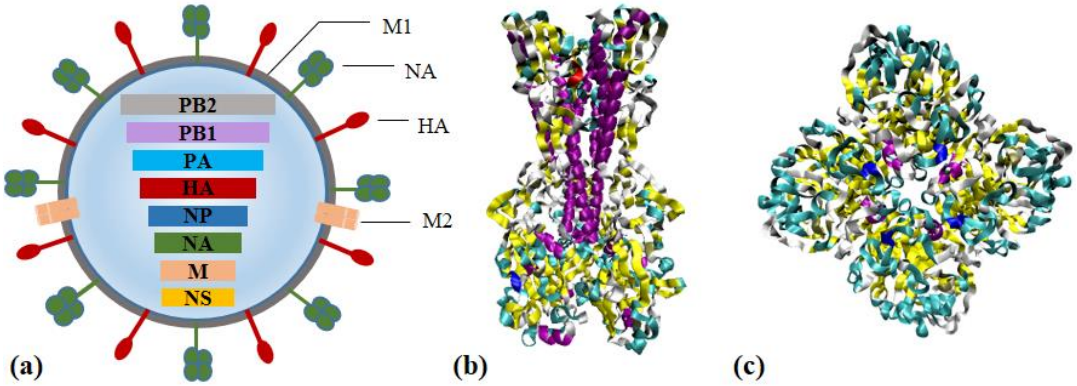



Şekil 1.4: İnfluenza A virüsünün yapısı. (a) Virüs modeli: Yüzey proteinleri ve Ribonükleoproteinleri. (b) Elektron mikrografı, virüs partiküllerinin ince kesitlerini göstermektedir. Virionların çapı 100 nm'dir (von Itzstein 2012).

İnfluenza A ve B virüslerinin genomları, vRNA olarak bilinen sekiz ayrı negatif, tek iplikli RNA segmentinden oluşur. İnfluenza C virüs genomu, yedi segmente sahiptir. İnfluenza A, B ve C virüslerinin genomları, sırasıyla toplam 13.600, 14.600 ve 12.900 nükleotit uzunluğuna sahiptir. Bu segmentasyonlar virüslerin genetik varyasyona/değişkenliğe sahip olmasına neden olmaktadır. Her bölüm bir veya iki viral proteini kodlar. İnfluenza A virüsleri için ayırımı şu şekildedir: Polimeraz 2 Proteini için RNA segmenti 1, Polimeraz 1 Proteini için segment 2 ve bazı suşlarda Polimeraz 1 Proteini-F2, PA için segment 3, HA için segment 4, Nükleoprotein (NP) için segment 5, NA için bölüm 6, M1 ve M2 için segment 7 ve NS1 ve NS2 / Nükleer Çıkarma Proteini (Nuclear export protein) (NEP) için segment 8 (von Itzstein, 2012).

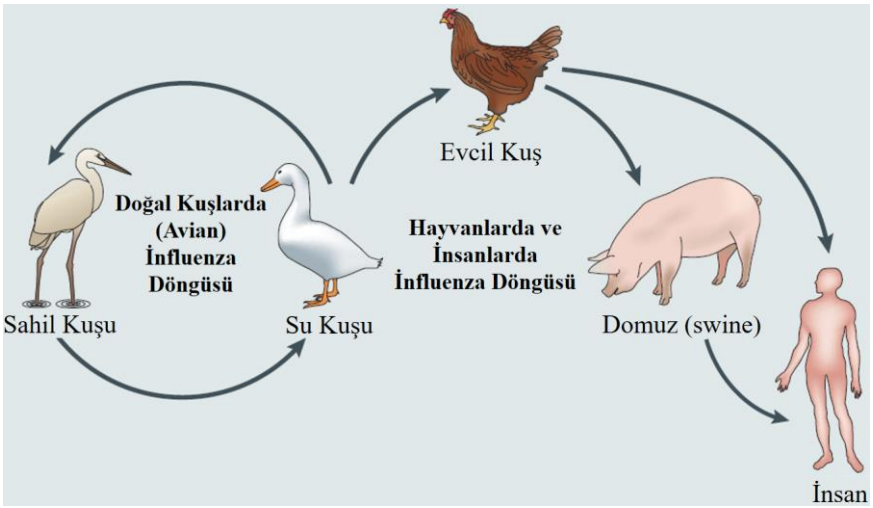
İnfluenza virüsünün yayılmaya sebep olabilecek proteinleri üzerinde birçok çalışma yapılmaktadır. Özellikle yüzey proteinlerinden hemaglutinin ve nöraminidaz en çok çalışılan proteinlerdir. Hemaglutinin (HA), virüs enfeksiyonu için gerekli olan homotrimerik bir integral membran glikoproteinidir. HA, radyal doğrultuda 35-70 Å arasında değişen, silindirik biçimli 135 Å uzunluğundadır (Isin, Doruker, & Bahar, 2002). Üç boyutlu yapısı Şekil 1.5 (b)'de görülmektedir. Virüs enfeksiyonunun ilk aşamalarını reseptör bağlanma ve membran füzyonu oluşturur. Konak hücre yüzeyindeki glikoproteinlerin ve glikolipidlerin sialik asitlerini tanıyarak o bölgelere bağlanırlar (Gamblin vd., 2004). HA homotrimer yapısı uzun, genişletilmiş bir kök bölgesi ve reseptör bağlanma alanı ve körelme esteraz bölgesini içeren bir küresel baş yapısına sahiptir. Kök bölgesi membran füzyon mekanizmasından, küresel baş kısmı ise reseptöre bağlanma mekanizmasından sorumludur (R. Xu vd., 2010).

Virüs enfeksiyonunda diğer konak hücrelere dağılımı sağlayan yüzey proteini NA'dır. NA proteini sialik asit ile komşu şeker kalıntısının arasında α-ketosidik bağını koparan bir ekzosiyalidaz enzimidir. İnfluenza A NA'nın dokuz alt tipi iki filogenetik gruba ayrılır. Birinci grup N1, N4, N5 ve N8 alt tiplerinden, ikincisi ise N2, N3, N6, N7 ve N9 alt tiplerinden oluşmaktadır. İnfluenza virüsü NA'nın polipeptit zinciri 470 amino asit içerir. NA'nın üç boyutlu yapısı birkaç alandan oluşur: sitoplazmik (6 amino asit), transmembran (7-29 amino asit), "baş" (19 amino asit enzimatik aktivite sağlar) ve ayrıca "gövde" (~50 amino asit), baş bölgesi transmembran alanına gövde bölgesiyle bağlanır. NA proteini influenza virüsünün yüzeyinde mantar şekline benzer homotetramerdir (Şekil 1.5 (c)). Baş kısmı 80x80x40 Å boyutunda gövde ise 15 Å genişliğinde 60 ile 100 Å uzunluğundadır. Bir monomerin ağırlığı ~60 kDa'dır. Bir virüs partikülü yaklaşık 50 tetramer içermektedir. (Shtyrya, Mochalova, & Bovin, 2009), (Jagadesh, Salam, Mudgal, & Arunkumar, 2016). NA'nın işlevi, hücre yüzeylerinde ve yeni oluşan virüslerin üzerinde mevcut olan terminal sialik asit moleküllerini parçalamak ve virüsün enfekte olmuş hücrelerden salımını kolaylaştırmaktır. NA, yeni oluşan virüslerin salınmasında ve yayılmasında önemli bir rol oynamaktır (Jagadesh vd., 2016).




Şekil 1.5: (a) İnfluenza virüsünün yapısı ve içerdiği proteinler, (b) Hemaglutanin proteininin yapısı (PDB: 1RUZ), (c) Nöraminidaz proteininin yapısı (PDB: 2HU4).

İnfluenza virüslerinin yüksek genetik değişkenliği, bu ajanların karmaşık ekolojisi ve epidemiyolojisi için en önemli belirleyicidir. Özellikle influenza A virüsleri çok sayıda HA ve NA alt tipine sahiptir. Bu alt tiplerin sadece bir kısmı insan, domuz, at ve diğer memelilerde gözlemlenirken, tüm tipler kuşlarda görülür (Şekil 1.6). Su kuşlarından köken alan virüsler, tavuk ve bıldırcınlar tarafından ara konakçı olarak replike edildiğinde, insan reseptörlerine bağlanma yatkınlıklarının fazla olduğu gözlenmiştir. Bu virüsler daha sonra insanları enfekte edebilir ve hem yüzey hem de iç proteinleri kodlayan genlerdeki mutasyonlarla daha fazla adapte olabilir. Bu mutasyonlarla, yeni yüzey glikoproteinleri olan virüslerin antijenik özelliklerinde belirgin bir değişiklik meydana gelebilir. Böyle bir antijenik kaymaya sahip yeni bir virüs insanda ortaya çıkarsa, bir pandemiye neden olabilir (von Itzstein, 2012).




Şekil 1.6: İnfluenza A virüsünün türler arası aktarımı (Shi, Wu, Zhang, Qi, & Gao, 2014).

İnfluenza A ve B virüsleri yıllık salgınlara neden olur ve influenza A virüsü belli aralıklarla pandemiye neden olmuştur (Bakınız Bölüm 1.2.6). İnfluenza C virüsleri ise insanları enfekte eder ancak hafif, semptomsuz enfeksiyonlara neden olmaktadır. Bazı alt tip virüsler (H5N1, H7N7, H9N2) yaygın salgınlara neden olmamıştır (Kawaoka ve Neumann 2012).

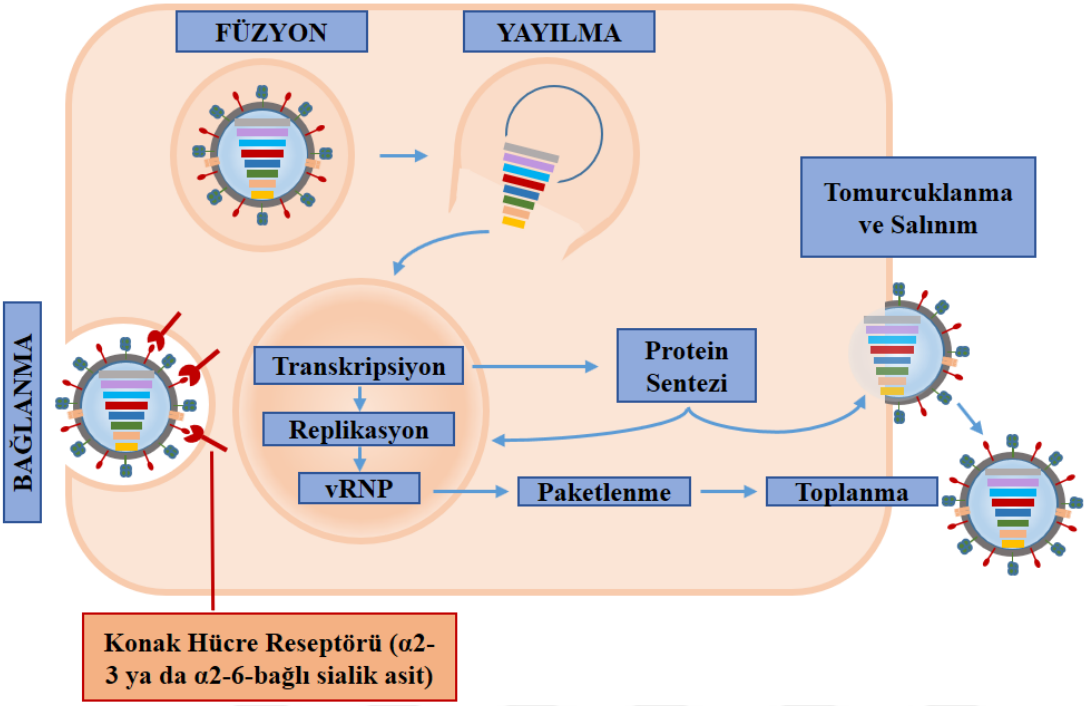
İnfluenza virüsü evriminin anlaşılma zorluğunun üstesinden gelmek için, İnfluenza Genom Dizileme Projesi gibi geniş çaplı bir işbirliği oluşturulmuştur. İnsan popülasyonunda dolaşan influenza tipleri (clade) ortaya çıkarılmıştır. Daha fazla sayıda influenza virüsü genom sekansı kullanılabilir hale geldiğinde ve antijenik değişim döngüleri hakkında bilgi sahibi olunduğunda, bu bilgi yıllık aşı gelişimine uygulanabilir. Gelecek influenza sekanslarının tahmin edilmesi, etkili aşıların ve tedavi yöntemlerinin zamanında gelişmesine olanak sağlar (Rappuoli & Giudice, 2011).

Dünya çapında toplanan örneklerden elde edilen mevcut influenza genom dizilerinin artan sayısı, özellikle virüs evrimi, popülasyon bağışıklığı ve virüs etkisi arasındaki etkileşim bilgisi, influenzanın global epidemiyolojisi hakkında daha tutarlı sonuçlar elde edilmesini sağlamaktadır (Rappuoli & Giudice, 2011).

#### 1.2.6.2 İnfluenza virüsünün yayılımı

Viral yaşam döngüsü, virüslerin hücre yüzeyindeki reseptörlere bağlanması ile başlar. Bağlanma, konak hücre yüzeyinde bulunan proteinlere ve lipitlere bağlı sialil-oligosakkaritler ile HA proteininin arasındaki etkileşim ile olur. Virüs, reseptör-aracı endositoz ile hücre içerisine girer. Geç endozomlardaki düşük pH, HA'da bir konformasyon değişikliğini tetikler (Kawaoka & Neumann, 2012). Ayrıca M2 iyon kanalı içeren İnfluenza A virüslerinde, M2 iyon kanalları virüs partikülü içindeki pH'ı düşürerek moleküllerin ayrışmasına yardımcı olur (von Itzstein, 2012). Böylece viral ve endozomal membranların füzyonu ve viral ribonükleoprotein (vRNP) komplekslerinin (vRNA ve polimeraz ve NP proteinlerinden oluşan) sitoplazmaya salınımı gerçekleşir. Çekirdek içerisinde gerçekleşen replikasyon ve transkripsiyon basamakları, vRNA'ların amplifikasyonu ve viral protein sentezi için mRNA'ların sentezine yol açar. Enfeksiyon döngüsünün sonlarında, yeni oluşan vRNP'ler, M1 ve NEP proteinlerinin yardımıyla sitoplazmaya verilir. Bu vRNP'ler plazma zarında yeni sentezlenmiş HA, NA ve iyon kanalı proteini M2 ile bir araya getirilir. NA

karbonhidratla bağlı sialik asidi ayırarak, yeni oluşan virüsün salınmasını kolaylaştırır. İnfluenza virionları tomurcuklanarak salınır (Compans & Oldstone, 2014). Şekil 1.7'de influenza A virüsünün yaşam döngüsü basamakları görülmektedir.




Şekil 1.7: İnfluenza A virüslerinin yaşam döngüsü (Shi vd., 2014).

### 1.2.6.3 İnfluenza virüsü için ilaç geliştirilmesi

İnfluenza virüsü yüksek hastalık ve ölüm oranlarında dünyadaki tüm yaş gruplarını etkilemektedir. İnfluenza hastalığının tedavisi ve önlenmesi için, çeşitli antiviral tedaviler uygulanmaktadır. İnfluenza virüsüne karşı etkili aşılar ve aşılama stratejileri ayrıca antiviral ilaçlar geliştirilmiş ve daha etkili sonuçlar elde etmek için geliştirilmeye devam edilmektedir (Englund, 2002).

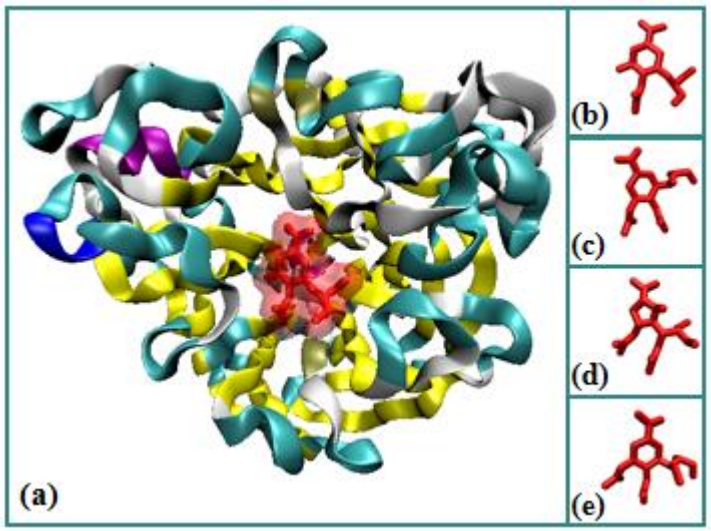
Mevsimsel suşlara (virüs alttürleri) karşı her yıl aşılar üretilmektedir. İmmünojenik proteinleri üretmek için kullanılan influenza aşısı farklı virüslerden alınan gen segmentleri içerir. İnfluenzada gen gruplaşma (gene constellation) etkisini anlamak, özellikle aşı üretimi için önemlidir. Bu birbirine benzer özellikleri taşıyan segmentlerin grup oluşturması, bir ataya ait virüslerden farklı fenotiplere sahip yeni virüslerin oluşmasına yol açar. Glikoproteinleri ve polimeraz proteinlerini kodlayan bazı segmentler arasında bu olay daha sık gerçekleşmektedir. Hücreye yönelim, hücreye yayılma ve yayılma hızı, büyüme ve patojenik etki gibi özellikler yeni oluşan

gen gruplarından dolayı değişmektedir. Gen segmentlerinde meydana gelen bu değişimleri anlamak için oluşan mutasyonların daha fazla analizi yapılmalıdır. Oluşan mutasyonların anlaşılması ve tanımlanması, viral proteinler arasındaki etkileşim ağını anlamamıza yardımcı olacaktır. Böylece, gen gruplaşma etkisinin anlaşılması, aşı üretimi için aday virüslerin seçilmesini sağlayabilir (Plant & Ye, 2013). Şuan kullanılan aşılar influenza virüsü üstünde bulunan bir yüzey proteini olan HA'yı bloke etmek için yapılmıştır. Standardize etme işlemleri NA proteini için gerçekleştirilememiştir. Bu nedenle, NA proteini antijen olarak kullanılamamaktadır. NA daha çok antiviral ilaçlar için bir hedef olarak görülmektedir. NA-özel antikorlar enfeksiyonu engellemek için değil daha çok virüs yayılımını engellemek için kullanılabilir. NA proteini HA proteinine göre daha yavaş evrimleşmekte ve değişmektedir. İlerisi için hem HA hem de NA'ya etkili kombine edilmiş aşıların etkili olacağı düşünülmektedir (Jagadesh vd., 2016).

İnfluenza virüs enfeksiyonlarının, aşıların kullanımı ile önlenmesi, grip virüsü kontrolünün en uygun maliyetli ve pratik yöntemidir, ancak bazı yüksek riskli popülasyonlarda veya bireylerde influenzaya karşı antiviral korunma ve tedavi yöntemleri sunulmaktadır (Englund 2002). Antiviral ilaçlar influenza virüsü enfeksiyonlarının kontrolünde önemli bir rol oynamaktadır. 1999'dan önce, influenza enfeksiyonlarından korunma ve influenza enfeksiyonlarının tedavisi için sadece adamantan türevi (amantadin ve rimantadin) ilaçlar kullanılmaktaydı. Adamantanlar influenza A virüslerinin M2 proteini tarafından oluşturulan proton kanalını hedefler ve influenza B virüslerine karşı etkili değildir. Adamantan dirençli virüslerin yakın zamanda ortaya çıkması ve yayılması, bu ilaç sınıfının yararlılığını büyük ölçüde azaltmıştır.

Amantadin ve Rimantadin'e karşı direncin artması ve etkili bir aşı olmaması nedeniyle, bu virüse karşı korunmak için NA inhibitörleri (NAI'lar) kullanılmaya başlanmıştır. Dünya Sağlık Örgütü oluşabilecek pandemik durumlara karşı önlem almak için üye ülkeleri NAI geliştirmeye teşvik etmiştir (Jefferson, Jones, Doshi, & Del Mar, 2009). NA inhibitörleri (Şekil 1.8), hem influenza A hem de B virüslerine karşı aktif olan bir ilaç sınıfıdır. Halen, inhale zanamivir ve oral oseltamivir, A ve B tipi influenza enfeksiyonlarına karşı kullanılan FDA onaylı NAI'lardır. Bir intravenöz (IV) formülasyon olarak geliştirilen Peramivir, ABD'de 2009 H1N1 pandemisi sırasında acil kullanım yetkisi kapsamında reçete edilmiş ve şu anda Japonya ve Güney Kore'de

lisanslanmıştır. Ayrıca, bir inhaler prodrug (soluk yoluyla alınan ön ilaç) olarak geliştirilen Laninamivir, Japonya'da lisanslıdır. NAI'lar, sialik asit (N-asetil nöraminik asit) NA'nın doğal substratını taklit eder ve korunan NA aktif bölgesine rekabetçi bir şekilde bağlanır.




Şekil 1.8: (a) NA proteininin ilaç ile etkileşimi (PDB:3TI6), (b) Oseltamivir, (c) Zanamivir, (d) Peramivir, (e) Laninamivir.

Influenza virüsleri, NAI'ların varlığında yayıldığı zaman, yeni oluşan virionlar, hücre zarına ve birbirlerine yapışır ve böylece komşu hücrelere enfeksiyonun yayılmasını sınırlar. 2007'den önce, küçük çocuklarda oseltamivir tedavisinin ardından toplanan virüs varyantlarının detaylı çalışmaları, test edilen örneklerin % 18'inde dirençli mutasyonların bulunduğunu ortaya koysa da, NAI'lara karşı sadece düşük direnç seviyeleri tespit edilmiştir. Son çalışmalar, N1 enziminin aktif bölgesinde bir diğer adıyla ortosterik bölgede, NAI'nın indüklenmesini önleyen amino asit değişikliklerinin, oseltamivir ve/veya zanamivire karşı direnç oluşturma potansiyeline sahip olduğunu göstermektedir. Bu nedenle, her yeni virüs suşunun, özellikle NA aktif bölgesi amino asit sekansı, NAI-duyarlı virüsler ile karşılaştırıldığında aynı olan fenotiplerin NAI duyarlılığına bakılmalıdır.

NA proteininin aktif bölge olarak adlandırılan bölgesinde 150-loop adında bir bölge bulunmaktadır. Bu bölge konformasyon değişikliklerine uğrayarak ya açık ya da kapalı konformasyonda bulunmaktadır. İlaç hedeflemesi olarak da şu an bu bölge üzerinde çalışmalar yapılmaktadır (Amaro vd., 2011).

Çünkü aktif bölgenin dışındaki amino asit değişimleri de, NAI'nın aktif bölgeye bağlanma eğilimini olumsuz yönde etkileyebilmektedir. (Kawaoka & Neumann, 2012). Aktif bölge dışında proteinin ilaca karşı direnç kazanmasına sebep olabilecek yeni ilaç bağlanma bölgeleri oluşabilmektedir, bu bölgeler allosterik bölge olarak adlandırılmaktadır. Protein üzerinde meydana gelen amino asit değişiklikleri dolaylı olarak proteinin fonksiyonunu etkilemektedir. Yapılan bir araştırmaya göre ortosterik ve allosterik bölgeler, diğer bölgelere göre daha çok birbirlerinden etkilenmektedirler (Ma, Meng, & Lai, 2016). Bu nedenle NA proteininin allosterik bölgeleri keşfedilirse yeni ilaç hedefleme bölgeleri olarak kullanılabilirler.

İlaç hedefleme çalışmalarında yüzey proteinlerinin birbirine olan etkisi de araştırılmaktadır. Bir deneysel çalışmaya göre hayvanlar üzerinde gözlemlenen influenza virüslerinde mutant hemaglutanin içerenlerin NAI'lara karşı daha duyarlı olduğu bulunmuştur, NA'nın enfeksiyon sürecinde reseptör yıkımı dışında hayati bir rol oynayabileceği düşünülmektedir (Garman & Laver, 2005).

### 1.2.7 Deneysel çalışmalar

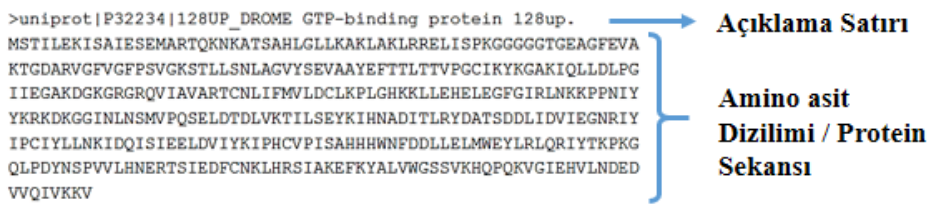
İnfluenza virüs mekanizmasının anlaşılması ve bu virüse karşı ilaç keşfi deneysel çalışmalar yardımıyla büyük bir hız kazanmıştır. Klinik olarak gözlemlenen virüsler deneysel ortamlarda incelenerek, enfeksiyon oluşturma kapasiteleri (viral fitness), çevre şartları değişimlerine karşı tepkileri ve tehlikeli mutasyon bölgeleri araştırılmaktadır (Wargo & Kurath, 2012). Kristal yapısı elde edilmiş influenza virüs yapıları X-ray, NMR gibi yöntemlerin kullanılmasıyla mevcuttur. Bunun dışında antijenik varyasyonların çoğu Hemaglutanin İnhibisyonu (HAI) analiziyle ölçülmektedir. Bunun yanında plak ve mikronötrleme tahlilleriyle de desteklenmektedir (Petrova & Russell, 2018).

#### 1.2.7.1 Protein sekans veri bankaları

Protein sekansı, amino asitlerin sırasını ve dolayısıyla polipeptit zincirinin kovalent yapısını tarif etmektedir. Protein sekansı verilerinin büyük çoğunluğu amino asit dizisi ile temsil edilmektedir. Amino asit dizisi, protein yapısı ve işlevi hakkında temel bilgiyi içerdiği için önemli bir yere sahiptir (Bakınız Bölüm 1.2.3). Protein sekans verilerinin kullanımı biyokimya, ekoloji, etimoloji, evrim, genetik, genetik mühendisliği, genomik, moleküler filogenetik ve sistematik, farmakoloji ve toksikoloji

gibi alanlarda yaygındır (Edwards, Stajich, & Hansen, 2009). Protein ve genomik sekans analizleri, hücresel sistemlerin yapısını, işlevini ve organizasyonunu anlamada yardımcı olmaktadır. Protein sekans analizi, sekans benzerliğini, fonksiyonel motifleri ve desenleri tanımlamayı içermektedir. Ortak bir atadan evrimleşmiş protein sekansları benzer yapıyı ve işlevi paylaşır. Korunmuş sekans bölgeleri sekans motifler ya da yapı motifleri olarak adlandırılır. Bir proteinin üç boyutlu yapısı biliniyorsa, bilinmeyen protein için geometrik bilgi elde etmek için karşılaştırmalı modelleme teknikleri uygulanmaktadır. Benzer ortak ataya sahip veya katlanma yapısı bilinen proteinler biyolojik bilginin etkili bir şekilde ele alınması için kritik öneme sahiptir (Mathura & Kangueane, 2009). Bu nedenle protein sekans bilgisi kullanılarak genetik haritaların çıkarılması, polimorfizm tanımlama, protein-protein etkileşimleri ve ilaç tasarımı gibi konularda çalışmalar yapılabilir.

Biyoenformatik alanında yapılan çalışmaların artmasının en büyük sebebi dünya çapında herkesin ulaşabileceği biyolojik veri bankalarının oluşturulması ve verilerin depolanmasıdır. Moleküler biyoloji tekniklerinde ilerleme ve yüksek verim (high-throughput) yöntemleri, genomik ve proteomik verilerde üstel bir artışa neden olmuştur (Mathura & Kangueane, 2009). En önemli protein veri tabanları, İsviçre Protein Veri tabanı (SWISS-PROT) (Boeckmann vd., 2003), EMBL (TrEMBL), Protein Bilgi Kaynağı (PIR), ve 3 boyutlu protein yapılarının bulunduğu Protein Data Bankası'dır (PDB) (Berman vd., 2000). Protein veri tabanlarının büyüme oranı, DNA veri tabanlarına kıyasla daha doğrusal olmuştur. Şu anda, UniProt Bilgi Bankası adı altında Avrupa Biyoenformatik Enstitüsü'nü, İsviçre Biyoenformatik Enstitüsü'nü ve Protein Bilgi Kaynağı'nı içeren yaklaşık 39 milyar amino asit, 115 milyon sekans bulunmaktadır (Url-2). Bunlar FASTA formatında (Şekil 1.9) (Pearson & Lipman, 1988) ve özel ara yüzler aracılığıyla sıkıştırılmış bir sekans dosyasında mevcuttur. Son kırk beş yılda sekans üretimindeki büyüme, protein dizisi benzerliğini değerlendirmek için otomatik prosedürlere olan talebi arttırmıştır.




Şekil 1.9: Fasta formatı (Edwards vd., 2009).

### 1.2.7.2 İnfluenza virüsünün antiviral ilaçlara karşı direnç gelişimi

Düşük duyarlılık (fidelity) ve sık genetik sürüklenme, influenza virüsünde çok çeşitlilik görülmesine sebep olmuştur. İnfluenza genomundaki tüm segmentlerin mutasyona uğrama dağılımları çıkarılmıştır. Nöraminidaz proteini için bu oran % 10.4'tür (Visher, Whitefield, McCrone, Fitzsimmons, & Lauring, 2016). İnfluenza sahip olduğu bu özellikler ile antikorlardan daha rahat kaçabilmekte ve ilaçlara karşı direnç göstermektedir.

Virüslerin antiviral ilaçlara karşı gösterdikleri etki ya da direnç iki tip deneysel analiz ile anlaşılabilmektedir. Bunlardan birincisi fenotipik analizlerdir. Fenotipik analizler viral yayılım sonucu % 50 inhibitör konsantrasyon ($IC_{50}$) değerini belirlemektedir ve NAI'lar için enzimatik tahliller tercih edilmektedir. $IC_{50}$ değeri, inhibe edilmemiş enzim (kontrol grubu) ile karşılaştırıldığında enzim aktivitesinin % 50'sini inhibe eden konsantrasyon olarak tanımlanmaktadır. NA'daki bir mutasyon nedeniyle proteinin ilaca karşı duyarlılığı azalmış ise $IC_{50}$ değeri yüksek çıkmaktadır. Dünya Sağlık Örgütü (World Health Organization, WHO)'ne göre yapılan tahlil sonuçlarında 10-100 kat ya da üst sınırların daha üstünde $IC_{50}$ değeri artıyorsa, NA proteini ilaca karşı direnç göstermektedir. İkinci yaklaşım olan genotipik analizler için en çok RT-PCR tercih edilmektedir. DNA Sanger sekanslama yöntemiyle sekanslar çoğaltılmakta ve ilaca karşı direnç gösterecek potansiyel mutasyonlar tespit edilebilmektedir (Boivin, 2013). Yapılan klinik çalışmalarda, 2008-2009 yılları arasında H1N1 virüsü tarafından enfekte olmuş hastaların % 90'nı ilaçlara karşı direnç kazandığı tespit edilmiştir (Mckimm-Breschkin, 2013).

Protein sekansında meydana gelen değişimler, amino asitlerin tek harf kısaltmaları ve değişimin olduğu pozisyonun sırası ile ifade edilir. Örneğin; 275. pozisyon için var olan amino asit, Histidin (H) iken değişim sonucu Tirozin (Y) amino asidine dönüşmüştür. Bu durum H275Y olarak ifade edilir. Tüm amino asit kısaltmaları ekler bölümünde verilmektedir. NA'nın H275Y mutasyonu başta olmak üzere direnç gösteren birçok amino asit değişikliği aşağıdaki Çizelge 1.1'de verilmiştir. H275Y mutasyonuna sahip NA proteinlerinin başka mutasyonlar ile birlikte $IC_{50}$ değeri üzerinde sinerjik etkiye sebep olup antiviral ilaçlara karşı duyarlılığın daha da azalmasına sebep olduğu bulunmuştur (Mihajlovic & Mitrasinovic, 2008);(Bloom, Gong, & Baltimore, 2010);(Hayden & De Jong, 2011);(Wu vd., 2013);(Baek vd., 2015). Çizelgede görülen bazı mutasyonlar ise klinik olarak gözlenmemiş ancak ters

genetik yöntemleriyle o mutasyonlar oluşturulup ilaç direnci kazandıkları bulunmuştur (Y Abed, Goyette, & Boivin, 2004);(Boivin, 2013).

Çizelge 1.1: Dirençli NA (H1N1 virüsü) mutasyonları.

| **Mutasyon** | **Lokasyon** | **Virüs Kaynağı** | **Referans** |
|---|---|---|---|
| I117R | Allosterik | Deneysel | (Gregory vd., 2017) |
| E119A | Ortosterik | Ters Genetik | (Baek vd., 2015) |
| E119A/H275Y | Ortosterik | Ters Genetik | (Baek vd., 2015) |
| E119D | Ortosterik | Ters Genetik | (Baek vd., 2015) (Yacine Abed vd., 2016) |
| E119D/H275Y | Ortosterik | Ters Genetik | (Baek vd., 2015) |
| E119G | Ortosterik | Klinik | (Baek vd., 2015) |
| E119G/H275Y | Ortosterik | Klinik | (Baek vd., 2015) |
| E119Q | Ortosterik | Ters Genetik | (Y Abed vd., 2004) |
| E119V | Ortosterik | Ters Genetik | (Abed vd. 2006) |
| Q136K | Allosterik | Klinik | (Nisn, 2010) (Boivin, 2013) |
| Q136K/D151E | - | Deneysel | (Okomo-Adhiambo vd., 2010) |
| Q136K/H275Y | - | Deneysel | (Okomo-Adhiambo vd., 2010) |
| Q136K/D151N/H275Y | - | Deneysel | (Okomo-Adhiambo vd., 2010) |
| Q136R | Allosterik | Deneysel | (Pizzorno, Abed, Rhéaume, Bouhy, & Boivin, 2013) |
| G147R/H275Y | - | Klinik | (Gregory vd., 2017) |
| T148I | Allosterik | Klinik | (Gupta, 2015) |
| I149V/H275Y | - | Klinik | (Yongkiettrakul vd., 2013) |
| D151E | Ortosterik | Deneysel | (Pizzorno vd., 2013) |
| D151N | Ortosterik | Deneysel | (Gregory vd., 2017) |
| D151E/N/H275Y | Ortosterik | Deneysel | (Okomo-Adhiambo vd., 2010) |
| Y155H | Allosterik | Klinik | (Monto vd., 2006) |
| R194G/H275Y | - | Klinik | (Wu vd., 2013) |
| D199E | Ortosterik | Deneysel | (Takashita vd., 2015) |
| D199G | Ortosterik | Deneysel | (Pizzorno, Bouhy, Abed, & Boivin, 2011) |
| D199N | Ortosterik | Klinik | (Baek vd., 2015) |
| D199N/H275Y | Ortosterik | Klinik | (Baek vd., 2015) |
| E214D/H275Y | - | Klinik | (Wu vd., 2013) |

| | | | |
|---|---|---|---|
| V234M/R222Q/H275Y | - | Klinik | (Bloom vd., 2010)<br>(Wu vd., 2013) |
| I223K | Ortosterik | Klinik | (Huang vd., 2014) |
| I223R | Ortosterik | Klinik | (Hayden & De Jong, 2011)<br>(Huang vd., 2014) |
| I223R/H275Y | Ortosterik | Klinik | (Boivin, 2013) |
| I223V | Ortosterik | Ters Genetik | (Hayden & De Jong, 2011) |
| I223T | Ortosterik | Klinik | (Huang vd., 2014) |
| V234M | Allosterik | Klinik | (Gupta, 2015) |
| F239Y/H275Y | - | Klinik | (Wu vd., 2013) |
| V241I | Allosterik | Klinik | (Gupta, 2015) |
| S247G | Ortosterik | Klinik | (Takashita vd., 2015) |
| S247N | Ortosterik | Klinik | (Gupta, 2015) |
| S247N/H275Y | Ortosterik | Klinik | (Boivin, 2013) |
| S247R | Ortosterik | Klinik | (Gregory vd., 2017) |
| G248R/I266V | Allosterik | Klinik | (Monto vd., 2006) |
| H275Y | Ortosterik | Klinik | (Monto vd., 2006)<br>(Boivin, 2013)<br>(Baek vd., 2015) |
| L250P/H275Y | - | Klinik | (Wu vd., 2013) |
| Q313R/I427T | Allosterik | Klinik | (Tu vd., 2017) |
| R293K | Ortosterik | Ters Genetik | (Baek vd., 2015) |
| N295S | Ortosterik | Ters Genetik | (Boivin, 2013)<br>(Baek vd., 2015) |
| D344N | Ortosterik | Klinik | (Gupta, 2015) |
| D354G | Allosterik | Klinik | (Gupta, 2015) |
| N369K | Allosterik | Klinik | (Gupta, 2015) |
| I427T | Allosterik | Klinik | (Nisn, 2010)<br>(Tu vd., 2017) |

Gözlemlenen nokta mutasyonları ya da ikili, üçlü mutasyonlar dışında proteinde oluşan tüm değişimleri inceleyebilmek için protein mutasyon haritaları oluşturulmaktadır. Bu haritaların çıkarılması için deneysel olarak iki yaklaşım vardır. İlk yaklaşım, tanımlanmış tekli mutasyonların karakterizasyonunu içerir. Enzimatik aktivitelerin ölçümleri yüksek performanslı sıvı kromatografisi ya da UV spektrofotometre ile yapılabilmektedir. Örneğin, bir enzimin farklı varyasyon yapıları alınarak farklı substratlarla etkileşime sokularak proteaz etkinliği ve stabilitesi serbest enerji hesaplamalarıyla elde edilebilmektedir. Ayrıca yeni katalitik aktivite ya da

enantiyo seçicilik seviyesi karakterizasyon ile görülebilmektedir. Akış sitometrisi, mikro akışkanlar, faj gösterimi veya büyüklük seçimi gibi yöntemlerle bir enzimin tüm tek amino asit mutantlarını kodlayan gen koleksiyonu oluşturulabilmektedir. Örneğin, Alanin tarama yöntemi ile bir sekansta bulunan tüm amino asitler alanin amino asidine dönüştürülerek fonksiyonel olarak etkisi ölçülmektedir. Ancak, kapsamlı ve doğru sonuçların elde edilmesi için çoklu örneklemelerin yapılması gerekmektedir. Buna çözüm olarak derin hizalama yöntemleri geliştirilmiştir. Mutasyon haritalarının çıkarılması için derin mutasyon taraması yapılarak protein-DNA ya da protein-protein etkileşimleri incelenebilmektedir. Bu yöntemlerle oluşturulan mutasyon haritaları, nötr, faydalı ve zararlı amino asit değişimleri hakkında sistematik bilgi sağladıkları için, enzimlerdeki sekans-fonksiyon ilişkilerini anlamamıza yardımcı olmaktadır (van der Meer, Biewenga, & Poelarends, 2016). Yeni metodolojiler ve teknik gelişmeler deneysel çalışmaların maliyetini düşürmekte ve daha önce yapılması imkansız olan araştırmaları mümkün kılmaktadır. Ancak, amino asitlerin değişimlerini ve davranışlarını incelemek için geliştirilen deneysel yöntemler henüz yeterli değildir. Bilgisayar tabanlı yöntemlere ihtiyaç duyulmaktadır (Hecht, Bromberg, & Rost, 2013).

### 1.2.8 Modelleme çalışmaları

Birçok gen hasarı ve proteinlerin üç boyutlu yapısı deneysel olarak tanımlanabilmektedir. Deneysel olarak tanımlanmış yapılar ve moleküler modeller mutasyonların yorumlanması için bir yol göstericidir. Ancak sekans varyasyonlarının ve genetik hastalıkların moleküler mekanizmalarının özellikle geniş sayıda mutasyona sahip kanser hastalığının anlaşılması, deneysel olarak çok pahalı, zaman alıcı ve zordur. Amino asitlerin birbirlerine dönüşmelerinin yani mutasyonların etkileri teorik yöntemlerle daha kolay bir şekilde incelenebilmektedir. Genotip-fenotip korelasyonunun anlaşılması için gerekli olan protein yapı-fonksiyon ilişkileri farklı teorik çalışmalarla analiz edilmiştir (Thusberg & Vihinen, 2009).

#### 1.2.8.1 Protein sekans hizalama

Ortak bir atadan evrimleşen diziler, eşdeğer amino asit pozisyonlarında benzer amino asitleri paylaşır. Ortak ataya sahip dizilerin evrimi sırasında olan değişimlerin yorumlanması için sekans hizalama yöntemlerine ihtiyaç vardır. Sekans hizalama, alt

alta sıralanmış iki sekansta karşılık gelen konumları eşlemeyi ifade eder. İki sekansın özdeş olması durumunda, her konumdaki alfabe (amino asit), başka bir sekanstaki alfabeyle (amino asit) eşleşecektir. Evrimleşme sırasında muhtemel konumlarındaki amino asitlerin birbirine dönüşmesi ihtimali dışında ekstradan bir amino asit eklenebilir veya olan amino asit silinebilir. Bu durumda, iki protein sekansının hizalanması ile yukarıda bahsedilen durumlar karşısında en doğru alt alta eşleşme ile sekans dizilerinin benzerlikleri incelenebilir. Optimal hizalamanın elde edilmesi için alt alta hizalanmış amino asit çiftlerinin puanlandırılması gerekir. Çoğu hizalama algoritması, iki sekans için benzerlik skorunu maksimuma çıkararak anlamlı hizalamalar üretmeye çalışır. Bu işlemler uzun sekans dizileri için zor bir işlem haline gelir. Hizalama işlemlerinin daha verimli bir şekilde çözüme ulaşması için dinamik programlama yöntemleri kullanılır (Mathura & Kangueane, 2009).

En çok kullanılan iki tip hizalama yöntemi vardır. Global hizalama yönteminde, iki sekans bir bütün halinde alınarak hizalama yapılır. Çoğunlukla elde edilen global hizalama sonuçları, farklı amino asitler veya boşluklar ile eşleştirilen uzun sekans uzantıları içerebilir. Tersine, eğer algoritma korunmayan (değişime uğrayan) bölgeleri göz ardı ederek korunan alt dizileri hizalamaya çalışırsa, o zaman lokal hizalama olarak adlandırılır. İki sekansın lokal hizalamasıyla birçok alt hizalama grupları oluşabilir. Şimdiye kadar, sadece iki dizinin karşılaştırıldığı durum tarif edilmiştir. Buna çift yönlü (pairwise) hizalama denir. İkiden fazla dizinin eş zamanlı olarak karşılaştırıldığı duruma çoklu hizalama denir (Edwards vd., 2009). Genetik materyallerin evrimsel zamana göre nasıl değiştiğini incelemek için çoklu dizi hizalamaları kullanılır (Şekil 1.10). Puanlama yapılırken kullanılan skorlama fonksiyonlarının pozisyona özgü puanlama yapması ve sekansların filogenetik ağaç ile evrimsel ilişkilerinin çıkarılması dikkate alınan iki önemli özelliktir. Evrimsel ilişki modelinin çıkarılması oldukça karmaşıktır. Evrimsel süreçte doğal seleksiyonun getirdiği pozisyona özgü yapısal ve fonksiyonel kısıtlamalar ile sekanslar üzerinde belli bölgeler korunmaktadır. Bu bölgelerin hizalanmasıyla birlikte değişime uğrayan bölgeler de en ideal şekilde hizalanır (Durbin, Eddy, Krogh, & Mitchison, 1998). Çoklu sekans hizalama için birçok sayıda yöntem mevcuttur. Clustal W (Thompson, Higgins, & Gibson, 1994) gibi yöntemler benzer sekanslar için makul doğrulukta sonuç verir. Ancak uzak akrabalığa sahip sekanslar için doğru hizalamanın yapılması zordur ve yapılan çalışmalar çok kapsamlı bir MSA yönteminin mevcut olmadığını ve

kendine özgü güçlü ve zayıf yönleri olduğunu göstermektedir. Bu durum, en uygun hizalama yönteminin seçimini zorlaştırır. En yaygın olarak kullanılan dizi hizalama yöntemleri Çizelge 1.2'te listelenmiştir. (Thusberg & Vihinen, 2009).

Çizelge 1.2: Çoklu sekans hizalama (MSA) metotları.

| | | |
|---|---|---|
| Clustal Omega | https://www.ebi.ac.uk/Tools/msa/clustalo/ | Sieverd F., 2011 |
| MAFFT | https://www.ebi.ac.uk/Tools/msa/mafft/ | (Katoh vd.,2002) |
| PROBCONS | http://probcons.stanford.edu/ | Do vd., 2005 |
| PROMALS | http://prodata.swmed.edu/promals/ | Pei vd, 2007 |
| T-Coffee | https://www.ebi.ac.uk/Tools/msa/tcoffee/ | Notredame vd., 2000 |
| MUSCLE | https://www.ebi.ac.uk/Tools/msa/muscle/ | Edgar, 2004 |
| WebPRANK | https://www.ebi.ac.uk/goldman-srv/webprank/ | Löytynoja, A., Goldman, N. (2010) |


Şekil 1.10: Çoklu sekans hizalama örnek seti. Rastgele alt alta dizilmiş 4 sekans hizalama işlemi ile aynı ya da benzer olan amino asitler alt alta gelecek şekilde düzenlenmiştir.

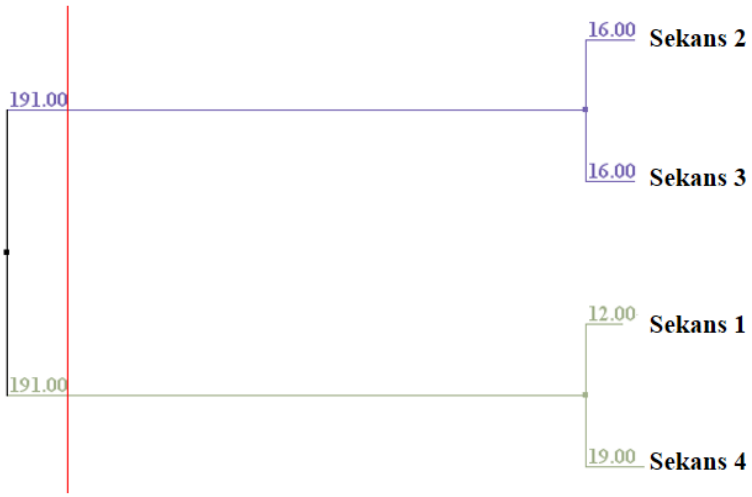
### 1.2.8.2 Filogenetik ağaç oluşturma

Moleküler filogenetik, biyolojik sekansların evrimini ve aralarındaki tarihsel ilişkileri inceler. Çoklu sekans hizalama sonuçlarının ağaç üzerinde görselleştirilmesidir. Filogenetik ağaçlar, incelenen tüm sekansların ortak bir atayı paylaştığını ve ağaçtaki tüm dallar boyunca evrimleşen sekansların bağımsız olarak geliştiğini gösterir (Keith, 2008).

Filogenetik ağaç oluşturma yöntemleri ya uzaklığa bağlı ya da karakter tabanlı olmaktadır. Uzaklığa bağlı yöntemlerde, her bir sekans çifti arasındaki mesafe hesaplanır ve sonuçta elde edilen mesafe bir matris olarak ifade edilir ve her bir

hesaptan sonra yeniden bu matris yapılandırılır. Örneğin, komşu birleştirme (neighbour-joining) yöntemi, tamamen çözülmüş bir filogenetiğe ulaşmak için mesafe matrisine bir küme algoritması uygular. Karakter tabanlı yöntemler, maksimum parsimony, maksimum olabilirlik ve Bayesci çıkarım yöntemlerini içerir. Bu yaklaşımlar, aynı anda, her bir ağaç için bir skor hesaplamak için bir karakter (hizalamadaki bir bölge) göz önünde bulundurularak, hizalamadaki tüm dizileri karşılaştırır. Ağaç skoru, maksimum parsimony için minimum değişiklik sayısı, maksimum olabilirlik için log-olabilirlik değeri ve Bayesci çıkarım için önsel (posterior) olasılıktır (Yang & Rannala, 2012). Maksimum olabilirlik, filogenetik ağaçta evrim modeline özgü tanımlanan parametreleri kullanarak elde edilen olasılık değerinin, dal uzunluğuna bölünmesi ile elde edilir. Bu sonuç, sekansların zaman içerisinde nasıl bir değişime uğradığını ve soylarından ne kadar farklılık gösterdiğini tanımlar. Bölüm 1.2.8.1'de verilen dört sekans için oluşturulmuş filogenetik ağaç Şekil 1.11'de görülmektedir.

Teorik olarak, mümkün olan tüm ağaçları karşılaştırarak en iyi skoru olan ağaç tanımlanmalıdır. Bu istatiksel yöntemler genetik alanından dünya ekonomisine kadar pek çok araştırmaya uygulanmış, yerleşik ve güvenilir bir metodolojiye sahiptir (Keith, 2008).




Şekil 1.11: Komşu birleştirme metodu ile oluşturulan filogenetik ağaç (Jalview).

### 1.2.8.3 Proteinlerin evrimsel korunma mekanizmaları

Çoklu sekans hizalamaları, diziler arasındaki yapısal, fonksiyonel veya evrimsel ilişkileri belirlemek için yaygın olarak kullanılmaktadır. Bir hizalamada gözlemlenen amino asit değişimlerinin çoğu nötrdür. Bu durum, proteinin bu pozisyonda ne kadar

toleranslı olduğunu belirtmektedir. Örneğin, hemoglobin proteininin en işlevsel kısmı heme grubu, proteindeki diğer amino asit bölgelerine göre daha az değişime uğrar. Tolerans gösteremeyen proteinlerde işlevsel olarak farklılıklar ortaya çıkabilmektedir (Valdar, 2002). Protein sekans-yapı-fonksiyon çalışmalarından biri proteinlerin pozisyona bağlı evrimsel korunma düzeyi ve tipini belirlemeye dayalıdır. Evrimsel korunma, protein içerisindeki amino asit pozisyonlarının protein yapı ve işlevine ne kadar etki ettiğine göre derecelendirilmektedir. Protein fonksiyonları için tehlikeli ve etkisiz mutasyonların belirlenmesi, buna ek olarak amino asitlerin fizikokimyasal özelliklerinin (hidropati, yük, boyut vs.) yapısal bütünlüğe etkisinin incelenmesi evrimsel koruma çalışmaları içerisindedir (Miller & Kumar, 2001).

Protein sekanslarının konumsal (amino asit pozisyonu) koruma analizi için çeşitli yöntemler kullanılmaktadır. En sık kullanılan konumsal koruma yöntemlerinden biri sum-of-pairs (SP) skorlarıdır. Bu yöntem hizalanmış sekansların her kolonundaki amino asit çiftleri arasındaki benzerliklerin toplamını hesaplayarak korumayı tanımlar. Benzerlik değerleri ise skorlama matrislerinden elde edilir (Karlin). Kullanılan bir başka konumsal koruma yöntemi amino asitlerin hizalanmış sekans içerisindeki frekans oranlarını (entropi bilgisini) kullanarak sekans değişkenliğini hesaplamaktadır (Sander Schneider). Sekans seti içerisinde tekrar eden sekansların sistemi baskılamaması için ağırlıklı skor hesaplamaları da kullanılmaktadır. Bunun dışında olasılık hesaplaması yapan maksimum olabilirlik yöntemi gibi istatistiksel yöntemler de kullanılmaktadır. Şu anda ise makine öğrenimi (machine learning) tabanlı yöntemler kullanılmakta ve geliştirilmektedir.

ConSurf adındaki internet tabanlı program sekans içerisindeki her bir pozisyonun evrimsel mutasyon skorunu 6 farklı skorlama matrisi kullanma seçeneği sunarak bayesci çıkarım yöntemiyle ya da maksimum olabilirlik yöntemiyle hesaplamaktadır. ConSurf, sonuçları referans alınan bir protein/DNA/RNA üzerinde görsel olarak renklendirerek göstermektedir. Bir proteindeki bir amino asit sadece sekans diziliminde bulunan çevre amino asitleriyle etkileşimde olmaz. Katlanma yapıları olduğundan etkileşim içerisinde olan amino asit grupları ancak üç boyutlu protein yapısı üzerinden gözlenebilmektedir. Bu yöntemin kullanılmasıyla sekans üzerindeki bir amino asit değişimi ya da korunumu üç boyutlu yapı üzerinde tanımlanabilmekte ve daha kolay bir şekilde incelenebilmektedir (Landau vd., 2005);(Ashkenazy vd., 2016).

Bu alanda yapılan başka bir çalışma ise proteinlerin amino asit pozisyonlarında meydana gelen her olası değişim (kendisi dışında 19 amino asit) için mutasyon haritası çıkarılarak SNP'lerin bir diğer adıyla tek amino asit değişimlerinin sekans üzerindeki etkisinin araştırılmasıdır.

SIFT (Sorting Intolerant From Tolerant) adındaki bir program, protein fonksiyonuna etki edecek amino asit değişimlerini var olan skorlama matrislerini kullanmadan kendisi pozisyona özel matrisler oluşturarak ifade eder. SIFT ilk olarak verilen bir protein sekansı için benzer sekanslardan bir set oluşturur. Bu protein sekanslarının, çoklu sekans hizalanmasını elde eder. Hizalamadaki her pozisyonda görünen amino asitlere bakarak olasılık hesabı yapar. Belirlenen eşik değeri altında hesaplanan olasılıklara sahip amino asitleri tehlikeli amino asit olarak kabul eder. Bu şekilde yapılan tahminlerle protein fonksiyonunda değişikliğe sebep olabilecek olası amino asit değişimleri belirlenir (Sim vd., 2012);(Ng & Henikoff, 2001).

SNP'lerin etkilerini inceleyen bir başka program SNAP'tir (Screening for non-acceptable polymorphisms). SNAP nöral ağ tabanlı bir program olup girdi olarak sadece sekans bilgisini alarak tehlikeli SNP'lerin fonksiyonel etkilerini tahmin etmeye çalışır. SIFT'te olduğu gibi pozisyona özel matris kullanılır. Sekans bilgisini kullanarak ikincil yapı ve çözücü erişilebilirliği (solvent accessibility) gibi bilgileri de tahmin etmeye çalışır. SIFT ile SNAP karşılaştırıldığında tehlikeli ve nötr SNP'leri SNAP % 78 doğrulukta bulurken SIFT % 74 doğrulukta bulmaktadır (Bromberg & Rost, 2007). Bilgisayar tabanlı yöntemlerle mutasyon haritalarının çıkarılması yeni deney ihtiyaçlarının belirlenmesi için ve yeni nesil sekanslama yöntemlerinin yorumlanması için yol göstermektedir.

Başka bir koruma mekanizması, kovaryasyonlardır. Kovaryasyon, protein içinde olan bir mutasyona bağlı olarak başka bir bölgede mutasyon olması anlamına gelmektedir. Kovaryant amino asitlerin pozisyonlarından dolayı fonksiyonel olarak işlevleri belirgin olmayabilir ancak proteinlerin önemli konformasyonlarında fiziksel bir etkileşim oluşturarak bir işlev gösterebilirler (Thusberg & Vihinen, 2009). Fiziksel temas içerisinde bulunan amino asitler birbirlerine bağlı olarak mutasyon geçirirler. Bir amino asit mutasyona uğrarsa temasta bulunduğu diğer amino asit de mutasyona uğrama eğilimi gösterir. Bir çalışmada çoklu sekans hizalaması yapılarak her bir pozisyon için değişim matrisi çıkarılmıştır. Bu matrislerden yararlanılarak her iki pozisyon arasındaki temas haritası çıkarılmıştır. Çıkarılan haritalar deneysel olarak

elde edilmiş temas haritalarıyla karşılaştırılmıştır. 11 protein ailesi için yapılan karşılaştırmada % 37 ile % 55 arası doğru tahmin sonuçları elde edilmiştir. Doğruluk oranının arttırılması için korelasyona sahip mutasyonların çoklu hizalamada yapısal ve fonksiyonel grupların ayrılabilmesiyle olabileceğini belirtmektedirler (Gobel, Sander, Schneider, & Valencia, 1994). Bu sayede yapısal olarak değişime uğramış olsa da fonksiyonel kısımda bir etki yaratmayan amino asitler tanımlanabilir. Bir başka çalışmada 4 farklı kovaryans algoritmalarının performansları ve doğruluk oranları karşılaştırılmıştır. Pfam veri seti ile yapılan hizalamalar için, McBASC ve OMES algoritmalarının kovaryasyona sebep olan amino asit eşlerini bulmada SCA ve MI algoritmalarına göre daha başarılı olduğu sonucuna varılmıştır. Kovaryans algoritmaları genellikle farklı derecelerde background korunma frekanslarını dahil ettikleri için McBASC ve OMES algoritmaları Pfam veri bankasındaki proteinlerin evrimine daha uygun bir algoritmaya sahip olduğundan daha doğru tahminler vermiştir. Sonuç olarak farklı algoritmaların kombine hale gelmesiyle diğer veri bankaları için de uygulanabilecek algoritmaların oluşturulabileceği öne sürülmüştür (Fodor & Aldrich, 2004).

Hopf ve arkadaşları mutasyonların etkilerini incelemek için mutasyonlar arası genetik ilişkiyi ifade eden epistaz bilgisini de dahil ederek amino asitlerin pozisyonlarına bağlı birbirleriyle olan etkileşimlerini ifade eden bir istatistiksel yöntem kullanarak mutasyonların etkilerini tahmin etmişlerdir. Bu istatiksel yöntemde bir pozisyonda görülen amino asitlerin değişimini ve o pozisyonuna komşu olan diğer amino asitlerin oluşan değişime karşı etkilerini bir enerji fonksiyonu oluşturarak incelemişlerdir. Elde ettikleri tahminlerin önceki biyolojik verilerle tutarlı olduğunu, epistatik etkileşimin sisteme dahil edilmesiyle daha doğru mutasyonların değerlendirildiğini ve yeni protein sekans kütüphanelerinin tasarımına yardımcı olacağını belirtmişlerdir. Çalışmalarını ayrıca deneysel olarak elde edilen yüksek verimli mutagenez sonuçlarıyla karşılaştırmışlardır. Şu anda Evmutation evmutation.org platformu içerisinde yaklaşık 7.000 insan proteini için tahmin yapılmıştır (Hopf vd., 2017).

Bu bölümde bahsedilen evrimsel örnekleme çalışmalarında sekansların bir kısmının kendi içinde birçok kere tekrar ediyor olması bir başka deyişle sistemi domine ediyor olması, evrimsel sürecin bağlı olduğu parametrelerin fazla olması, yeterli deneysel sekans verisinin olmaması bilgisayar tabanlı sonuçların yorumlanmasını zorlaştırmaktadır (Hopf vd., 2017). Genetik varyasyon ve evrim mekanizmalarının

analizlerinin daha doğru sonuç vermesi için, yukarıda bahsedilen limitasyonlara çözüm bulunması gerekmektedir.

Antijenik sürüklenmeyle yeni sekans yapılarının oluşması nedeniyle, mevsimsel grip aşılarının etkili kalması için sık sık güncellenmeye ihtiyacı vardır. Neher ve arkadaşları, virüsler arası antijenik mesafenin genetik farklılıklarla ilişkisine dayanarak, ölçülen antijenik verileri yorumlamak ve antijenik olarak karakterize edilmemiş virüslerin özelliklerini tahmin etmek ve gelecekteki grip virüsü popülasyonlarının bileşimini tahmin etmek için bir model geliştirmişlerdir. Nextflu adında geliştirilen bu program influenza virüsleri arasında genetik ilişkileri neredeyse gerçek zamanlı olarak incelemektedir (Neher, Bedford, Daniels, Russell, & Shraiman, 2016). Bu program sekansların filogenetik ağacını oluşturarak maksimum olabilirlik metodunu kullanarak mutasyon frekanslarını çıkarmaktadır. Ayrıca deneysel olarak elde edilmiş hemaglutanin proteini inhibisyon testleriyle birlikte olası sekansları tahmin etmektedir. Filogenetik ağaç üzerindeki antijenik özellikleri ve Hemaglutanin İnhibisyon (HI) titer verilerinin modellerini influenza virüsünün evrimini görselleştirmek için interaktif izleme aracı olan nextflu.org platformuna entegre etmişlerdir. Mevsimsel grip aşılarında kullanılacak suşların seçimi için nextflu önemli bir program haline gelmiştir (Neher & Bedford, 2015). Diğer virüslerin de incelenebileceği bir platforma sahip olan Nextflu daha sonra Nextstrain nextstrain.org olarak yeni bir platformda, endemik viral hastalıkları (mevsimsel influenza, dang) ve pandemik viral salgınları (kuş gribi, Zika, Ebola) incelemekte ve sekans tahmin sonuçlarını güncellemektedir (Hadfield vd., 2018). Ancak oluşturdukları model aynı amino asit pozisyonunun, filogenetik ağacın farklı bölümlerinde birçok kez mutasyona uğraması durumunda tahminlerde hatalara sebep olmaktadır.

Şu ana kadar oluşturulan modellerde ve fonksiyonlarda tam olarak virüslerin nasıl bir evrimsel süreçten geçtiği kesin olarak bulunamamıştır. Rastsal mutasyonların meydana gelmesi sistemi komplike bir hale getirmektedir. Bu nedenle tam doğru sonuç veren bir tahmin metodunun oluşturulması zordur. En yakın tahmin modelinin oluşturulması için birçok çalışma grubu çalışmalarına devam etmektedir. Bu tez kapsamında biz de zaman bilgisini var olan bazı modellere ve yeni oluşturduğumuz modellere entegre ederek sistem üzerinde nasıl bir etkiye sahip olduğunu inceledik. Kullanılan metodlar gelecek bölümlerde detaylı olarak anlatılmaktadır.

# 2. MATEMATİKSEL MODELLEME VE SAYISAL YÖNTEMLER

Giriş bölümünde de detaylı bir şekilde anlatıldığı üzere, evrimsel süreçte meydana gelen değişimler sonucu virüs yayılımına engel olmak için yeni ilaç tasarımları ya da virüs üzerinde yeni ilaç hedefleme bölgeleri araştırılmaktadır. Bu proje kapsamında influenza A grip virüsünün yüzey proteini nöraminidaz (NA), hedef olarak alınmıştır. Bu yüzey proteini, virüsün konak hücreden dışarı çıkmasına yardımcı olmaktadır. Bu proteinin işlevini yerine getirememesi durumunda, virüs diğer hücrelere yayılamamaktadır. Bu nedenle nöraminidaz proteini birçok bilimsel çalışmada hedef olarak kullanılmıştır. Bu projede hedef olarak kullanılmasının sebebi nöraminidaz proteini değişime uğradığı için ilaçların etkisi azalmakta ya da ilaçlar etki gösterememektedir. Bu durum grip hastalıklarının dünyada yaygınlaşmasına sebep olmaktadır. Grip virüsünün evrimsel değişiminin anlaşılması bu nedenle büyük bir önem arz etmektedir.

NA proteinindeki değişimlerin incelenmesi için farklı şehirlerden ve yıllardan toplanmış olan sekans verilerinin evrimsel ilişkisinin çıkarılması gerekmektedir. Bu bilgiden yararlanılarak NA proteini üzerinde değişime karşı hassas olan ve korunan bölgeler, farklı modeller kullanılarak tespit edilecektir. Kullanılan ve geliştirilen modeller (fonksiyonlar) bu bölümde detaylı olarak bahsedilmektedir. Elde edilen model sonuçları haritalar üzerinde resmedilmiştir. Bu haritalar, NA protein sekansı üzerindeki pozisyonlarda görülebilecek değişme olasılıklarını göstermektedir.

NA proteinin evrimsel değişiminin tahmin edilebilmesi için pozisyon bilgisine ek olarak o pozisyonlarda görülebilecek olası amino asitlerin bilinmesi gerekmektedir. Bu bilgiyi ise amino asit frekansları hesaplanarak elde edilebilir. Doğada her amino asidin birbirine dönüşme olasılığı aynı değildir. Her amino asit sadece bir tane 3'lü nükleotid tarafından değil birkaç kodondan üretilebilmektedir. Bu nedenle kodon tablosu yardımıyla her amino asidin üretilme olasılıkları hesaplanmış ve beklenen frekansları elde edilmiştir. Ayrıca veri seti içerisindeki amino asit dağılımlarına bakılarak doğal frekans hesabı ile amino asitlerin birbirlerine dönüşme olasılıkları elde edilmiştir.

Her virüs için evrimsel süreç aynı hızda ilerlememektedir. NA proteininin mutasyon hızı bilgisi gelecekte karşılaşabileceğimiz farklı NA proteini varyasyonlarının tahmin edilebilmesi için önemlidir. Bu bilgi ile protein üzerinde gözlemlenebilecek bir değişimin ne kadar sürede meydana gelebileceğini ya da belli bir süre içerisinde NA üzerinde kaç farlı yerde değişim gözlemlenebileceği bilgisi elde edilecektir.
Mutasyon haritaları, amino asit frekansları ve mutasyon hızı bilgisi ile NA proteinin evrimsel değişiminin tahmin edilmesi amaçlanmıştır.

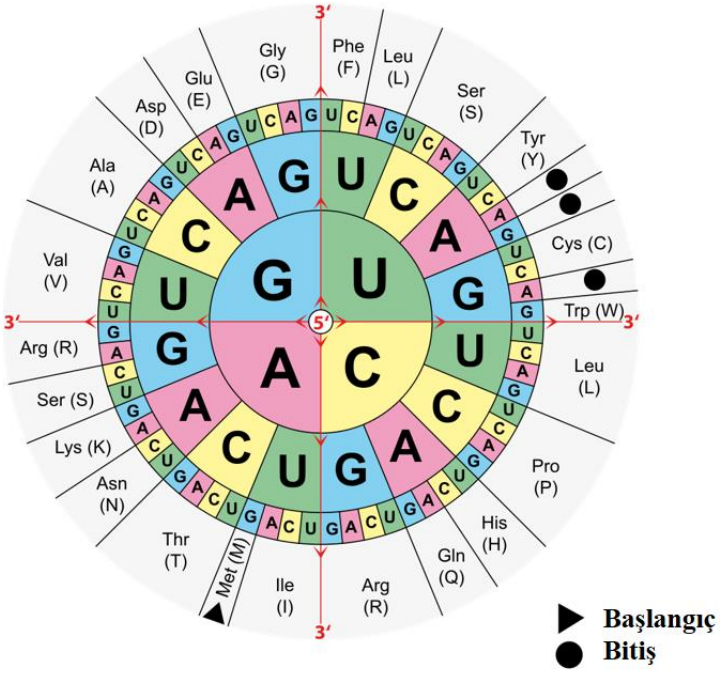
## 2.1 Nöraminidaz Proteininin Evrimsel İlişkisinin İncelenmesi

Örnek olay incelemesi olarak seçilen influenza A virüsü yüzey proteini nöraminidaz (NA) için, ilk olarak sekans verileri toplanmıştır. Bunun için geniş ve kapsamlı virüs verisine sahip olan Influenza Research Database (IRD veya fludb) ve Influenza Virus Resource kullanılmıştır. Fludb, kuşlardan ve memelilerden izole edilen influenza virüslerinin fenotipik, genomik ve proteomik verilerini içermektedir (Url-3). Bu veri bankalarından yararlanarak, 1918-2018 yılları arasında insandan izole edilen NA (H1N1 virüsü) protein sekanslarını içeren bir veri seti oluşturulmuştur.
Eksik tanım içeren, ülke, yıl bilgisi eksik olan, tanımlanamayan amino asit sayısı çok olan (X içeren) sekanslar, oluşturduğumuz veri seti içerisinden elenmiştir. Eleme sonrasında sekansların yıllara ve bölgelere göre nasıl bir dağılım gösterdiği incelenmiştir.
Veri seti içerisinde farklı yıllarda ya da farklı bölgelerden gelen ve tamamen aynı sekansa sahip veriler de bulunmaktadır. Bu tip aynı sekansların oluşturulacak olan sistem içerisinde baskın gelerek hatalı sonuçlara neden olmamaları için bu sekanslar geliştirilen bir kod kullanılarak veri seti içerindeki her bir sekans birbirinden farklı olacak şekilde sadeleştirilmiştir. Bu sadeleştirilmiş (non redundant) set bu tez kapsamında ana girdi seti olarak kullanılmıştır.
Derlenen veri setindeki sekansların evrimsel ilişkileri filogenetik ağaç oluşturularak incelenmiştir. Filogenetik ağaç, aynı evrim ağacı gibi, virüsler arası yakınlık ilişkilerini vermektedir. Yıllar içinde NA proteininde meydana gelen değişimler ile bu ağacın dallanma sayısı artmıştır. Filogenetik ağaç oluşturulmadan önce çoklu sekans hizalama yöntemlerinden biri olan Clustal Omega programı kullanılarak veri setindeki sekanslar skorlama matrislerine göre hizalanmıştır. Jalview adında bir program aracılığıyla ortalama mesafe metoduna göre filogenetik ağaç oluşturulmuştur. Bu

program içerisinde ayrıca dizi hizalamaları görüntülenebilmektedir. Bu sayede filogenetik ağaçta görülen gruplaşmalar ya da ayrışmalar sekans dizileri üzerinde renklendirilerek görüntülenebilmektedir.

## 2.2 Amino Asitler Arası İlişki ve Skorlama Matrisleri

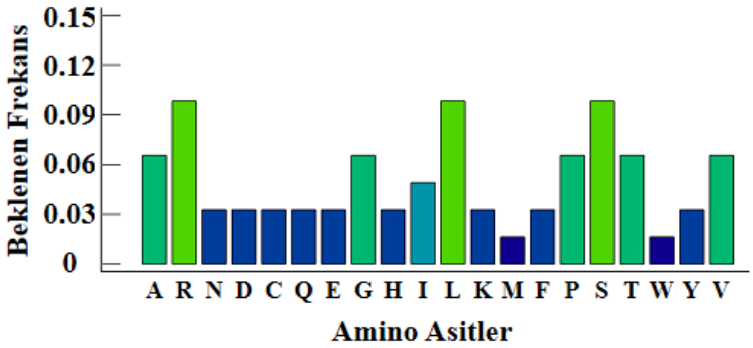
Ortak bir atadan evrimleşen protein sekansları, benzer amino asit dizileri ve pozisyonları içermektedirler. Bir proteinin evrimi sırasında, belirli bir pozisyondaki amino asit, farklı bir amino asidin (ör. izolösin, valinin yerini alır) yerine geçebilir. Şekil 2.1'de bulunan tabloda her bir amino asidi oluşturan 3'lü nükleotidlerin (kodon) farklı kombinasyonları bulunmaktadır. Bu tablo okunurken içten dışa doğru toplam 3 adımda her bir adımda 4 nükleotidden biri seçilerek gerçekleştirilir. Bu kombinasyonlara bakıldığında bir amino asit birden fazla kodon tarafından oluşturulabilmektedir ve her amino asidin oluşma olasılığı aynı değildir. Kodon tablosundan yararlanılarak elde edilen beklenen amino asit frekansları Şekil 2.2'de bulunmaktadır.




Şekil 2.1: Kodon tablosu (Url-9).

Beklenen amino asit frekanslarıyla doğada her zaman karşılaşılamamaktadır. Farklı protein ailelerinde her amino asit için farklı bir frekans dağılımı görülebilmektedir.

Görülen bu farklılıkların bazıları, beslenme koşulları ve çevre gibi dış faktörlerden, bazıları ise, proteinin ikincil yapısı, büyüklüğü ve yüzey/hacim oranı gibi fiziksel faktörler ile açıklanabilir. Bu nedenle her sekans seti için ayrıca gözlemlenen amino asit frekansları hesaplanmaktadır.




Şekil 2.2: Amino asitlerin doğada beklenen frekansları.

Belirli bir amino asit veya nükleotidin bir diğerinin yerini alma olasılığı sayısal olarak ifade edilmelidir. Bu bilgi bir skorlama matrisi (benzerlik matrisi) olarak adlandırılan bir matriste tanımlanmaktadır. Bu matrisler amino asitler arasında gözlemlenen değişim frekanslarına dayalı olarak yapılandırılır.

$Q_{i,j}$, $i$ ve $j$ amino asitleri arasındaki değişim frekansı, bir diğer adıyla doğal frekansı $P_i$ ve $P_j$ sırasıyla $i$ ve $j$ amino asitlerinin beklenen frekansı olduğu kabul edilirse yer değiştirme skoru $S_{i,j}$, doğal ve beklenen frekans değerlerine bağlıdır. Bu bağıntı Eşitlik (2.1)'de görülmektedir (Lipman, Wilbur, Smith, & Waterman, 1984).

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

Yakın derecede evrimleşmiş protein dizileri hem yapıyı hem de işlevi korur (Wilson, Kreychman, & Gerstein, 2000). Eğer iki sekans, ortak bir atadan evrimleşirse, o zaman homolog sekanslar olarak adlandırılır. Homolog dizilerin % 100 özdeş olması gerekmez ve benzer bir amino asit eşdeğer konumda yer alabilir. Genel olarak yapıları ve işlevleri için önemli olan kısımlar korunur. Benzer fizikokimyasal özellikleri paylaşan amino asitler sıklıkla daha yüksek frekanslarla değişmektedir. Bir diğer deyişle, benzer olan amino asitlerin birbirlerine dönüşme olasılıkları yüksektir (Mathura & Kangueane, 2009). Örneğin, hidrofobik ve nötr bir amino asit olan izolösinin, bir başka hidrofobik ve nötr amino asit olan valine dönüşme frekansı,

pozitif yüklü bir amino asit olan arjinine dönüşme frekansından yüksektir. Margaret Dayhoff, doğal olarak gözlenen amino asit frekansından ilk skorlama matrisini türetmiştir (Çizelge Ek 1). Hemen hemen aynı sekanslara sahip olan birkaç protein ailesinin her bir pozisyonunda meydana gelen değişimleri manuel olarak hesaplayarak evrimsel benzerliklerini ortaya çıkarmıştır. Böylece, gözlemlenen mutasyonlardan/değişimlerden, iki amino asidin birbirinin yerine geçme olası frekansını bulmaktadır (Dayhoff, Schwartz, & Orcutt, 1978). Dayhoff, Point-Accepted Mutation'ı (PAM) evrimsel bir ayrışma birimi olarak tanıttı. Bir PAM birimi, bir protein dizisinde 100 pozisyon arasında 1 amino asit değişimi olarak tanımlanır. Bu gibi dizi kümelerine dayalı olarak yapılan skorlama matrisi PAM1 matrisidir. PAM250 ise çok farklı sekanslar arasında beklenebilecek değişimlere karşılık gelir. PAM matrisi yakınlık derecesi yüksek olan sekanslar için kullanılır. Steven ve Jorja Henikoff, BLOSUM (Blok Değiştirme Matrisi) matris setini oluşturdu (Çizelge Ek 2) (Henikoff & Henikoff, 1992). Bir protein ailesinin global hizalanmasına (Bakınız Bölüm 1.2.8.1) dayanan amino asitlerin değişim frekansını türetmek yerine, lokal hizalama veya bloklar kullanmışlardır. BLOSUM, PAM matrislerinin oluşturulmasındaki gibi, bir ekstrapolasyon tekniği ile protein dizileri arasında gözlemlenen değişimler kullanılarak yapılandırılmıştır. Bu bloklar protein ailesi içindeki korunmuş segmentleri temsil eder. % 62 özdeş amino asit paylaşan diziler BLOSUM62 matrisi olarak adlandırılır. Birkaç özdeş sekansı paylaşan oldukça farklı sekanslar için BLOSUM30 kullanılmalıdır. Her ikisi de 24 × 24 matristir ve her iki durumda da korunan amino asitler yüksek puanlar verirken, beklenmeyen yer değiştirmeler daha düşük puanlar vermektedir. Orijinal PAM veya BLOSUM matrislerine bağlı çeşitli skorlama matrisleri türetilmiştir. GONNET (Benner, Cohen, & Gonnet, 1994) (Çizelge Ek 3) ve PET (Jones, Taylor, & Thornton, 1992) (Çizelge Ek 4) adlı skorlama matrisleri mevcut dizilerin kapsamlı bir listesini kullanılarak türetilmiştir.

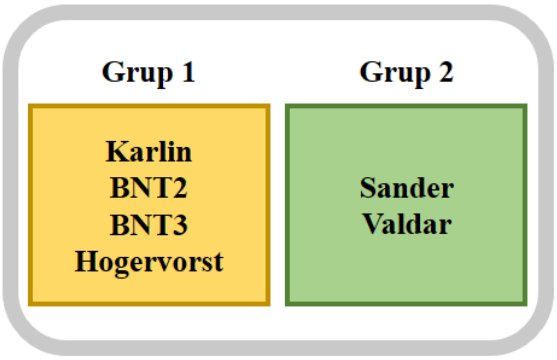
## 2.3 Nöraminidaz Proteininin Bölgesel Mutasyon Eğilimlerinin Hesaplanması

Bir proteinin yapısını ve işlevini koruyabilmesi, protein sekansındaki amino asitlerin korunmasına bağlıdır. Çoklu hizalama yöntemleri kullanıldıktan sonra her bir amino asit pozisyonunda meydana gelmiş olan amino asit değişimleri, bir skor ile ifade edilmektedir. Bu skorlar amino asit korunumu hakkında bilgi vermektedir. Son otuz

yılda, pek çok skorlama fonksiyonu oluşturulmuştur, ancak her bir fonksiyon farklı derecelerde parametreleri dahil ettikleri için standart bir fonksiyon ortaya çıkarılamamıştır (Valdar, 2002). Bu bölümde farklı skorlama fonksiyonlarının özellikleri ve formülasyonları incelenmektedir ve tanımlanmaktadır. Bunun yanında bu projede oluşturulmuş skorlama fonksiyonları bulunmaktadır ve bu fonksiyonlar var olan diğer skorlama fonksiyonlarıyla karşılaştırılmıştır.

### 2.3.1 Skorlama fonksiyonları

Valdar ve çalışma grubu arkadaşları var olan skorlama fonksiyonlarını özelliklerine göre gruplara ayırarak incelemişlerdir. İncelemeye göre sadece frekans içeren fonksiyonların veya sadece benzerlik matrisi ile hesap yapan fonksiyonların, yani sadece tek parametre kullanılarak oluşturulan fonksiyonların diğer fonksiyonlara göre daha fazla hata payına sahip olduğu görülmüştür (Valdar, 2002).
Bu sebepten dolayı en düşük hata payına sahip olduğu gösterilen 2 fonksiyon grubu bu tez kapsamında incelenmiştir (Şekil 2.3). Bu iki grup içerisindeki fonksiyonlar skorlama matrislerini kullanarak hesap yapmaktadır. İkinci grubu, birinci gruptan ayıran özellik her sekans için bir ağırlık katsayısı eklenmektedir.




Şekil 2.3: Skorlama fonksiyonlarının gruplanması.

Aşağıdaki veri seti örneği kullanılarak bu iki grup içerisindeki fonksiyonların elemanları açıklanmıştır. Veri seti içerisinde $L$ uzunluğunda amino asit dizilimine sahip, $N$ farklı protein sekansı bulunmaktadır.

| **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** |

Karlin fonksiyonu, $f_{karlin}(x)$, birinci grup içerisinde bulunan bir fonksiyondur (Eşitlik (2.3)). Skorlama matrisi, $m$, Eşitlik (2,2)'deki $M$ hesabı yapılarak karlin modeline entegre edilir. $s_i(x)$, $i$ sekansındaki $x$ pozisyonunda bulunan amino asidi ifade etmektedir. Aynı şekilde, $s_j(x)$, $j$ sekansındaki $x$ pozisyonunda bulunan amino asidi ifade etmektedir. Toplam formülü ile veri setindeki her bir sekans diğer bir sekans ile karşılaştırılarak mutasyon skoru hesaplanmaktadır. $x$ kolonundaki her amino asit çifti için hesaplama yaparak o sonuçların toplamını mutasyon skoru olarak ifade eder. $\frac{2}{N(N-1)}$ ile çarparak modeli normalize eder. Örneğin; bir kolondaki tüm aminoasitler eş ise sonuç 1 çıkar.

$$M(a,b) = \frac{m(a,b)}{\sqrt{m(a,b)m(a,b)}} \tag{2.2}$$

$$f_{karlin}(x) = \frac{2}{N(N-1)} \sum_{i}^{N} \sum_{j>i}^{N} M\left(s_i(x), s_j(x)\right) \tag{2.3}$$

İkinci grup fonksiyonlarda, skor hesaplanırken veri setindeki sekanslar arası benzerlikler, bir katsayı ile sisteme dahil edilmektedir. Hesaplanan katsayılar sekansların veri seti içerisindeki ağırlıklarını göstermektedir. İkinci grup fonksiyonlardan biri olan Sander, $f_{sander}(x)$ olarak ifade edilmektedir (Eşitlik (2.7)). Sekansların birbirlerine olan benzerlikleri, $d\left(s_i(x), s_j(x)\right)$ formülü ile mesafe (distance) hesabı yapılarak katsayı hesabına dahil edilir. $i$ ve $j$ birbirinden farklı sekanslar olmak üzere $x$ pozisyonundaki amino asitler için, amino asitler aynı ise 1

değerini alırken farklı amino asitler bulunuyorsa 0 değerini almaktadır (Eşitlik (2.4)). Eşitlik (2.5)'te katsayı hesabının formülü görülmektedir. Eşitlik (2.6)'daki $L$ bir sekansın amino asit sayısını ifade etmektedir. Her iki sekans için hesaplama yapıldıktan sonra, hesaplanan her bir değer $\lambda$ hesabı ile tüm mesafelerin toplam değerine bölünmüştür (Eşitlik (2.5)).

$$d\left(s_i(x), s_j(x)\right) = \begin{cases} 1 & s_i(x) = s_j(x) \\ 0 & s_i(x) \neq s_j(x) \end{cases} \tag{2.4}$$

$$w_{ij} = 1 - \frac{1}{L}\sum_{x}^{L} d(s_i(x), s_j(x)) \tag{2.5}$$

$$\lambda = \left( \sum_{i}^{N} \sum_{j>i}^{N} d\left(s_i(x), s_j(x)\right) \right)^{-1} \tag{2.6}$$

$$f_{sander}(x) = \lambda \sum_{i}^{N} \sum_{j>i}^{N} w_{ij} m\left(s_i(x), s_j(x)\right) \tag{2.7}$$

Valdar fonksiyonunda, $f_{valdar}(x)$, Eşitlik (2.11), mesafe hesabı birim matris kullanılarak değil skorlama matrisi kullanılarak hesaplanmıştır (Eşitlik (2.8)). Sander'de olduğu gibi $d(s_i(x), s_j(x))$, mesafe hesabı için skorlama matrisi sonuçları, $M(s_i(x), s_j(x))$, total amino asit dizilimi sayısına, $L$, bölünmüştür (Eşitlik (2.9)). Sander fonksiyonunda karşılaştırılan iki sekans için ortak bir değer hesaplanırken, Valdar fonksiyonunda katsayı, $w_i$, bulunurken her bir sekans için tek tek hesap yapılmaktadır. Her bir sekansın diğer sekanslara göre mesafesi hesaplanmış ve sonunda $N-1$ değerine bölünmüştür (Eşitlik (2.10)).

$$M\left(a,b\right) = \begin{cases} \dfrac{m\left(a,b\right) - \min\left(m\right)}{\max\left(m\right) - \min\left(m\right)} & a \neq boşluk,\ b \neq boşluk \\ 0 & a = boşluk,\ b = boşluk \end{cases} \tag{2.8}$$

$$d(s_i(x), s_j(x)) = 1 - \frac{1}{L}\sum_{x}^{L} M(s_i(x), s_j(x)) \tag{2.9}$$

$$w_i = \frac{1}{N-1}\sum_{i \neq j}^{N} d(s_i(x), s_j(x)) \tag{2.10}$$

$$f_{valdar}(x) = \sum_{i}^{N} \sum_{j>i}^{N} \frac{w_i w_j M\left(s_i(x), s_j(x)\right)}{\sum_{i}^{N} \sum_{j>i}^{N} w_i w_j} \tag{2.11}$$

Birinci gruba dahil olan diğer 3 fonksiyon, sırasıyla $f_{BNT2}(x)$, $f_{BNT3}(x)$ ve $f_{hogervorst}(x)$, bu proje içerisinde oluşturulmuştur. Eşitlik (2.12)-Eşitlik (2.14)'te bulunan $m\left(s_i(x), s_j(x)\right)$, diğer fonksiyonlarda olduğu gibi skorlama matrisindeki skoru ifade etmektedir. $x$ pozisyonunda bulunan amino asitler sırasıyla $j > i$ olmak üzere her $i$ sekansı, $j$ sekansı ile karşılaştırılarak hesaplamalar yapılmaktadır. Hesaplamalar sonucunda, $a$, seçilen fonksiyon olmak üzere, her fonksiyon sonucu, $f_a{}^{norm}(x)$, Eşitlik (2.15)'e göre normalize edilmektedir.

$$f_{BNT2}(x) = \sum_{i}^{N} \sum_{j>i}^{N} \left[ \frac{2m\left(s_i(x), s_j(x)\right)}{m\left(s_i(x), s_i(x)\right) + m\left(s_j(x), s_j(x)\right)} \right] \tag{2.12}$$

$$f_{BNT3}(x) = \sum_{i}^{N} \sum_{j>i}^{N} \left[ \frac{\left| m\left(s_i(x), s_j(x)\right) \right| m\left(s_i(x), s_j(x)\right)}{m\left(s_i(x), s_i(x)\right) m\left(s_j(x), s_j(x)\right)} \right] \tag{2.13}$$

$$f_{hogervorst}(x) = \sum_{i}^{N} \sum_{j>i}^{N} \left[ \frac{m\left(s_i(x), s_j(x)\right)\left[ m\left(s_i(x), s_i(x)\right) + m\left(s_j(x), s_j(x)\right) \right]}{2m\left(s_i(x), s_i(x)\right) m\left(s_j(x), s_j(x)\right)} \right] \tag{2.14}$$

$$f_a{}^{norm}(x) = \frac{f_a(x) - \min(f_a)}{\max(f_a) - \min(f_a)} \tag{2.15}$$

Yukarıdaki skorlama fonksiyonları kullanılarak NA proteininde hassas ve korunan amino asit bölgeleri mutasyon skorları hesaplanarak tespit edilmiştir. Her bir mutasyon fonksiyonu sonucu olan mutasyon skorlarının haritaları çıkarılmıştır. Bu sonuçlar Bölüm 3'te verilmektedir.

### 2.3.2 Metotlar arası korelasyon incelemesi

Korelasyon haritaları oluşturulurken Eşitlik (2.19)'de $Corr(A, B)$ formülasyonu kullanılarak elde edilmiştir. Bu eşitlik kovaryasyon hesabının standart sapma hesaplarına bölümüyle elde edilmektedir. Kovaryasyon hesabı Eşitlik (2.18), standart sapma hesabı Eşitlik (2.17) ve ortalama hesabı Eşitlik (2.16)'da görülmektedir.

$$mean(\mathrm{A}) = \frac{x_1 + x_2 + \ldots + x_N}{N} \tag{2.16}$$

$$stdev(\mathrm{A}) = \left(\frac{1}{m \cdot n - 1}\right) \cdot \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left(A_{i,j} - mean(\mathrm{A})\right)^2 \tag{2.17}$$

$$C\,\mathrm{var}\left(A, B\right) = \frac{1}{m \cdot n} \cdot \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left(\left(A_{i,j} - mean(A)\right) \cdot \left(B_{i,j} - mean(B)\right)\right) \tag{2.18}$$

$$Corr\left(A, B\right) = \frac{C\,\mathrm{var}\left(A, B\right)}{\left(stdev\left(A\right) \cdot stdev\left(B\right)\right)} \tag{2.19}$$

## 2.4 Proteinlerin Evrimsel Değişimlerinin Tahmini

NA proteininde meydana gelebilecek tehlikeli mutasyonların bilinmemesi ileride oluşabilecek pandemik ve endemik salgınların öngörülememesine sebep olmaktadır. Bu salgınlar meydana gelmeden önlem almak büyük bir önem taşımaktadır. İleride oluşabilecek mutasyonların tahmin edilmesi ile tehlikeli salgınlara karşı önlem alınabilir. Bu bölümde elde edilen mutasyon haritalarından, amino asit frekanslarından ve NA proteininin mutasyon hızından yararlanarak ileride oluşabilecek mutasyonlar tahmin edilmiştir. Doğada proteinler üzerinde meydana gelen değişimler rastsal olarak meydana gelmektedir. Bu sebeple rastgele yürüyüş metodu kullanılarak NA proteininde meydana gelebilecek mutasyonlar tahmin edilmiştir. Kullanılan yöntemin ve parametrelerin detaylı açıklamaları aşağıdaki alt bölümlerde bulunmaktadır.

### 2.4.1 Nöraminidaz proteininin mutasyon hızının hesaplanması

Virüslerin mutasyon hızları, virüslerin evrimini anlamak için gerekli önemli bir veridir. Mutasyon hızı, genetik bilgideki bir değişikliğin sonraki jenerasyonlara geçme olasılığını ifade etmektedir. Mutasyon hızı, yeni konak hücreye adapte olma, yeni yayılım yolları bulma ve immün ataklardan kaçma hızına da etki etmektedir. Eğer bir virüsün mutasyon hızı yüksek ise yukarıda bahsedilen parametre hızları da yüksek olacaktır. Böyle bir durumda yeni oluşan virüsler ölümcül ya da tehlikeli etkilere sebep olabilmektedir (Sanjuan, Nebot, Chirico, Mansky, & Belshaw, 2010);(Sanjuán & Domingo-Calap, 2016).

RNA genetik materyalini taşıyan virüsler yüksek genetik değişkenliğe sahiptir. Yüksek oranda değişim gösteren RNA virüsleri değişen ortamlara hızlı bir şekilde adapte olabilir ve böylece ilaca direnç göstermelerine veya bağışıklık sisteminden kaçmalarına yardımcı olur. RNA taşıyan influenza virüsünün hücresel enfeksiyon (7 saat) ve kopyalanmış nükleotid başına ortalama $2.5\times10^{-5}$ mutasyon hızına sahip olduğu deneysel çalışmalar sonucu bilinmektedir (Sanjuán & Domingo-Calap, 2016). Bu bilgiden yararlanılarak yıllık mutasyon hız hesabı referans alınan makale içerişimde verilmiş. Ancak orda yapılan hesaplamaya göre olasılık değeri yıl sayısı arttıkça 1 değerinin üzerine çıkmaktadır. Bu nedenle doğru sonuçlar elde edebilmek için deneysel ölçüm verisinden yaralanarak mutasyon hız hesabı için aşağıdaki eşitlikler oluşturulmuştur. Ayrıca influenza genomunda meydana gelen mutasyonların sadece %10'luk kısmı nöraminidaz kısmında gerçekleşmektedir (Visher vd., 2016) Eşitlik (2.20)'de NA proteinini kodlayan nükleotid bölgesinin 7 saatlik hücresel enfeksiyon siklusunda sahip olduğu mutasyon hızı hesaplanmaktadır. 7 saatlik deneysel mutasyon hız hesabından yılda nükleotid başına düşen mutasyon hızı Eşitlik (2.21) ile hesaplanmıştır. Nükleotidden protein hızını hesaplamak için Eşitlik (2.22) oluşturulmuş ve 3 nükleotid 1 kodon (1 amino asit) denkliğinden yararlanılmıştır. Bu bilgiler kullanılarak nöraminidaz proteininin yılda ne kadar mutasyona uğradığı aşağıdaki Eşitlik (2.23) oluşturularak hesaplanmıştır.

$$H_{nuc'} = 2.5\times10^{-5}\cdot 0.1 \tag{2.20}$$

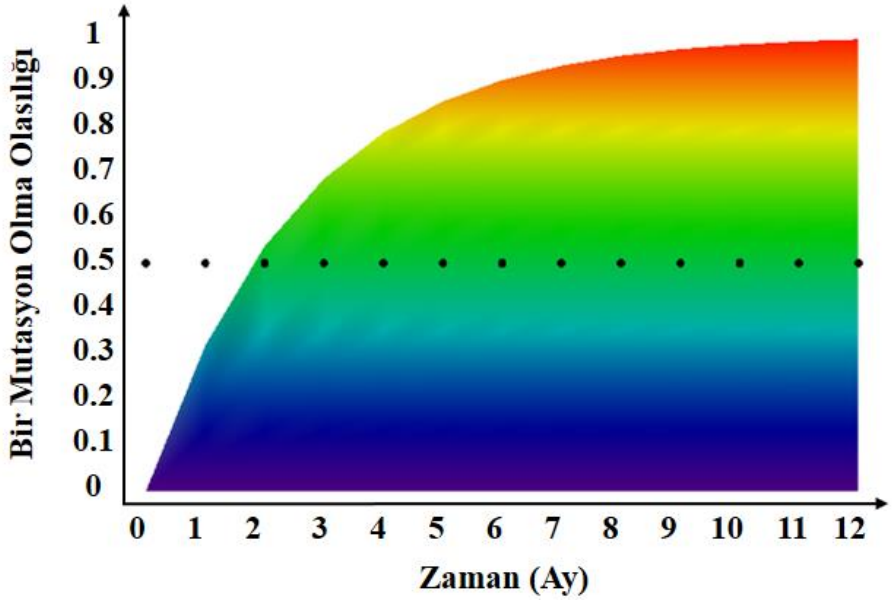
$$H_{nuc} = \frac{H_{nuc'}\cdot 24\cdot 365}{7} \tag{2.21}$$

$$H_{aa} = 1-(1-H_{nuc})^3 \tag{2.22}$$

$$H_{pr} = 1-\left[1-\left(H_{aa}\right)^z\right]^l \tag{2.23}$$

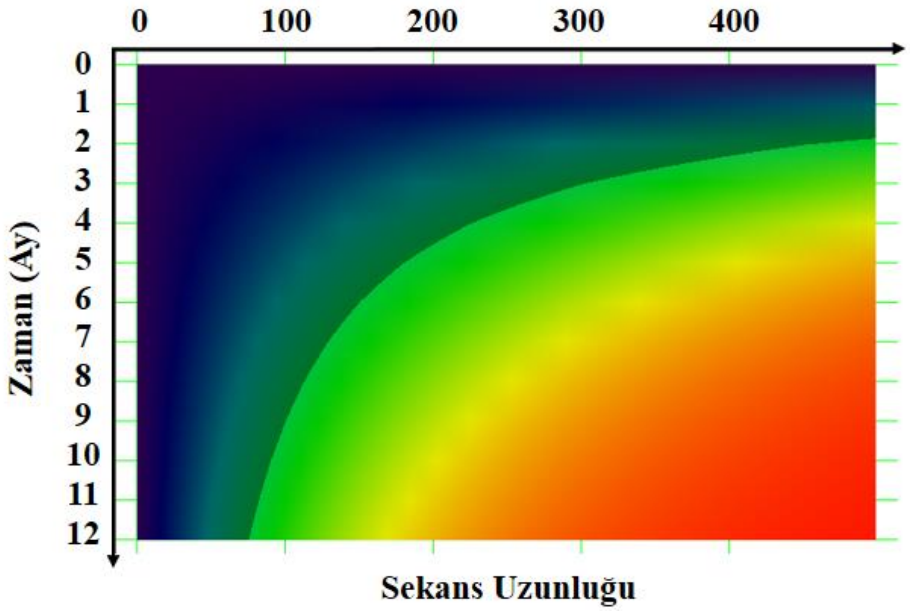
Eşitlikte bulunan $l$, amino asit uzunluğunu, $z$, zamanı ifade etmektedir. Hesaplamalar sonucunda, 388 amino asit dizilimine sahip nöraminidaz proteininin yılda yaklaşık bir kere mutasyona uğrama olasılığına sahip olduğu bulunmuştur (Şekil 2.4).

Bu bilgi skorlama fonksiyonlarına bir koşul olarak eklenmiştir. Bu koşul şu şekilde işlemektedir: Mutasyon skoru hesabında sekansların yıl bilgileri karşılaştırılarak, ancak sekanslar arasında 0-1 yıl fark olması durumunda hesap yapılmış ve ona göre mutasyon haritaları çıkarılmıştır.




Şekil 2.4: Nöraminidaz proteininin mutasyona uğrama olasılığı.

Eşitlik (2.23)'te görüldüğü üzere bir proteinin mutasyona uğrama olasılığı sekans uzunluğuna bağlıdır. Amino asit değişim hızı aynı olan farklı sekans uzunluklarındaki proteinler içerisinde uzun sekansa sahip olanlar kısa olanlara göre daha çok mutasyona uğrar (Şekil 2.5).




Şekil 2.5: Sekans uzunluğu ve zaman değişiminin mutasyon olasılığına etkisi.

### 2.4.2 Rastgele yürüyüş metodu

Rastgele yürüyüş, bir nesnenin bir doğru, yüzey veya hacim içerisinde her bir adım birbirinden bağımsız olması koşuluyla meydana gelen rastgele hareketidir. Rastgele yürüyüş, olasılık teorisinde en çok çalışılan konulardan biridir (Casella, Fienberg, & Olkin, 2011).

Sistem içerisinde sistemi etkileyen parametreler bir toplam vektörü olarak ifade edilmektedir. Bu toplam vektörü, ana vektördeki değerlerin art arda toplanmasıyla oluşturulan stokastik süreçleri ifade eden bir vektördür. Rastgele yürüyüşün her bir adımı, sıfır ile toplam vektörün maksimum değeri arasında denk gelen rastsal sayıya göre değişir. Toplam vektörüne örnek bir eşitlik aşağıda verilmiştir. Atılan rastgele bir adımın denk gelebileceği değer aralığı, Aralık olarak tanımlanan eşitlikte görülmektedir (Eşitlik (2.24)). Atılan rastgele bir adımın, değer aralığı fazla olan pozisyona gelme olasılığı fazladır.

$$Toplam = \begin{bmatrix} x_1 \\ x_1 + x_2 \\ x_1 + x_2 + x_3 \\ \vdots \\ x_1 + x_2 + \ldots + x_N \end{bmatrix} \qquad Aralık = \begin{bmatrix} 0 \le A < T_1 \\ T_1 \le A < T_2 \\ T_2 \le A < T_3 \\ \vdots \\ T_{N-1} \le A < T_N \end{bmatrix} \tag{2.24}$$

İnfluenza virüsünün evrimsel değişiminin tahmini için iki ayrı değişken vektörü kullanılarak rastgele yürüyüş metodu kullanılmıştır. Skorlama fonksiyonları kullanılarak elde edilen mutasyon skorları için bir toplam vektörü ve amino asit frekansları için ayrı bir toplam vektörü oluşturulmuştur. Amino asitlerin frekansları için Şekil 2.6'daki sıralama kullanılmıştır.

**Amino Asitler**

| A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |

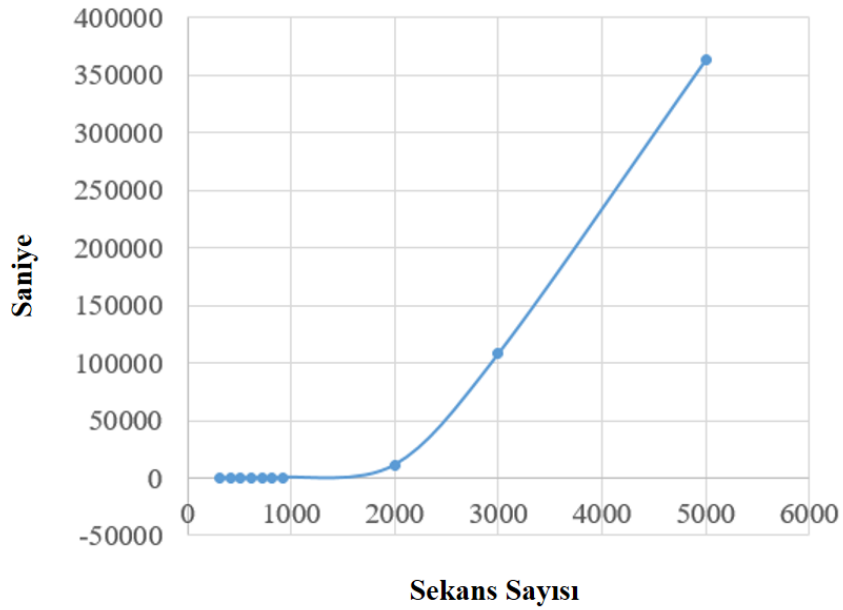
Şekil 2.6: Tahmin modeli için kullanılan amino asit sıralaması.

İlk oluşturulan skor vektörü, mutasyona uğrayacak pozisyonu belirlerken, frekans vektörü tahmin edilen pozisyonun hangi amino aside dönüşeceğini belirlemektedir. Örneğin; 100 amino asit sekans uzunluğuna sahip bir protein için, eğer rastsal olarak

atanan değer, ilk vektörün (mutasyon skorları) değer aralıklarından ikincisine denk gelmiş ise sekansın ikinci pozisyonunda bir değişiklik olacağı anlamına gelmektedir. Atanan ikinci rastsal sayı, ikinci vektörün değer aralıklarından birincisine denk gelmiş ise sekansın ikinci pozisyonundaki amino asit Alanin (A) amino asidine dönüşecektir. Tahmin modeli genel olarak bu şekilde ilerlemektedir.

Oluşturulan bu model içerisinde işlemler birbirinden bağımsız bir şekilde gerçekleştiği için belli periyotlarda rastgele yürüyüş tekrarlanarak olasılıkların yakınsayan sonuçları kullanılmıştır. Rastgele yürüyüş sisteminde atılan bir adımlarda meydana gelecek mutasyon sayısı hesaplanan mutasyon hızı kullanılarak kararlaştırılmıştır. Bir yılda bir mutasyon olma olasılığı olduğu Bölüm 2.4.1'de hesaplanmıştır, böylece tahmin modelinde bir adım atıldığında bir mutasyon gerçekleşeceği anlamına gelmektedir.

Tahmin etmek istenen süreç bir yıldan fazla ise sistem şu şekilde ilerlemektedir: İlk olarak tahmin edilen ilk yıl için periyot sayısı seçilmiştir. Periyot seçimi yapılırken 300 sekans ile 5000 sekans tahmini için geçen süre aralıklarına bakılarak hesaplama süresi en uygun olan 2000 sekanslık periyotlarda tahmin yapılmıştır. Şekil 2.7'te 2000'lik sekans tahmini yaklaşık 3 saat sürerken periyot sayısı 3000'e çıkarıldığında bu süreç 30 saate çıkmaktadır. 5000 olduğunda ise yaklaşık 4 gün sürmektedir. Bu nedenle her atılan adımda (tahmin yılında) 2000 tane sekans tahmini yapılmaktadır.

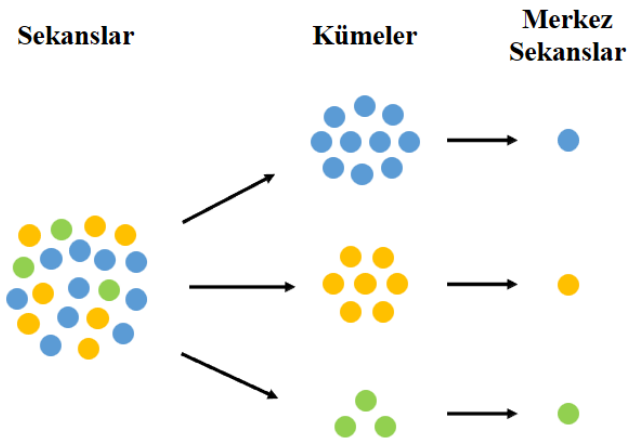



Şekil 2.7: Tahmin modelinin hesaplama süresi ile kullanılan periyotlar arası ilişkisi.

Periyot sayısı kadar tahmin yapılmış ve bu grup içerisinde kümeleme yapılarak küme eleman sayısının en çok olduğu ilk üç kümeden merkez sekans (temsilci sekans)

seçilerek bir sonraki yıl için referans sekans olarak verilmiştir. İkinci yıl için değişim ilk verilen referans sekans üzerinden değil de yeni oluşmuş tahminlerden merkez seçilen sekans üzerinde olmuştur. Bu döngü tahmin edilmek istenen yıl sayısı kadar gerçekleştirilmiştir. Bu sayede sistem tamamen rastsal olarak değil evrimsel sürecin işleyişi şeklinde ilerlemiştir (Şekil 2.9).

Kümeleme, nesnelerin belirli özelliklerine göre farklı kümelere ayrılmasıdır. Farklı kümeleme yöntemleri bulunmaktadır. Bu proje içerisinde hiyerarşik kümeleme yöntemi kullanılmıştır (Şekil 2.8). Hiyerarşik kümeleme, her nesneyi bir kümeye atar ve yinelemeler yapılarak en yüksek uyumu olan küme çiftlerini birleştirir ve kümeler genellikle bir dendrogram olarak gösterilir (Bar-Joseph, Gifford, & Jaakkola, 2001). Kümeleme işlemi için R programlama dili kullanılarak girdi olarak mesafe matrisi (PDS) verilmiştir. Mesafe matrisi, global hizalama yöntemlerinden biri olan Needleman Wunsch yöntemi kullanılarak oluşturulmuştur (Needleman & Wunsch, 1970). Global hizalama yöntemi ile sekansların birbirlerine olan benzerlikleri hesaplanarak benzerlik matrisi elde edilir. Elde edilen hizalama skoru ne kadar yüksek ise sekanslar birbirine o kadar benzer anlamına gelmektedir. Bu değeri sekanslar arası oluşan mesafe olarak tanımlamak için Eşitlik 2.25 oluşturulmuştur. Mesafe, benzerlik değerinin tam tersini ifade etmelidir. Birbirlerine en benzer sekansların mesafe skorları, diğer sekanslar arası skorlara göre daha düşük olmalıdır. Mesafe değeri az olan sekanslar birbirlerine daha benzerdir.




Şekil 2.8: Kümeleme işlem basamakları.

Kümele yapılırken sisteme bir eşik değeri verilmiştir, bu eşik değeri altında mesafeye sahip olan sekanslar, bir grup içerisine girmiştir. Eşik değeri, grup içerisinde merkez sekansın yakınsama durumuna göre belirlenmiştir. Örneğin, eşik değerinin

değişmesiyle bir küme içerisinde eleman sayısı değişmeye devam etse bile merkez sekans değişmiyorsa sistem yakınsamaya başlamıştır, bu bilgiden yararlanılarak eşik değeri seçilmiştir.

$$PDS_{i,j} = \frac{1}{PSS_{i,j}} \tag{2.25}$$

### 2.4.3 Tahmin performanslarının incelenmesi

Yapılan tahminler sonucu elde edilen sekansların doğruluğunun analiz edilmesi gerekmektedir. İleride karşılaşılabilecek sekanslar şu an için var olmadıklarından, yapılan tahminler ile olan tutarlılığının anlaşılması mümkün değildir. Bu nedenle şu an var olan sekanslar üzerinden performans analizi yapılmalıdır. Analiz için ilk olarak veri seti içerisinde eğitim ve doğrulama setleri oluşturulmuştur. Eğitim seti, tahmin için gerekli olan bilgiyi sunan, bir başka deyişle, gelecekte oluşabilecek veriyi tahmin etmesi için eğitilen veri setidir. Doğrulama seti ise, gerçekte gözlemlenmiş olan sekans seti, yani, tahmin sonuçlarıyla varılmak istenen veri setidir.

#### 2.4.3.1 Toplam benzerlik skoru hesabı

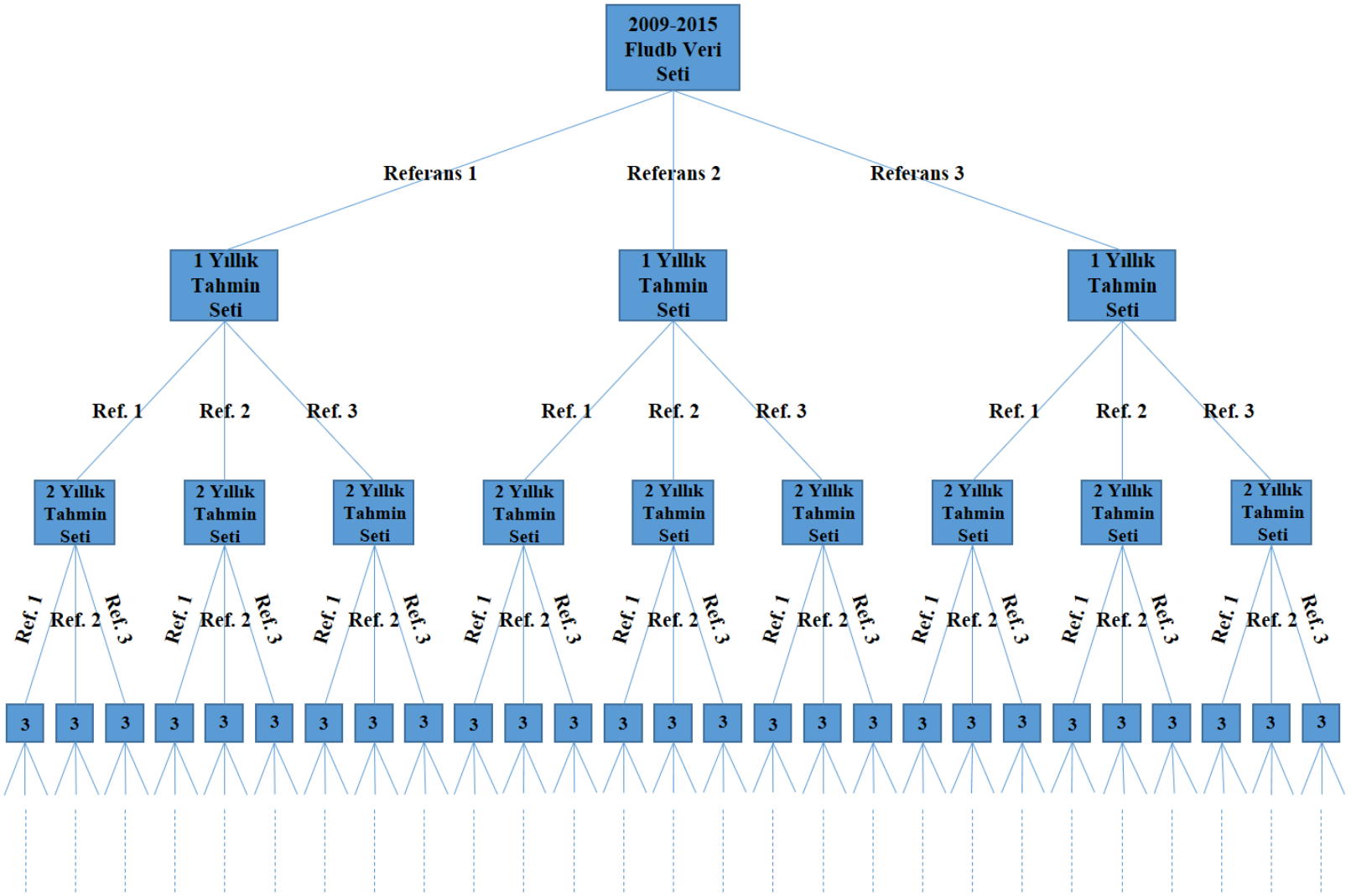
Tahminlerin, doğrulama seti ile olan benzerliklerinin hesaplanması için Needleman-Wunch yöntemi kullanılmıştır. $N_A$, eleman sayısına sahip doğrulama seti (A) ile $N_B$, elemanlı tahmin seti (B) içerisindeki her sekans birbirleriyle ikili olarak karşılaştırılıp toplanarak $PSS$, benzerlik skoru elde edilmiştir. Bu skor daha sonra eleman sayılarına bölünerek normalize edilmiştir. Normalize olan skor, toplam benzerlik skorunu vermektedir (Eşitlik 2.26) (Oren vd., 2007).

$$TSS_{A,B} = \frac{1}{N_A N_B} \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} PSS_{i,j} \tag{2.26}$$

Toplam benzerlik skoru ne kadar yüksek ise setler arası benzerliğin o kadar yüksek olduğu anlamına gelmektedir.

#### 2.4.3.2 Pozisyona bağlı ortalama amino asit farkı hesabı

$A$, doğrulama seti, $B$, tahmin seti olmak üzere, $A$ setindeki her $i$ sekansı, $B$ setindeki her $j$ sekansı ile karşılaştırılarak, $V_{i,j}$, farklılık matrisi elde edilmektedir (Eşitlik 2.28). Farklılık matrisinin elde edilmesi için Eşitlik 2.27 kullanılmıştır.



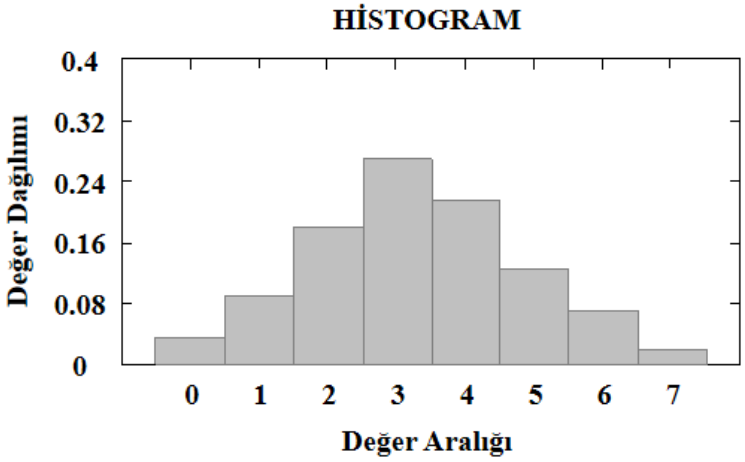

Şekil 2.9: Tahmin yılına göre tahmin akış şeması

Her $x$ pozisyonunda bulunan amino asit $i$ ve $j$ sekansı arasında karşılaştırılarak eşleşenler için 1 eşleşmeyenler için 0 değerini almaktadır. Farklılık matrisi için elde edilen eşleşme matrisi 1'den çıkarılarak farklılık matrisi oluşturulmaktadır.

$$d\left(s_i(x), s_j(x)\right) = \begin{cases} 1 & s_i(x) = s_j(x) \\ 0 & s_i(x) \neq s_j(x) \end{cases} \tag{2.27}$$

$$V_{i,j} = \sum_{x=1}^{L} 1 - d\left(A_i(x), B_j(x)\right) \tag{2.28}$$

Oluşan bu matriste her bir kolon için histogram oluşturulmuştur. Oluşan her bir histogram toplanarak eleman sayılarına, bölünerek normalize edilmiştir (Eşitlik 2.29). Bu normalizasyon ile oluşan her bir histogramın alanı 1 olmuştur (Eşitlik 2.30). Histogramlar arası karşılaştırma yapabilmek için histogramların ortalaması alınmıştır. Pozisyona bağlı ortalama amino asit farkı hesabı için histogramdaki toplam $DA$, değer aralığındaki, her $i$ değer aralığı, değerlerin dağılımı $Hist_i$ ile çarpılarak toplanmıştır (Eşitlik 2.31). Örnek bir histogram Şekil 2.10'da bulunmaktadır.




Şekil 2.10: Histogram örneği.

$$Hist = \frac{\sum_{j=1}^{N_B} Histogram(V^{\langle j \rangle})}{N_A N_B} \tag{2.29}$$

$$\sum Hist = 1 \tag{2.30}$$

$$TH = \sum_{i=0}^{DA} i \cdot Hist_i \tag{2.31}$$

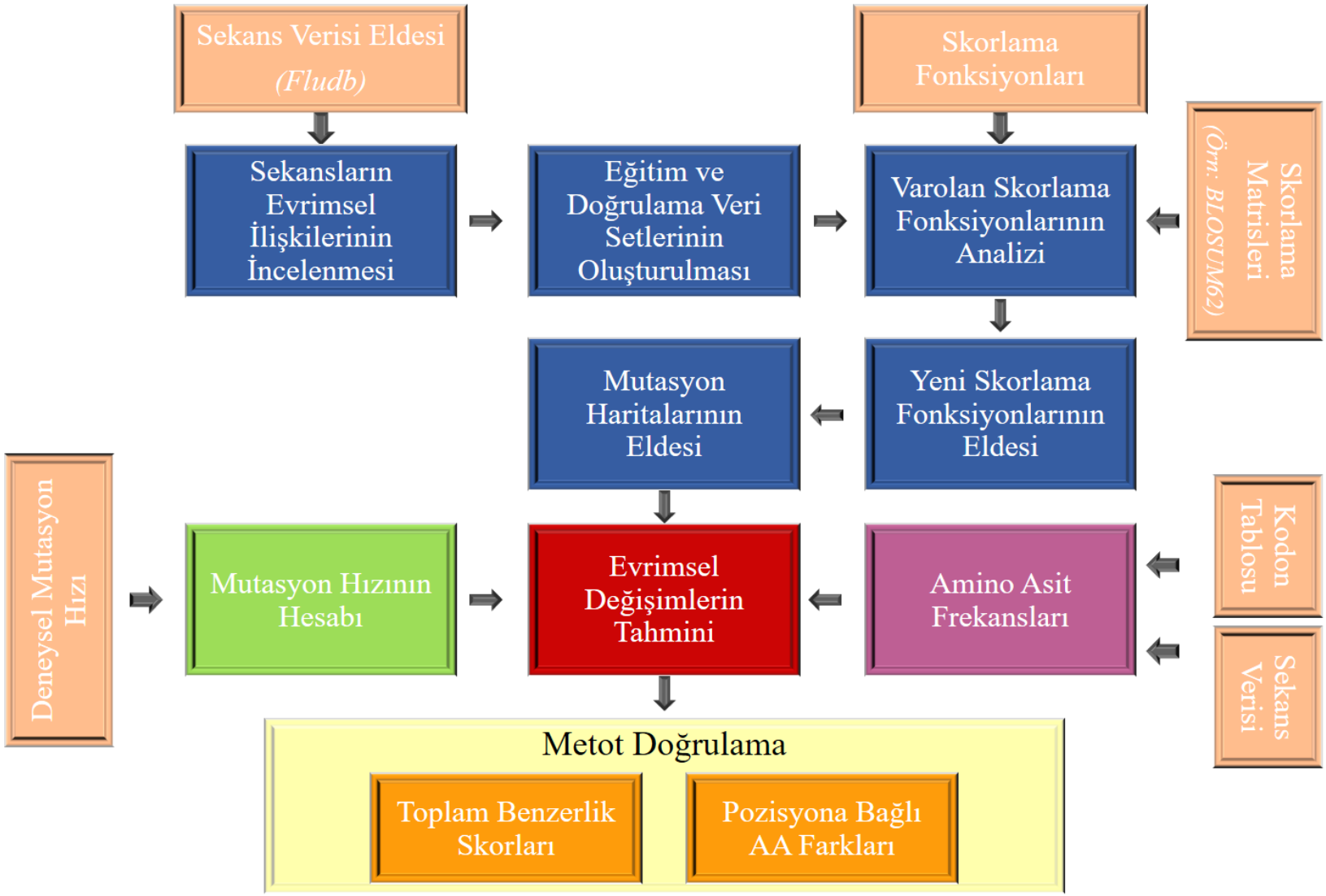
# 3. MODELLEME SONUÇLARI VE YORUMLAR

Matematiksel modelleme ve sayısal yöntemler bölümünde bahsedilen metotlar kullanılarak, ilk olarak NA proteini sekans verisi detaylı olarak incelenmiş, filogenetik ağaç oluşturulmuştur. Bir nevi evrim ağacı olarak da adlandırabileceğimiz filogenetik ağaçtan yararlanılarak, NA proteini gruplara ayrılmıştır. Her bir grup üzerinden mutasyon haritaları ve amino asit frekansları çıkarılmış ve bir önceki bölümde hesaplanan mutasyon hızı bilgisi de kullanılarak tahminler yapılmıştır.

Şekil 3.1'de NA proteinin evrimsel değişiminin tahmin edilmesi için gerçekleştirilen adımlar akış şeması halinde verilmiştir. NA proteini için elde edilen tüm sekanslar incelenmiş ve bu sekanslar aşağıda bahsedildiği şekilde eğitim ve doğrulama setleri olarak gruplara ayrılmıştır. Farklı skorlama fonksiyonları kullanılarak, eğitim setlerindeki sekanslar üzerinde amino asitlerin pozisyona bağlı mutasyona uğrama olasılıkları yani mutasyon haritaları elde edilmiştir. Buna paralel olarak literatürden elde edilen deneysel veriler ile NA proteininin mutasyona uğrama hızı hesaplanmıştır. Bu iki bilgi, ne kadar sürede hangi bölgelerde mutasyon görüleceğini belirtirken, mutasyonun gerçekleşeceği bölgedeki amino asitlerin hangi amino aside dönüşeceği ise amino asit frekansları hesaplanarak bulunmuştur. Sonuç olarak, bu üç bilgi rastgele yürüyüş algoritması içerisinde kullanılarak karşılaşma olasılığı bulunan NA proteinleri tahmin edilmiştir. Son olarak, tahminler sonucu elde edilen verilerin tahmin performansları incelenmiştir.

## 3.1 Nöraminidaz Proteini Veri Setinin Analizi

Evrimsel tahmin metotlarının geliştirilebilmesi için ilk önce girdi olarak verilecek verinin detaylı bir şekilde incelenmesi gerekmektedir. Veri bankalarında bulunan veriler dünya üzerinde görülen tüm influenza virüslerini değil raporlanan virüs sekanslarını içermektedir. Bu nedenle değerlendirme yapılırken dikkat edilmesi gereken bir noktadır.

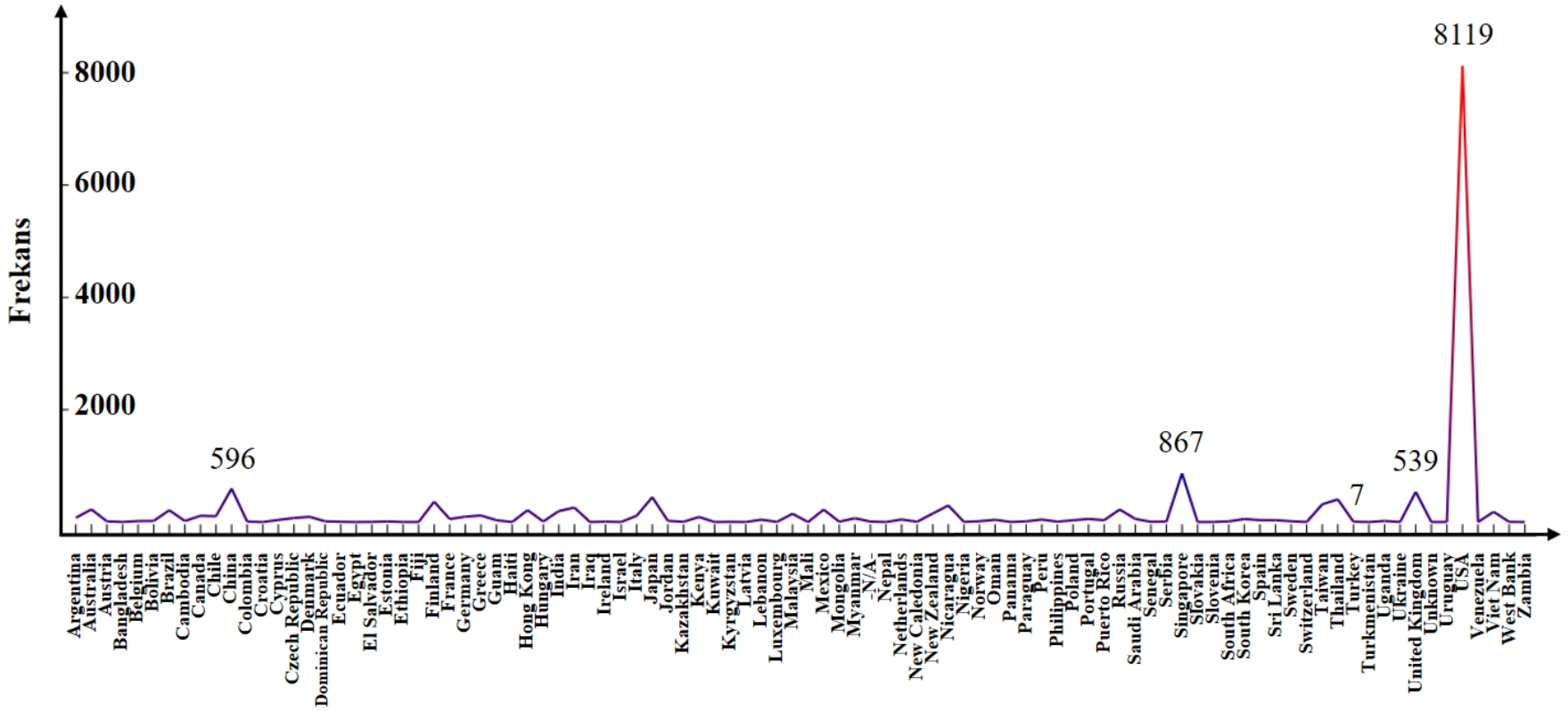



Şekil 3.1: Evrimsel değişimlerin tahmini için oluşturulan akış şeması.

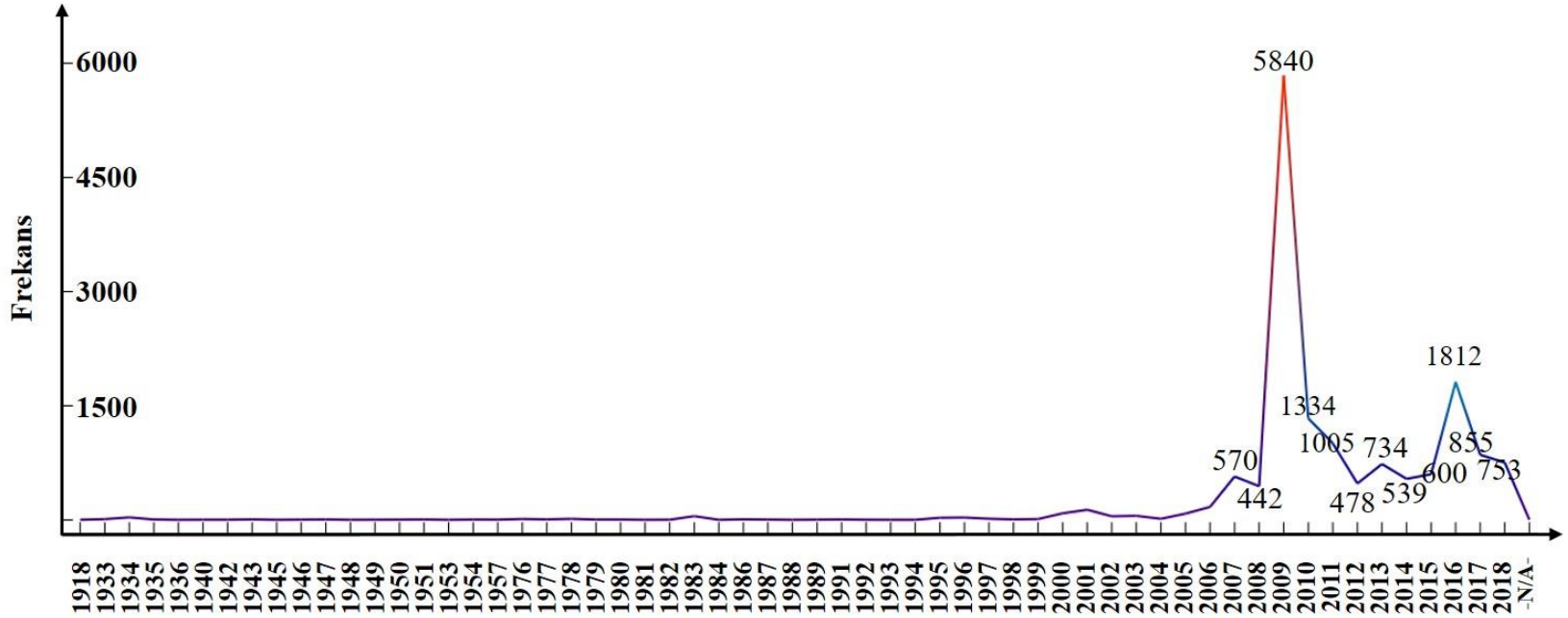
Verilerin toplanmasında etkili olan birçok parametre bulunmaktadır. Örneğin en çok verinin toplanmış olduğu Amerika Birleşik Devletleri'nin ve Singapur'un 2017 yılına ait kişi başına düşen gayrisafi yurt içi hasıla (GSYİH) değeri yaklaşık 60.000 dolardır. Bu nedenle ülkelerin gelişmişlik seviyeleri, gerekli laboratuvar donanımının sağlanmasına olanak vermesinden dolayı toplanabilen veri sayısına etki etmektedir. Veri sayısına etki eden bir başka parametre ise doğal olarak nüfustur. Çin gibi 1.42 milyar nüfusa sahip bir ülkede toplanan veri sayısı fazla olsa da nüfus göz önüne alındığında oldukça düşük olduğu görülmektedir. Etki eden bir başka parametre ise ülkeler arası etkileşimdir (turizm, ticaret). İnsanların seyahat etmesiyle hastalıklar daha kolay yayılmaktadır. Bu duruma örnek olarak Taiwan ve Thailand verilebilir, çünkü bu iki ülkenin gelişmişlik seviyesi Amerika ve Singapora göre daha azdır, ancak seyahatler sebebiyle virüs, bu bölgelere de yayılmıştır. Ayrıca ülkelerin bulundukları konumlar da hastalıkların yayılma oranlarına etki etmektedir. Nikaragua, 2009 salgınının başladığı Meksika'ya yakın bir konumdadır. İnfluenza virüsünün yakın mesafelere yayılma olasılığı daha fazla olduğu için Nikaragua'nın gelişmişlik düzeyi düşük olmasına rağmen toplanmış veri sayısı fazladır (Şekil 3.2).

Sekansların yıllara göre dağılımına bakıldığında (Şekil 3.3) veri toplama ve saklama teknolojilerinin giderek gelişmesiyle birlikte her yıl giderek artan sayıda verinin sisteme eklendiği görülmektedir. 2009 yılı ise tüm zamanlar içerisinde en fazla verinin toplandığı yıl olmuştur. Bunun nedeni ise, 2009 yılında karşılaşılan pandemidir. Modelleme bölümünde bahsedilen eleme metodu kullanıldıktan sonra yıllara göre dağılım tekrar incelenmiştir. Bu eleme yöntemi ile kendini tekrar eden sekanslar elenmiştir. Eleme yapılmasına rağmen en çok farklı sekansın görüldüğü yıl 2009 yılı olmuştur (Şekil 3.4).

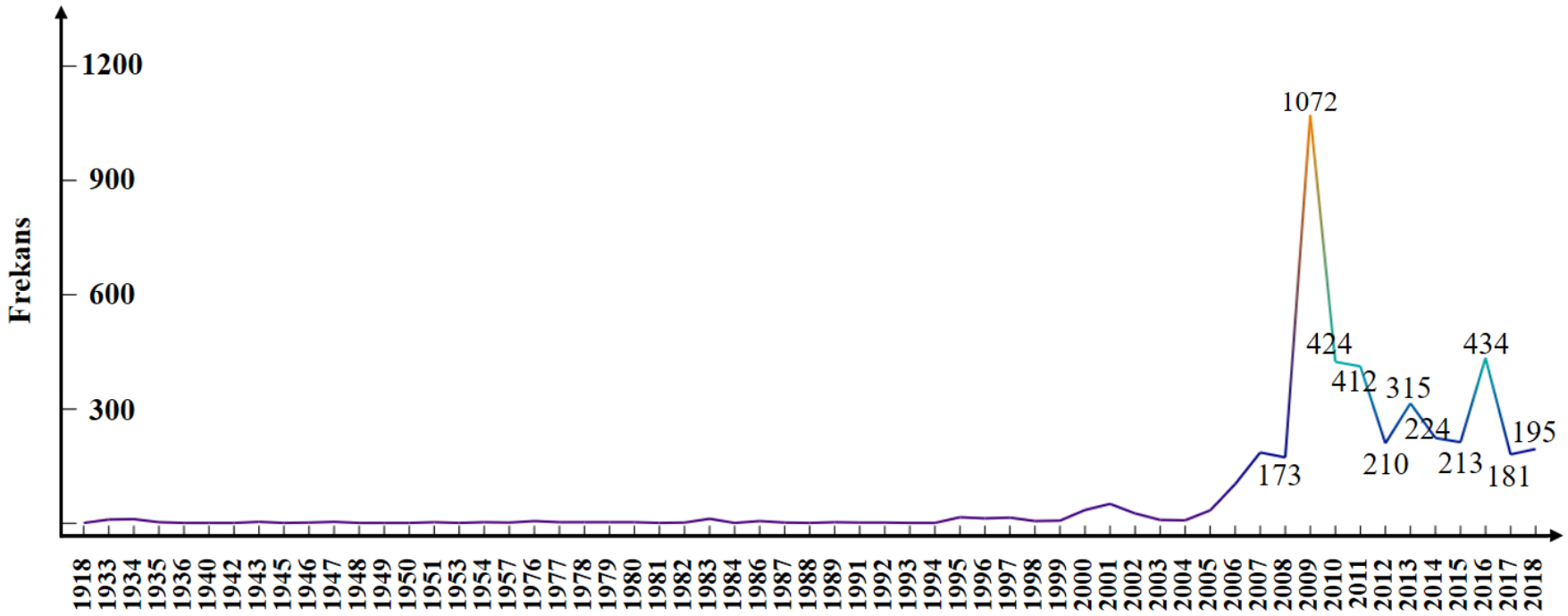
2009 yılında genetik kayma olması sonucu meydana gelen yeni H1N1 virüsü, insanları enfekte ederek, hızlı bir şekilde yayılarak pandemiye neden olmuştur. Oluşan bu yeni virüs, filogenetik ağaçta iki ana grubun oluşmasına sebep olmuştur. Filogenetik ağaçtan yararlanarak veri seti, tahmin modelinin doğru bir şekilde işlemesi için iki parçaya ayrılmıştır. Bu tez kapsamında geliştirdiğimiz yöntem eklemeli mutasyonları tahmin etmek için kullanılacak olup, 2009 pandemisine sebep olan genetik kaymalar için uygun değildir.




Şekil 3.2: 1918-2018 Fludb veri setinin ülke dağılımı.

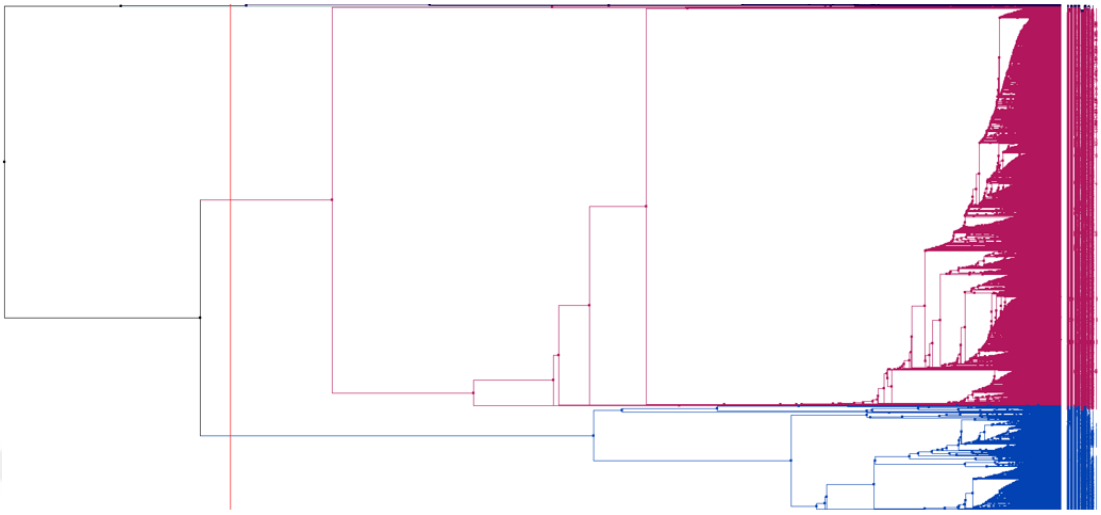



Şekil 3.3: 1918-2018 Fludb veri setinin yıl dağılımı.



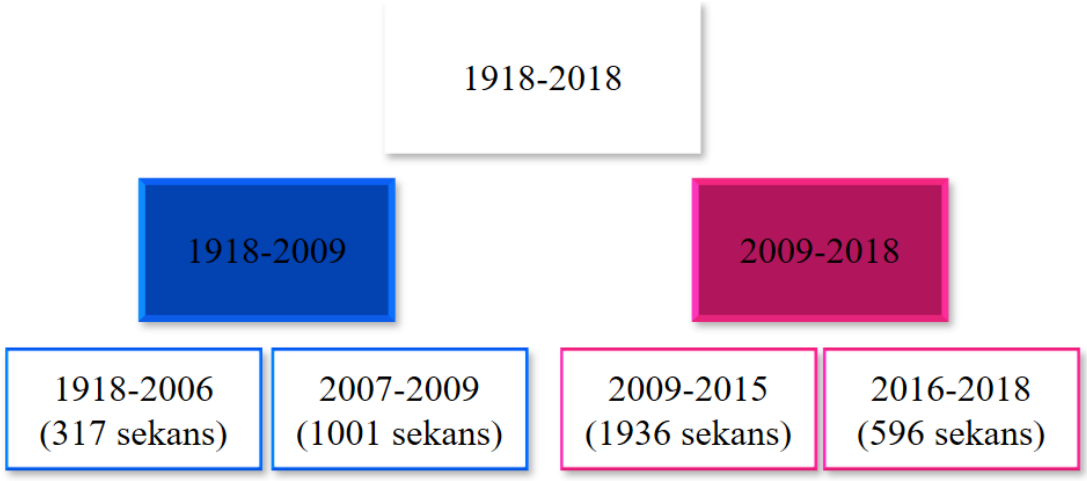

Şekil 3.4: 1918-2018 Fludb veri setinin sadeleştirilmiş yıl dağılımı.

Bu nedenle doğada meydana gelen evrimsel değişimlerin, bir başka deyişle genetik sürüklemelerin, evrimsel etkisini daha doğru bir şekilde inceleyebilmek için sistem iki gruba ayrılmıştır (Şekil 3.5).



Şekil 3.5: 1918-2018 veri setinin filogenetik ağaç üzerinde gruplandırılması. Burada oluşan iki ana dalın 2009 öncesi (mavi) ve sonrası (koyu pembe) olarak ayrıştığı gözlenmiştir.

Şekil 3.6 veri setinin gruplandırılmasını göstermektedir. 1918-2006 ve 2009-2015 veri setleri kullanılarak sırasıyla 2007-2009 ve 2016-2018 yıllarında görülme ihtimali olan yeni sekanslar tahmin edilmiştir.




Şekil 3.6: 1918-2018 Fludb veri setinin gruplandırılması.

Bu iki veri incelendiğinde 2009 pandemiğinin genetik kayma ile (domuz ve insan grip virüsleri arasındaki gen değişimi) oluştuğu görülebilmektedir. Bu nedenle, her iki grup kendi içerisinde alt gruplara ayrılarak eğitim ve doğrulama veri setleri oluşturulmuştur.

Bu durumda eğitim seti verileri 1918-2006 ile 2009-2015 yıllarını; doğrulama seti verileri ise, 2007-2009 ve 2016-2018 yıllarını içeren sekanslardır.

Karşılaştırma kriterleri için ilk olarak ilaçların bağlanma bölgesi olan aktif bölgedeki değişimler ele alınmıştır. Şekil 3.7'de 1918-2006, 2007-2009, 2009-2015 ve 2016-2018 veri setinde bulunan sekansların aktif bölgelerindeki korunan ve değişen amino asitler görülmektedir. 1918-2006 sekanslarında 344. pozisyon daha çok değişime uğrarken, 2009-2015 sekanslarında 275. pozisyon değişime uğramıştır. 2007-2009 yılı için 275. ve 344. pozisyonlar değişime uğramış, 2016-2018 veri seti kendi içinde korunmuştur. Bu iki belirgin değişim dışında genel olarak aktif bölge korunmaktadır.


Şekil 3.7: NA proteininin aktif bölgesindeki amino asitlerin farklı dağılımları.

Aktif bölge dışında tüm veri seti içerisinde tamamen korunan pozisyonlar bulunmaktadır. Çizelge 3.1 bu pozisyonları göstermektedir. Bu bilgiden yararlanarak tahmin metotları uygulandığında çizelgede bulunan bölgelerin değişime uğramaması gerektiği beklenmektedir.

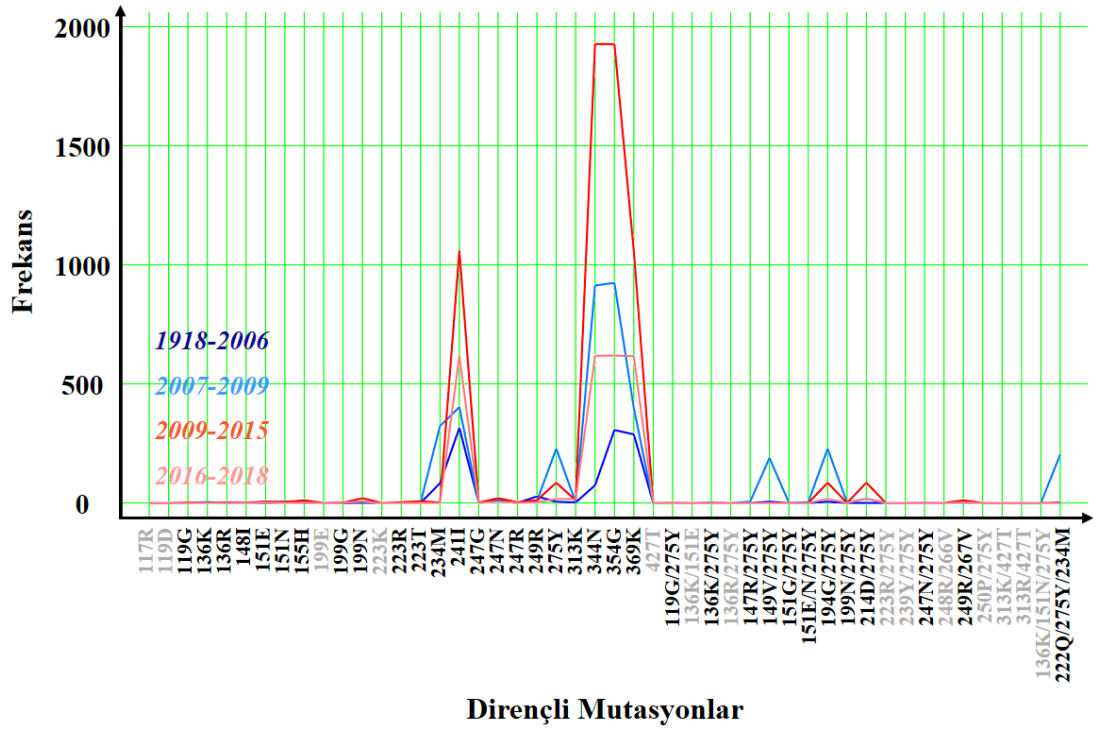
Literatürde dirençli olarak tanımlanan pozisyonların kullanılan veri seti içerisindeki sıklığı ise Şekil 3.8'de gösterilmiştir. Bazı pozisyonlardaki değişimler veri seti içerisinde bulunmamaktadır. Bunun sebebi, yapılan değişimlerin ters genetik yöntemi

ile deneysel olarak oluşturulması ve dirençli olduklarının gözlemlenmesidir. Bu yöntemle oluşturulan mutantlar Şekil 3.8'de bulunmamaktadır.

Çizelge 3.1: NA proteini ana baş kısmında tamamen korunan pozisyonlar.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 92 | CC | 179 | WW | 252 | SS | 302 | PP | 376 | DD | 433 | EE |
| 129 | CC | 187 | GG | 282 | YY | 310 | LL | 379 | GG | 442 | SS |
| 132 | FF | 192 | TT | 292 | CC | 312 | YY | 401 | GG | 456 | WW |
| 137 | GG | 194 | GG | 296 | WW | 324 | DD | 413 | TT | | |
| 161 | CC | 227 | QQ | 300 | NN | 345 | GG | 417 | CC | | |
| 167 | PP | 244 | DD | 301 | RR | 356 | GG | 425 | EE | | |

Gri renkte gösterilen sekanslar ise deneysel ortamda hücresel çoğalma ile gözlemlenmiş olan mutasyonlardır. Bu mutasyonlardan çoğu klinik olarak gözlemlenmediğinden ana veri seti içerisinde de sayıları oldukça azdır. Sonuç olarak 48 dirençli pozisyondan 34'ü ana veri seti içerisinde bulunmaktadır. Klinik ve deneysel çalışmalar ile tespit edilen dirençli pozisyonlar, tez kapsamında yapılan hesaplamalar sonucu bulunan dirençli mutasyonlar ile ileriki bölümlerde karşılaştırılacaktır.




Şekil 3.8: Klinik ve deneysel olarak gözlemlenen dirençli mutasyonların dağılımı.

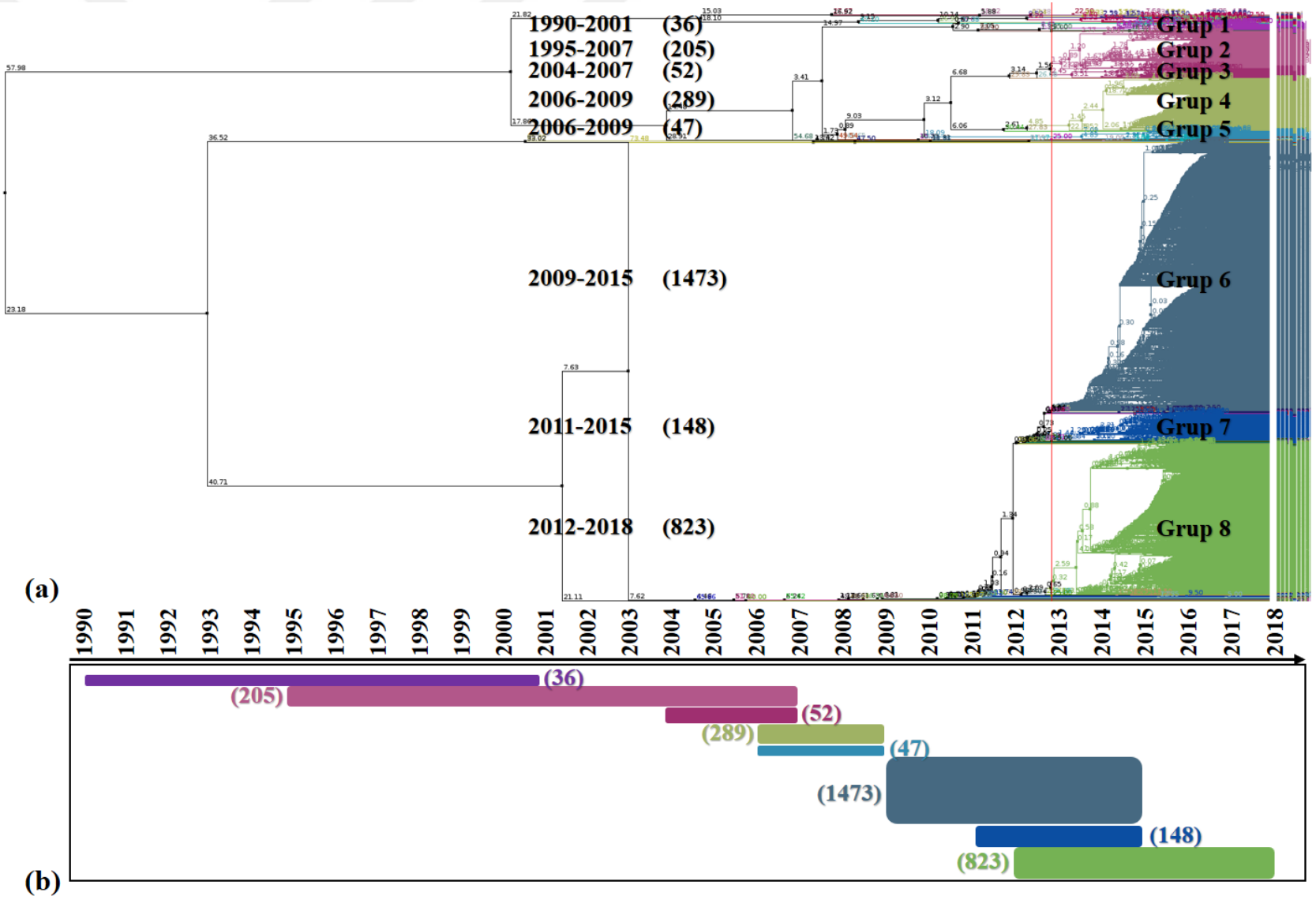
Gruplara ayrılmış veri setleri, kendi içlerinde sekans hizalama yapıldığında bazı pozisyonlara boşluk atanmaktadır. Boşlukların oluşma sebebi sekansta yer alabilecek yeni bir amino asit ya da var olan bir amino asidin yok olmasıdır. Boş olan

pozisyonların analizi bu tez kapsamına dahil edilmemiştir. Bu nedenle boşluk olan pozisyonların dağılımı yıllara göre incelenerek 435. pozisyondaki değişim yıllara göre Çizelge 3.2'de verilmiştir. Bu çizelge incelendiğinde, 435. pozisyonda zamanla bir amino asidin yok olduğu ve dolayısı ile sekans boyunda bir amino asitlik bir kısalma olduğu görülmüştür. Bu nedenle hizalama yapıldığında 435. pozisyona boşluk denk gelmektedir. 2016-2018 verisinde boşluk bulunmadığı için 2009-2015 veri seti içerisinde bulunan boşluk yerine T amino asidi içeren sekanslar çıkarılarak, boşluksuz hizalama olması sağlanmıştır. Bu sayede tahmin modeli modifiye edilerek daha doğru tahminler yapılması hedeflenmiştir.

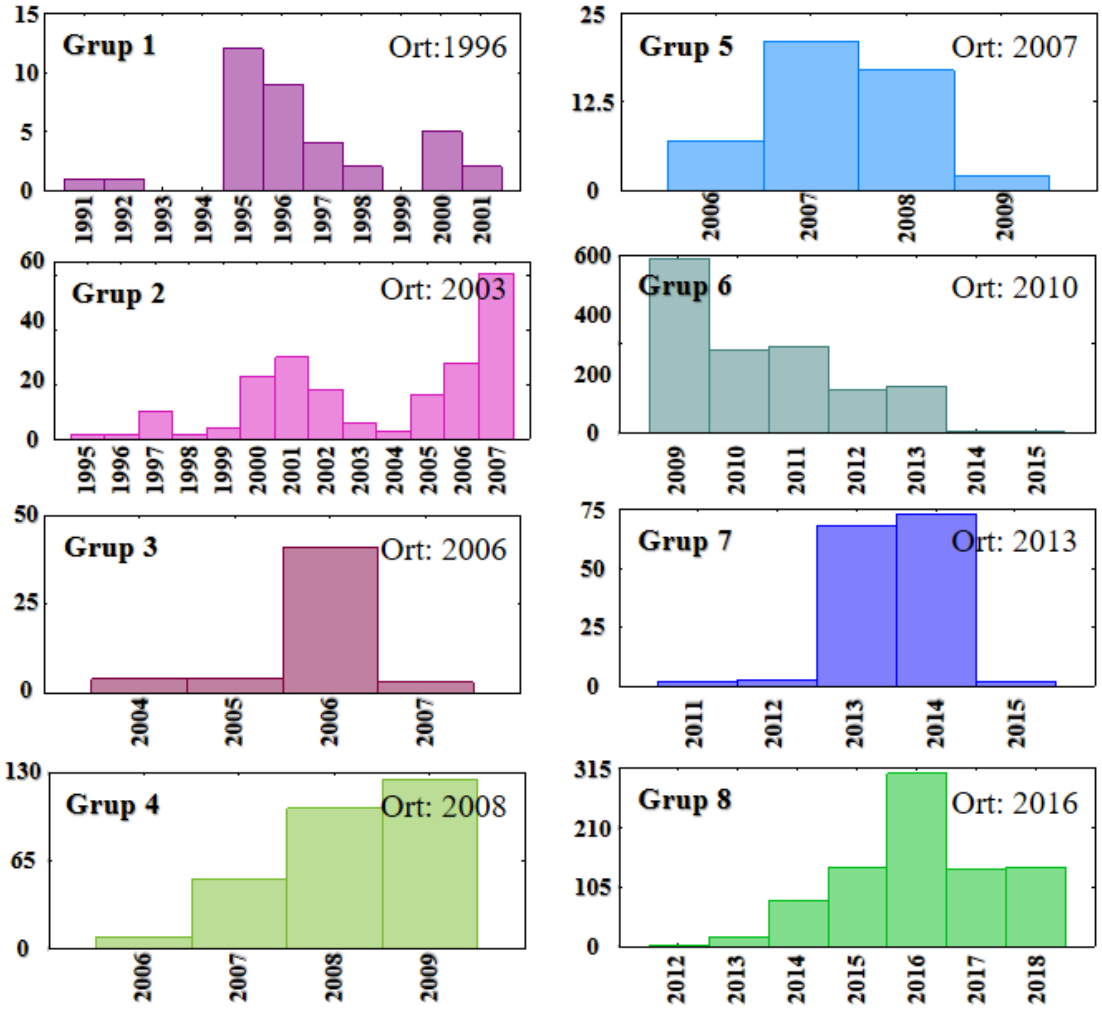
Çizelge 3.2: 435. pozisyonun yıllara göre uğradığı değişim.

| **Yıl** | **435. Pozisyondaki Amino Asit Değişimi** |
|---|---|
| 1918-2006 | T > Boşluk<br>1918-1947 (Boşluk) |
| 2007-2009 | T = Boşluk<br>En çok 2009 yılı içinde |
| 2009-2015 | Boşluk >> T |
| 2016-2018 | Boşluk ve T yok |

Yukarıda bahsedilen şekilde, 4 ana gruba ayrılan veri seti, filogenetik ağaç yardımıyla daha detaylı bir şekilde incelenmiştir (Şekil 3.9). Aralarındaki benzerliğe göre 4 ana grup daha küçük ve birbirleri ile benzer gruplara ayrılmıştır (Şekil 3.9 (a)). Bu durumda, 8 farklı grup oluşmuş ve bu grupların sekans sayısı ve yıl aralığı Şekil 3.9 (b)'de gösterilmiştir. Her bir grup içerisindeki sekanslar, belli yıllarda varlıklarını sürdürürken belli yıllardan sonra yok olmuşlardır. En belirgin gruplar arası farklılaşma Grup 5'ten Grup 6'ya geçerken görülmektedir. Her bir grup belli yıllar arasında beraber var olurken, Grup 5 ve öncesi, 2009 yılından sonra gözlemlenmemiştir. Bunun en büyük sebebi 2009 yılında meydana gelen pandemidir. 2009 yılından sonra farklı gruplar aynı yıllar içinde gözlemlenmiştir. Şekil 3.9 (b)'de gruplarındaki kalınlıklar veri sayısını ifade etmektedir. Örneğin; Grup 6'da 1473 sekans bulurken Grup 3'te 52 tane sekans bulunmaktadır. Bu nedenle Grup 6 kalın bir kutu ile görselleştirilmiştir. Her bir grup içindeki yıl dağılımı ise Şekil 3.10'da gösterilmiştir.




Şekil 3.9: NA sekanslarının gruplandırılması. (a) Gruplamanın filogenetik ağaçta gösterimi. (b) Yıllara göre veri dağılımı şeması.



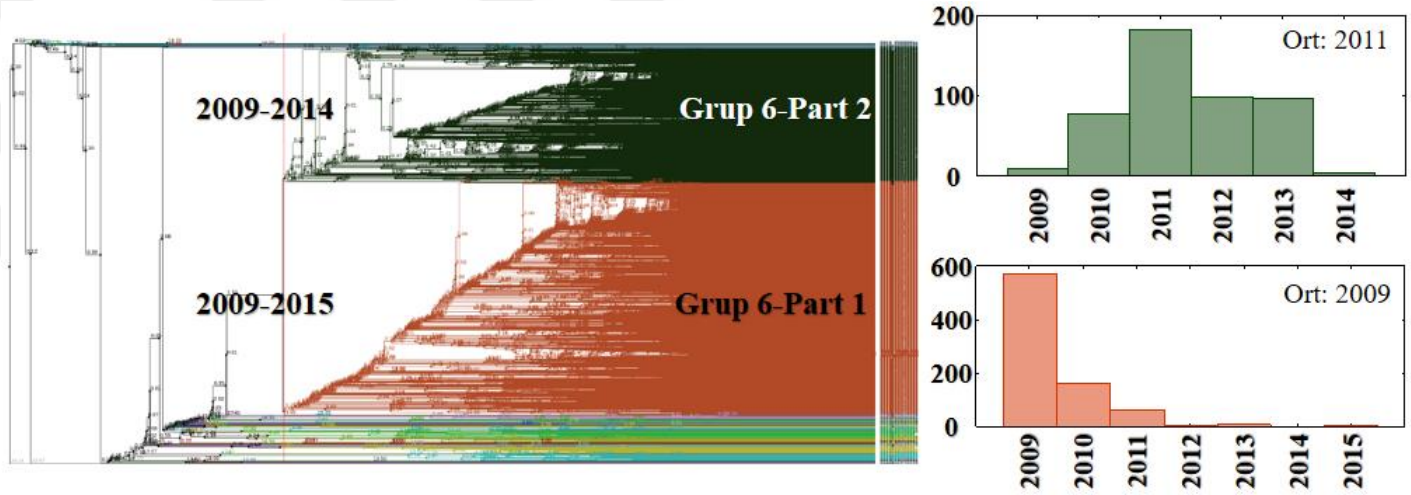

Şekil 3.10: NA proteini sekans gruplarındaki yıl dağılımı.

Veri sayısı ve yıl bilgisi tahmin modeli için önemli parametrelerdir. Veri sayısı büyük olduğunda tahmin sisteminin sahip olacağı bilgi daha fazla olacaktır. Zaman bilgisi de tahmin modelinde kullanıldığı için yıl dağılımı ile tahmin modelinde atılacak adım sayısı bir başka deyişle ilerlenecek yıl sayısı belirlenecektir. Örneğin; Grup 1'deki sekanslara 2001'den sonra rastlanmamıştır. Bu veri seti kullanılarak günümüzdeki sekanslara ulaşılması beklenmemektedir. Atılması gereken rastgele adım sayısı çok fazladır. Bu durum, başta sonuçların doğruluğu olmak üzere bilgisayar zamanı için de verimsiz olacaktır. Bu nedenle tahmin sisteminde eğitim seti ve doğrulama seti seçimi yapılırken, Grup 6 ve Grup 8 kendi içerisinde iki gruba ayrılmıştır. Yıl değerleri bir diğer gruba göre düşük olanlar, eğitim seti olarak, diğer grup ise doğrulama seti olarak kullanılmıştır. Bir başka eğitim seti olarak ise Grup 3 kullanılmış hem 2009 öncesi hem de 2009 sonrası sekanslar tahmin edilmiştir. Tahmin yapmak için gerekli bilgiler, eğitim setlerinden elde edilmektedir ve sonuç olarak elde edilen tahminlerin doğruluğu

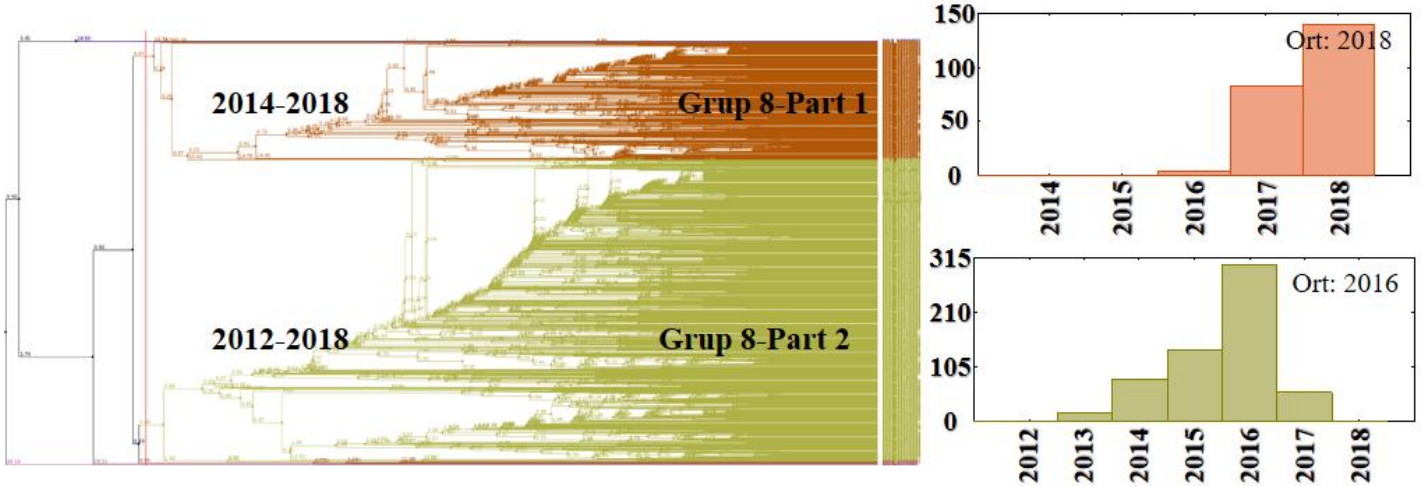
da doğrulama setleri ile karşılaştırılarak incelenmiştir. Bu grupları belirleme işlemi, filogenetik ağaç yardımıyla yapılmıştır ve elde edilen grup bilgileri Şekil 3.11 ve Şekil 3.12'de gösterilmiştir

Ana verinin detaylı analizinden sonra, tahmin sisteminin bir başka parametresi olan amino asit frekanslarının elde edilmesi için, iki temel yöntem uygulanmıştır. Bunlardan bir tanesi kodon tablosu kullanılarak amino asitlerin beklenen frekanslarının kullanılması, bir diğeri ise her set içerisinde bulunan amino asitlerin frekanslarının çıkarılarak bu frekansların kullanılmasıdır. Frekans bilgisi, tahmin modelinde mutasyon skorları yardımıyla belirlenen pozisyonun hangi amino aside dönüşeceğini belirleyecektir.

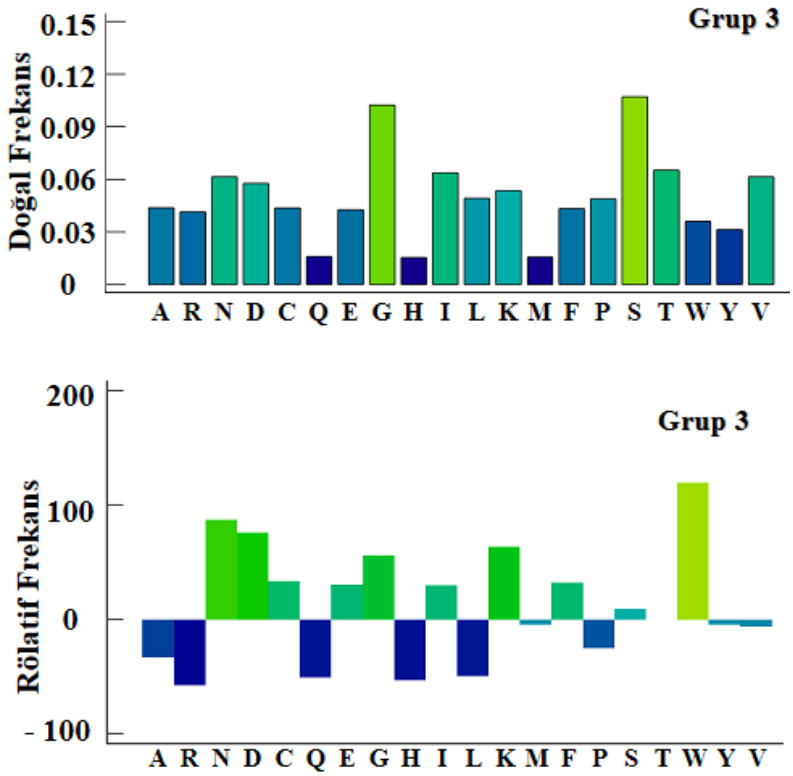
Seçilen eğitim grupları için hesaplanan amino asit frekansları ve beklenen frekanslara göre karşılaştırmaları, Şekil 3.13, Şekil 3.14 ve 3.15'te gösterilmiştir. Bu karşılaştırmalar yapılırken Eşitlik 2.1 kullanılmış ve göreceli frekanslar elde edilmiştir. Bu frekanslar incelendiğinde, bazı amino asitlerin frekansında düşüş gözlemlenirken bazı amino asitlerin frekansında artış gözlemlenmiştir (Şekil 3.13 (b)-3.15 (b)). Bu bilgi, rastgele yürüyüş modelinde hangi amino asitlerin hangi amino asitlere dönüştürüleceğinin belirlenmesinde kullanılacak olup NA proteininin evrimsel değişimi tahmininde önemli bir etki yaratabilir. Sonuç olarak bir tahmin modelinde amino asit frekansı için Şekil 2.2'de gösterilmiş olan beklenen frekans, bir diğerinde ise Şekil 3.13 (a)-Şekil 3.15 (a)'da verilen veri setlerindeki amino asitlerin doğal frekansları kullanılmıştır.



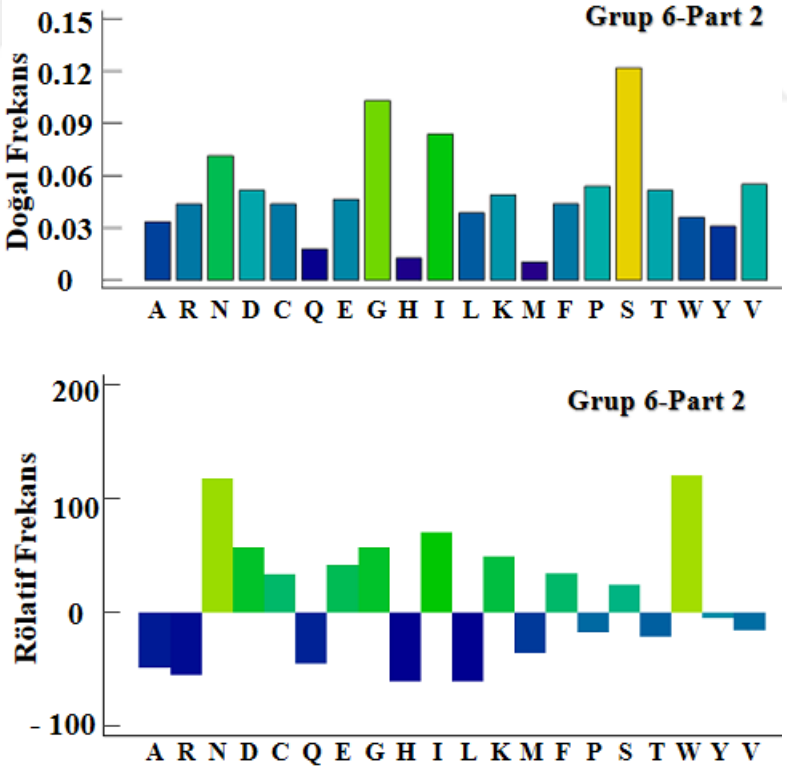

Şekil 3.11: Grup 6'daki sekansların filogenetik ağaçta gruplandırılması ve yıl dağılımı.



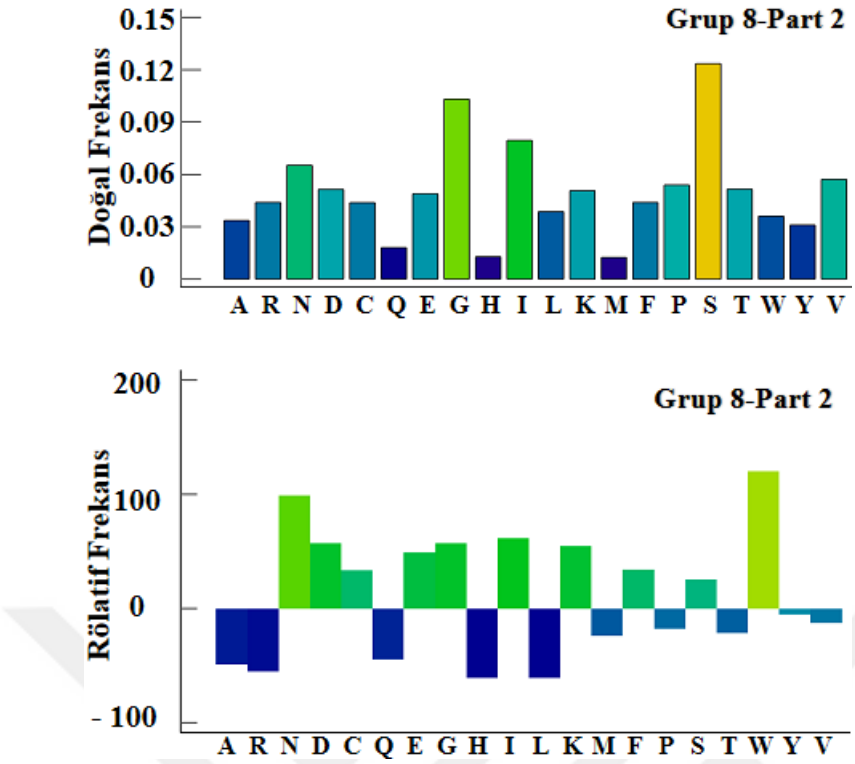

Şekil 3.12: Grup 8'deki sekansların filogenetik ağaçta gruplandırılması ve yıl dağılımı.




Şekil 3.13: Grup 3'teki sekansların doğal amino asit frekansları ve rölatif frekansları.




Şekil 3.14: Grup 6-Part 2'deki sekansların doğal amino asit frekansları ve rölatif frekansları.




Şekil 3.15: Grup 8-Part 2'deki sekansların doğal amino asit frekansları ve rölatif frekansları.

## 3.2 Metotlar Arası Korelasyon Analizi

Tez kapsamında, seçilen skorlama fonksiyonlarının hepsi bir skorlama matrisi kullanıp amino asitler arası evrimsel ilişkiyi hesaplayarak NA proteininin bölgesel mutasyon eğilimlerini çıkarmaktadır. Var olan 4 farklı skorlama matrislerinin etkisini ve ayrıca 6 farklı skorlama fonksiyonunun kendi içerisinde nasıl benzerlik ya da farklılık gösterdiklerini inceleyebilmek için korelasyon haritaları oluşturulmuştur.

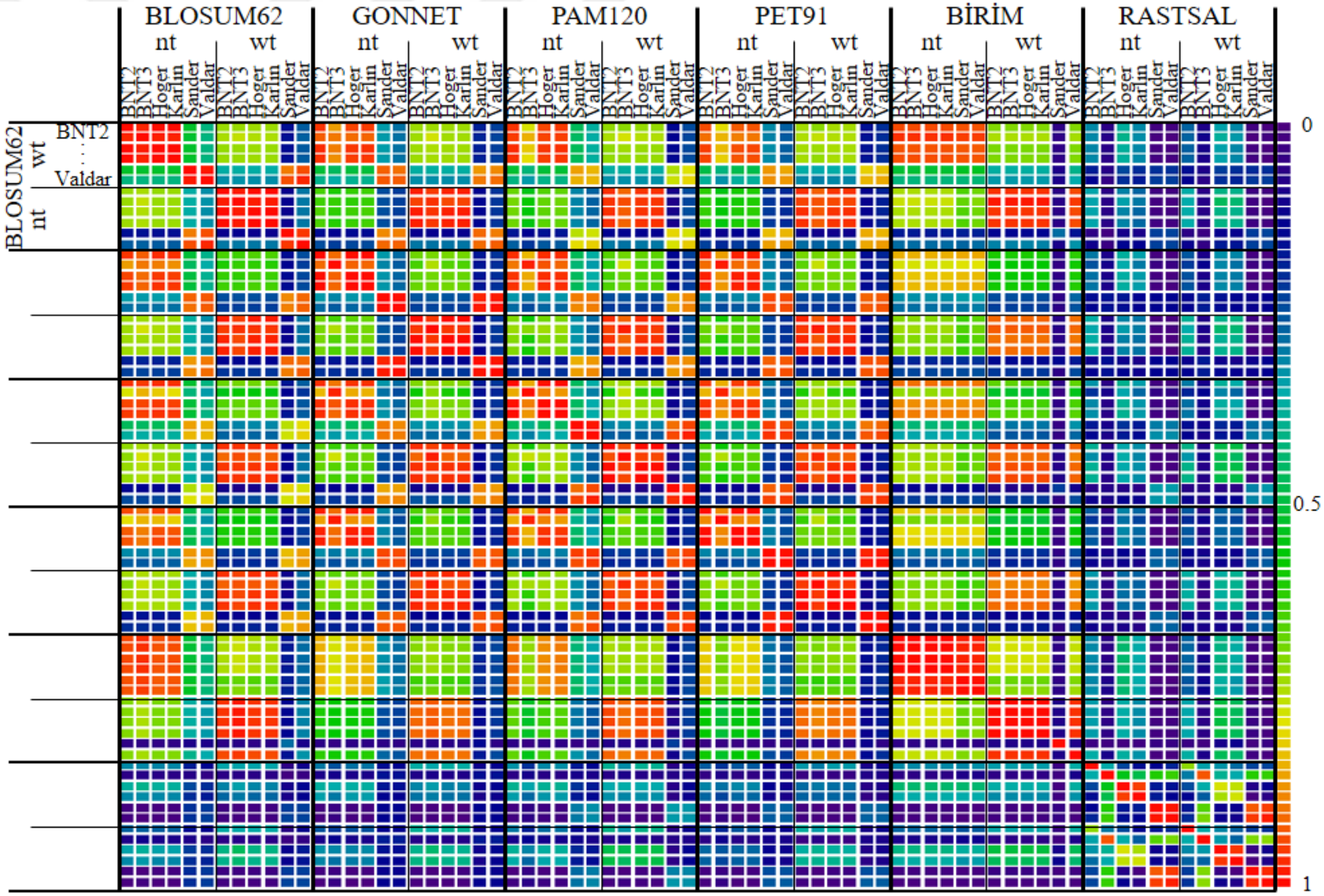
Skorlama fonksiyon sonuçlarının, yani mutasyon skorlarının, skorlama matrislerine bağlı olduğunun anlaşılabilmesi için skorlar, ayrıca birim matris ve rastsal olarak oluşturulmuş bir matris ile de hesaplanmıştır. Skorlar hesaplanırken iki ana veri seti girdi olarak kullanılmıştır. Bunlar, 1918-2006 yıllarına ve 2009-2015 yıllarına ait veri setidir.

Şekil 3.16 ve Şekil 3.17'de, kullanılan 6 skorlama matrisinin fonksiyonlar üzerindeki etkileri gösterilmiştir. Bu şekillerde, elde edilen sonuçların birbirleri ile tutarlılıkları kırmızı-mavi renk skalası ile belirtilmiştir: kırmızı tonlar tutarlılığın yüksek olduğunu mavi tonlar ise düşük olduğunu göstermektedir. Mavi tonlarının yoğun gözlemlendiği son kolon, rastsal matrisin sistemdeki tüm fonksiyonlar üzerinde farklı sonuçlar

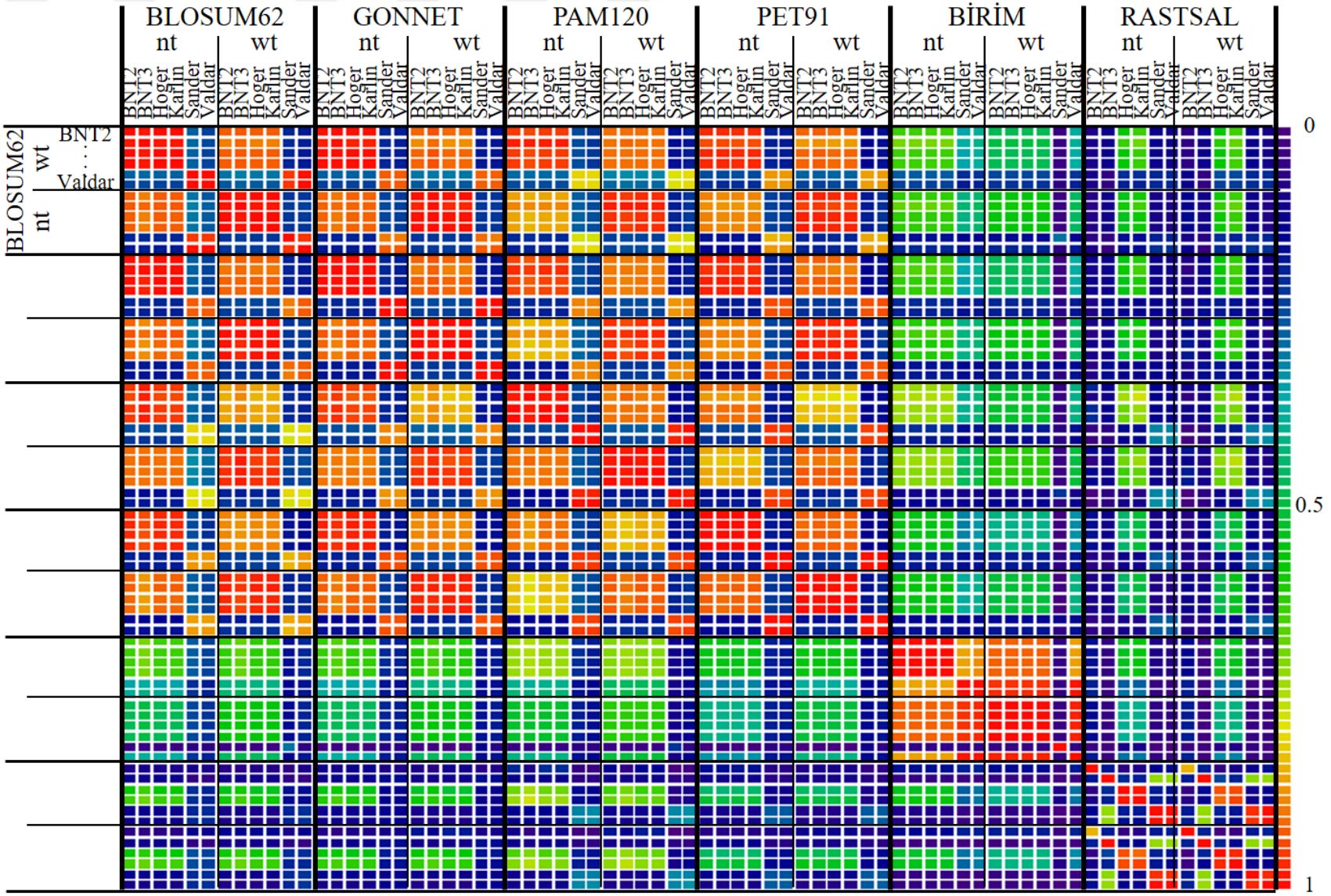
oluşturduğunu ve bu nedenle kendi içerisinde skorların benzerlik gösteremediğini belirtmektedir. Birim matris kullanıldığında ise 5. kolonda tüm matrislerin neredeyse aynı sonuçları verdiği, fonksiyonların ayırt edici özelliklerinin bulunmadığı gözlenmiştir. Siyah sınırlarla çevrilmiş her bir skorlama matrisinin sonuçları, diyagonal olarak bakıldığında, ilk 4 fonksiyonun sırasıyla BNT2, BNT3, Hogervorst ve Karlin'in kendi içerisinde benzerlik gösterdiğini ancak Sander ve Valdar'dan farklı sonuçlar oluşturduğu görülmektedir. Sander ve Valdar fonksiyonları ise kendi aralarında tutarlı skorlar oluşturmuştur. Bu iki gruplaşma kullanılan ilk 4 skorlama matrisi sonuçları için geçerlidir.

Şekil 3.16'da ve Şekil 3.17'de her bir kolon içerisinde iki parça bulunmaktadır. Bu parçalardan bir tanesi zaman bilgisi dahil edilmeden yapılan hesaplamaları, bir diğeri ise zaman bilgisi dahil edilerek hesaplanmış mutasyon skorlarını ifade etmektedir. Zaman bilgisi, mutasyon hızı bilgisi ile tahmin modeline eklenmiştir (Bakınız 2.4.1). Bu iki durum arasındaki fark, tek bir skorlama matrisi için incelendiğinde, zaman bilgisinin BNT2, BNT3, Hogervorst ve Karlin'nin korelasyonunda artış olduğu Sander ve Valdar'dan daha net bir şekilde ayrıldıkları görülmektedir. Sander ve Valdar da aynı şekilde kendi aralarında da benzer skorlar oluşturmuşlardır. Diyagonal olarak harita incelendiğinde tüm kullanılan matrislerde aynı eğilimin olduğu görülmektedir.

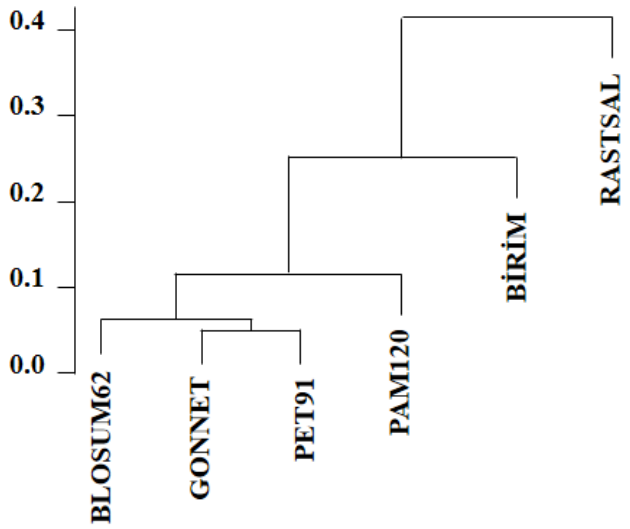
Skorlama matrislerinin mutasyon skorları, üzerinde etkisi olduğu gözlemlenmiştir. Bu bilgiye ek olarak matrislerin kendileri arasındaki benzerliklerine bakılmıştır (Şekil 3.18). Benzerlik sonuçları filogenetik ağaç ile oluşturularak incelendiğinde, GONNET ve PET91'in birbirine en yakın matrisler olduğu, bu grubu takiben BLOSUM62 matrisinin ve PAM120 matrisinin olduğu görülmektedir. Birim matris ve Rastsal matris ise bu oluşan yakın 4'lü gruba daha uzak bir benzerliğe sahiptir. Matrisler arası oluşan bu fark, mutasyon skorlarına da yansımaktadır. Bu skorlama matrislerinden BLOSUM62 matrisi seçilerek evrimsel değişim tahminleri yapılmıştır.




Şekil 3.16: 1918-2006 Veri seti ile elde edilen mutasyon skorları ve skorlama matrisleri arası korelasyon haritası.




Şekil 3.17: 2009-2015 Veri seti ile elde edilen mutasyon skorları ve skorlama matrisleri arası korelasyon haritası.




Şekil 3.18: Skorlama matrisleri arasındaki ilişkinin filogenetik ağaç ile gösterimi.

Mutasyon skorlarının daha detaylı incelenebilmesi için ilk olarak skor dağılımı oluşturulmuştur. Mutasyon skorları 0-1 aralığında değerler almaktadır. 0 değeri, veri seti içerisinde seçilen bir pozisyonun en yüksek korunma değerini, 1 değeri ise o pozisyonun en düşük korunmaya bir başka deyişle en çok mutasyona sahip olduğunu göstermektedir.

1918-2006 ve 2009-2015 yılları için zaman bilgisi kullanılarak mutasyon skorları elde edilmiştir. Skor dağılımı, bu bilgi yardımı ile incelendiğinde, 388 amino aside sahip NA proteini baş bölgesinin, 1918-2006 yılları arasında 219; 2009-2015 yılları arasında ise 148 tane pozisyonunun korunduğu görülmektedir (Şekil 3.19). Skor dağılımına bakıldığında korunan pozisyonların, mutasyona uğrayan pozisyonlara göre daha fazla olduğu görülmektedir. Mutasyona uğrayan bölgelere yoğunlaşılması için her bir fonksiyon için bir eşik değeri atanarak en düşük skora sahip ilk 10 pozisyon tespit edilmiştir. NA proteini üzerindeki tehlikeli bölgelerin tespit edilmesi için mutasyona uğrayan bölgelerin doğru bir şekilde ayırt edilmesi önemli bir bilgi sunacaktır. Eşik değerinin belirlenmesi için her bir fonksiyon sonucunda elde edilen skorlar, kendi içerisinde küçükten büyüğe doğru sıralandı. Şekil 3.20'de sıralanan skor değerleri görülmektedir. BNT2, BNT3, Hogervorst ve Karlin parabolik artış göstererek birbirlerine yakın değerlerde sıralanmıştır. Sander ve Valdar ise bu gruplaşan fonksiyonlardan ayrılmaktadır. Bu fonksiyonlar ise blok halinde belli bölgelerdeki pozisyonlar için aynı sonuçları bulmuşlardır. Bu durum ayırt edici olan bölgelerin

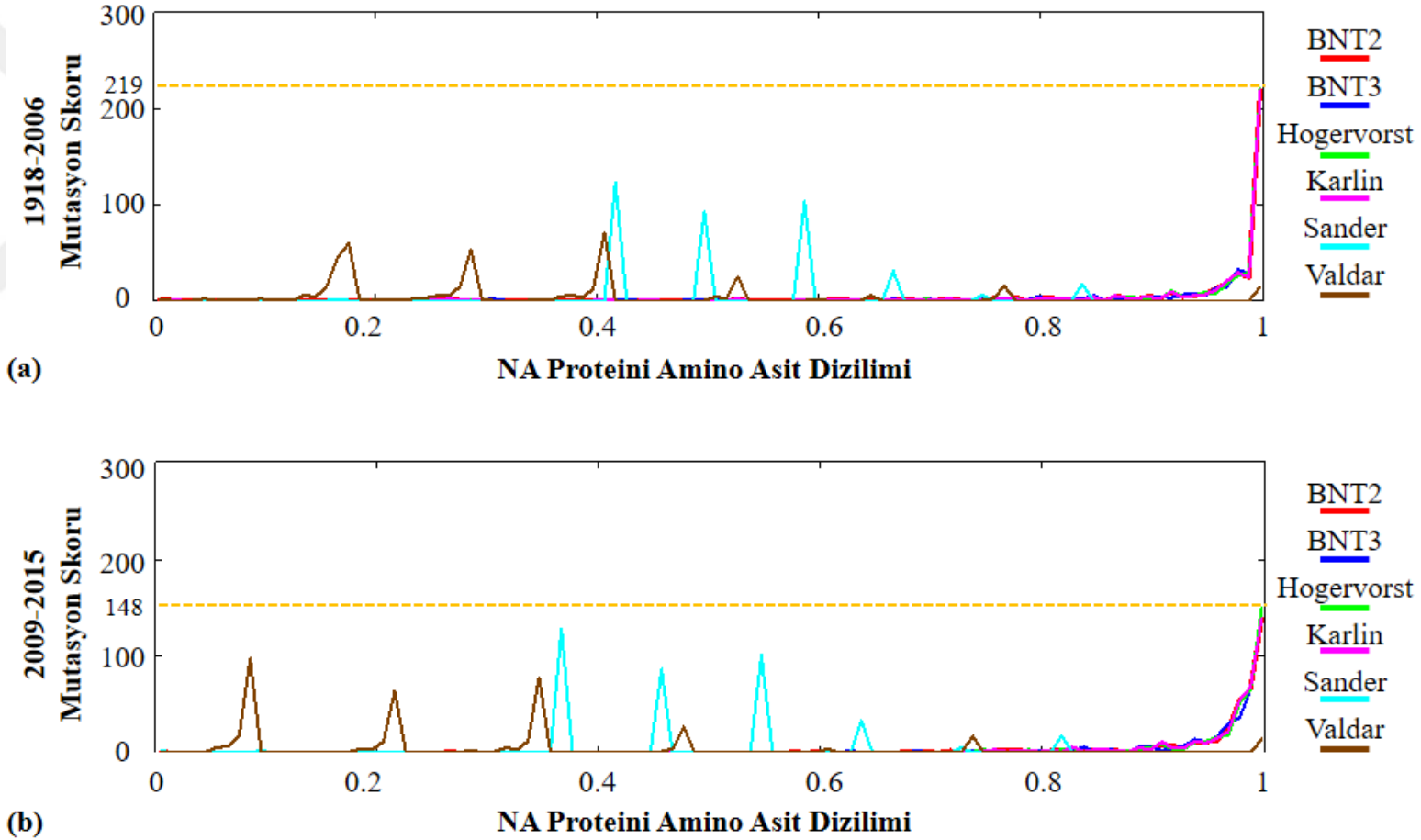
belirlenmesine engel olmaktadır. Bu nedenle yakın korele sonuç veren fonksiyonlar üzerinden çalışmalara devam edilmiştir.

Eşik değerlerine göre elde edilen pozisyonlardan, fonksiyonlar arası ortak çıkanların sayısı Şekil 3.21'de korelasyon haritası üzerinde gösterilmiştir. Mutasyona uğrama olasılığı yüksek olan bu pozisyonlar Çizelge 3.3'te görülmektedir. Sonuçlar göstermektedir ki fonksiyonlar arası 10 mutant pozisyondan en az 8'i ya da 9'u ortaktır.

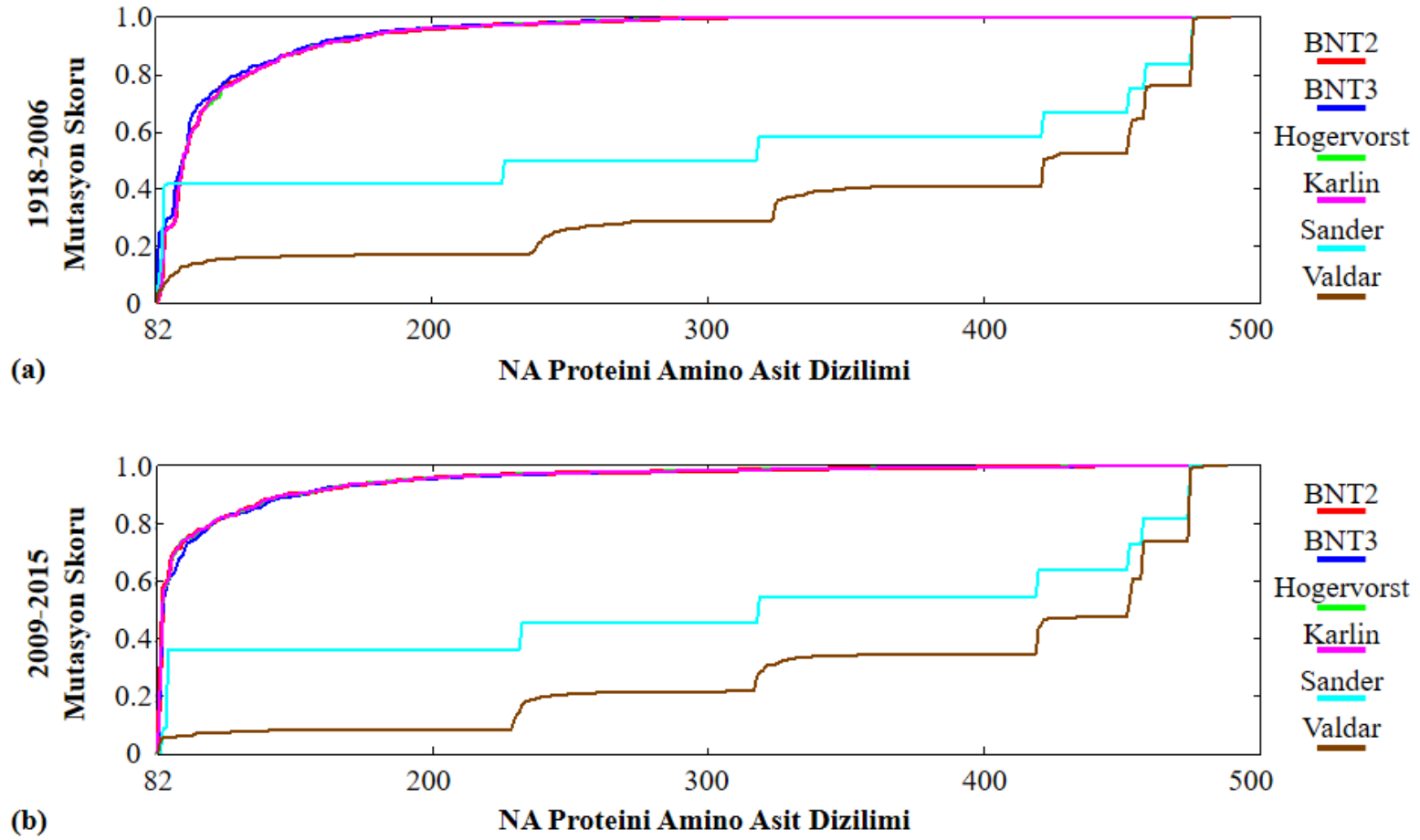
1918-2006 datası kullanılarak yapılan 4 farklı skorlama fonksiyonlarının sonuçlarına göre 222 pozisyonu mutasyona karşı en hassas pozisyondur. 2009-2015 verisi kullanılarak yapılan skorlama hesaplamalarına göre ise 369 pozisyonu en hassas pozisyondur. Yüksek mutasyon skorlarına sahip bu pozisyonlar Şekil 3.8'deki dirençli mutasyonlar ile karşılaştırıldığında 1918-2006 verisi sonuçlarından 234, 249, 344 pozisyonları; 2009-2015 verisi sonuçlarından 241, 275, 369 pozisyonları ile eşleşmektedir. Özellikle literatürde H275Y mutasyonu en çok çalışılan dirençli pozisyondur. Yapılan skorlama sonuçlarına göre 275 pozisyonu ilk 10 mutasyon arasına girmiştir. Ayrıca 275 ve 344 pozisyonları aktif bölgede bulunduğundan, yüksek mutasyon skoruna sahip olmaları ilaç direnci oluşmasında direkt (ortosterik) bir etkiye sahiptir. Bir pozisyonda meydana gelen bir mutasyon, çevresinde bulunan amino asitlere de etki etmektedir. Bu durum gerçekte var olan dirençli mutasyonlarda da görülmektedir. 117,119; 148, 151, 152; 222, 223; 247, 248, 249, 250 pozisyonları buna örnek verilebilir. Çizelge 3.3'te elde edilen sonuçlara bakıldığında ise yukarıdaki yakın pozisyondaki amino asitlere ek olarak 450, 451, 452 pozisyonları gelmektedir.

Çizelge 3.3'te 1918-2006 ve 2009-2015 sonuçları kendi içerisinde karşılaştırıldığında ilk 10 mutasyondan eşleşen bir pozisyon bulunmamaktadır. Bu durum yıllar içerisinde NA proteinin değişiminin fazla olduğunu göstermektedir. Buna etki eden en büyük faktör 2009 yılında görülmüş olan pandemidir.

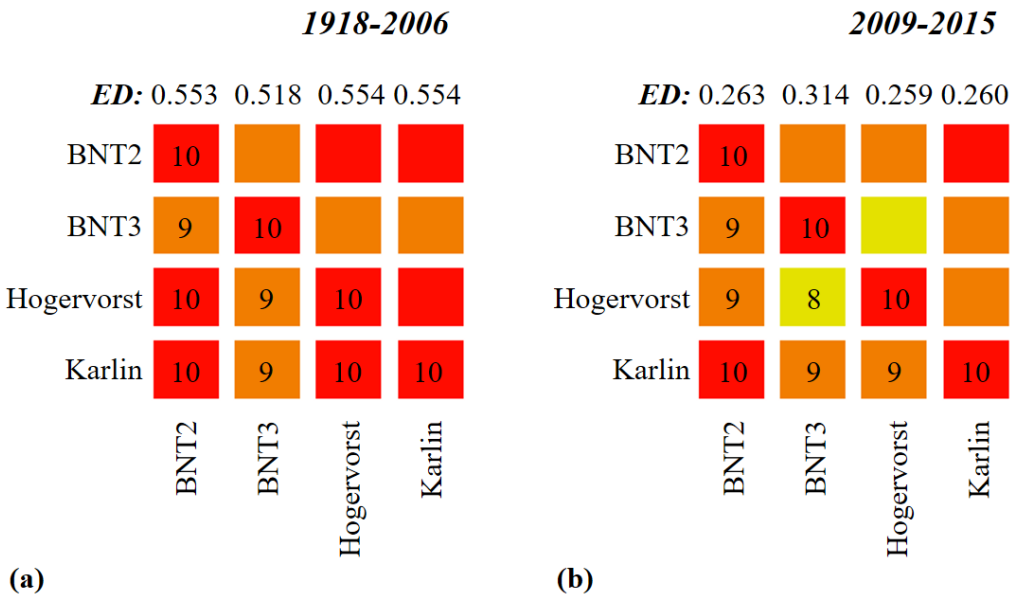
Korelasyon haritalarından yararlanarak elde ettiğimiz sonuçlar, BNT2, BNT3, Hogervorst ve Karlin fonksiyonlarının ortak sonuçlarının fazla olduğunu göstermektedir. Bu nedenle tahmin modeli için bu grup içerisinden bir tanesi seçilerek, fonksiyon sayısını dörtten bire indirgeyerek tahmin çalışmaları yapılmıştır. Tahmin çalışmalarında BNT3 skorlama fonksiyonu kullanılmıştır.




Şekil 3.19: Mutasyon skorlarının histogramı. (a) 1918-2006 yılı dağılımı. (b) 2009-2015 yılı dağılımı.




Şekil 3.20: En küçük skordan en büyük skora göre pozisyonların sıralanması. (a) 1918-2006 yılı sıralaması. (b) 2009-2015 yılı sıralaması.



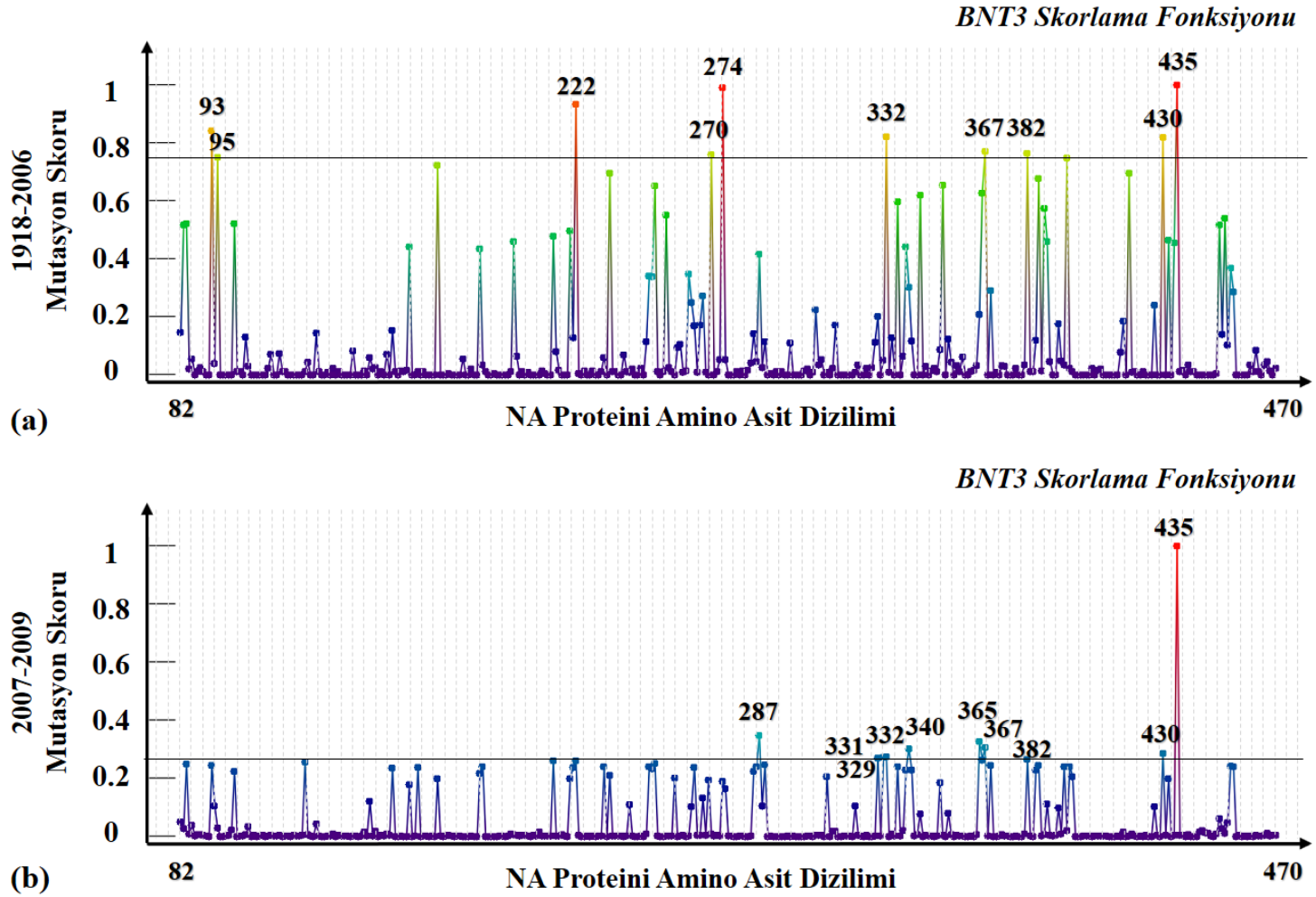

Şekil 3.21: Eşik değerine göre mutasyona uğrama olasılığı yüksek olan pozisyonların korelasyon haritası. (a) 1918-2006 sonuçlarına göre korelasyon haritası. (b) 2009-2015 sonuçlarına göre korelasyon haritası. *ED: Eşik Değeri

Çizelge 3.3: Mutasyon skoru yüksek olan ilk 10 pozisyon.

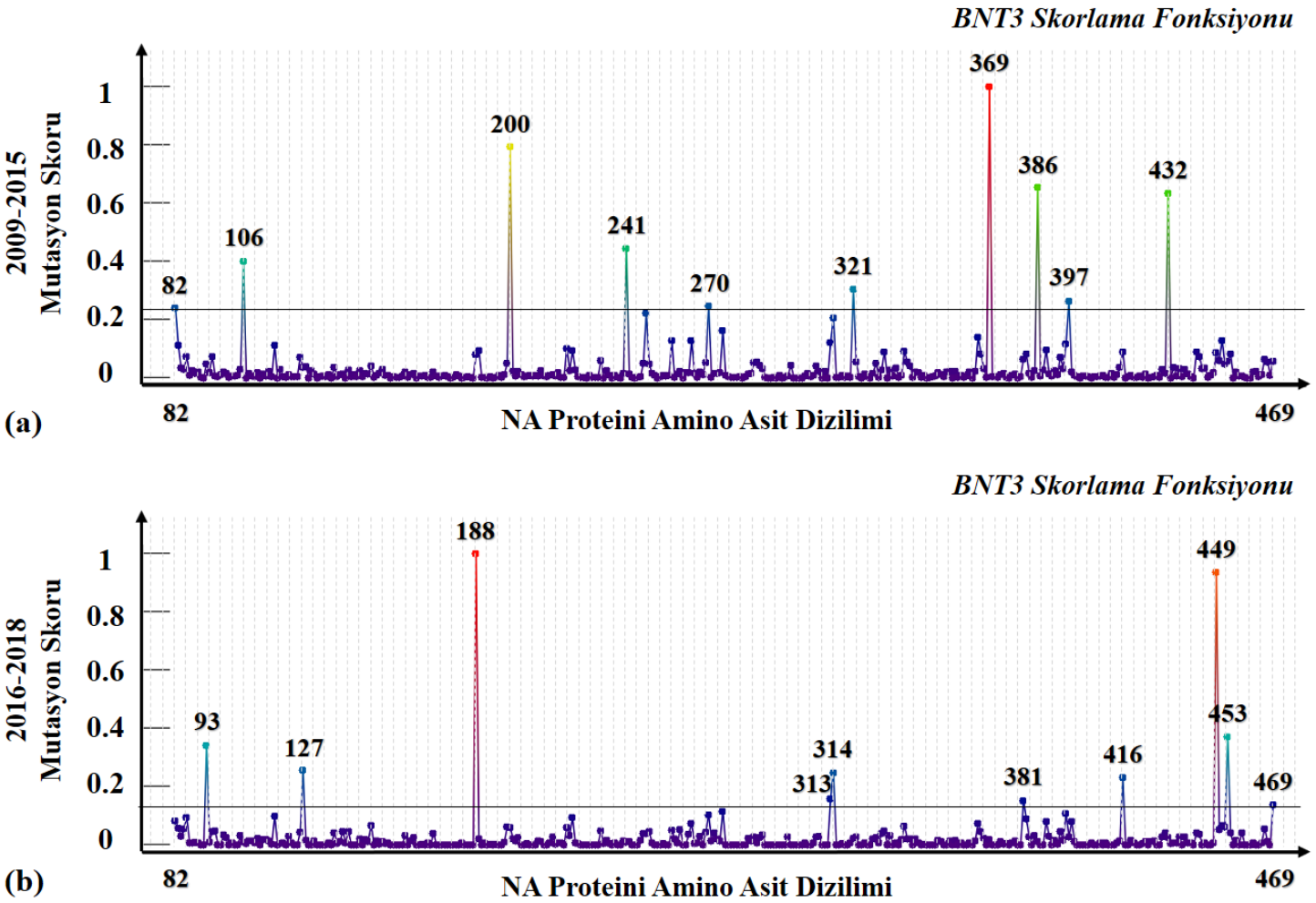
| 1918-2006 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| BNT2 | 214 | 222 | 234 | 249 | 267 | 332 | 344 | 382 | 450 | 452 |
| BNT3 | 173 | 214 | 222 | 234 | 267 | 332 | 344 | 382 | 450 | 452 |
| Hogervorst | 214 | 222 | 234 | 249 | 267 | 332 | 344 | 382 | 450 | 452 |
| Karlin | 214 | 222 | 234 | 249 | 267 | 332 | 344 | 382 | 450 | 452 |
| 2009-2015 | | | | | | | | | | |
| BNT2 | 82 | 200 | 241 | 248 | 275 | 365 | 369 | 386 | 397 | 432 |
| BNT3 | 82 | 106 | 200 | 241 | 248 | 275 | 369 | 386 | 397 | 432 |
| Hogervorst | 82 | 200 | 241 | 248 | 365 | 369 | 386 | 397 | 432 | 451 |
| Karlin | 82 | 200 | 241 | 248 | 275 | 365 | 369 | 386 | 397 | 432 |

## 3.3 Mutasyon Haritaları

Tahmin modelini oluşturan temel girdilerden bir tanesi, mutasyon skorlarıdır (Bakınız Bölüm 2.4.2). Bu bölümde, BNT3 fonksiyonunun NA proteini baş bölgesinde bulunan tüm pozisyonlar için hesaplamış olduğu mutasyon skorları, Şekil 3.22 ve Şekil 3.25 arasında görülmektedir. Mutasyon skoru yüksek ilk 10 pozisyon, her bir mutasyon haritası için gösterilmiştir. (a) şıklarında bulunan sekanslar eğitim seti olarak, (b)'dekiler ise doğrulama seti olarak kullanılmıştır.




Şekil 3.22: Zaman bilgisi içermeyen mutasyon haritaları. (a) 1918-2006 veri seti mutasyon skorları. (b) 2007-2009 veri seti mutasyon skorları.



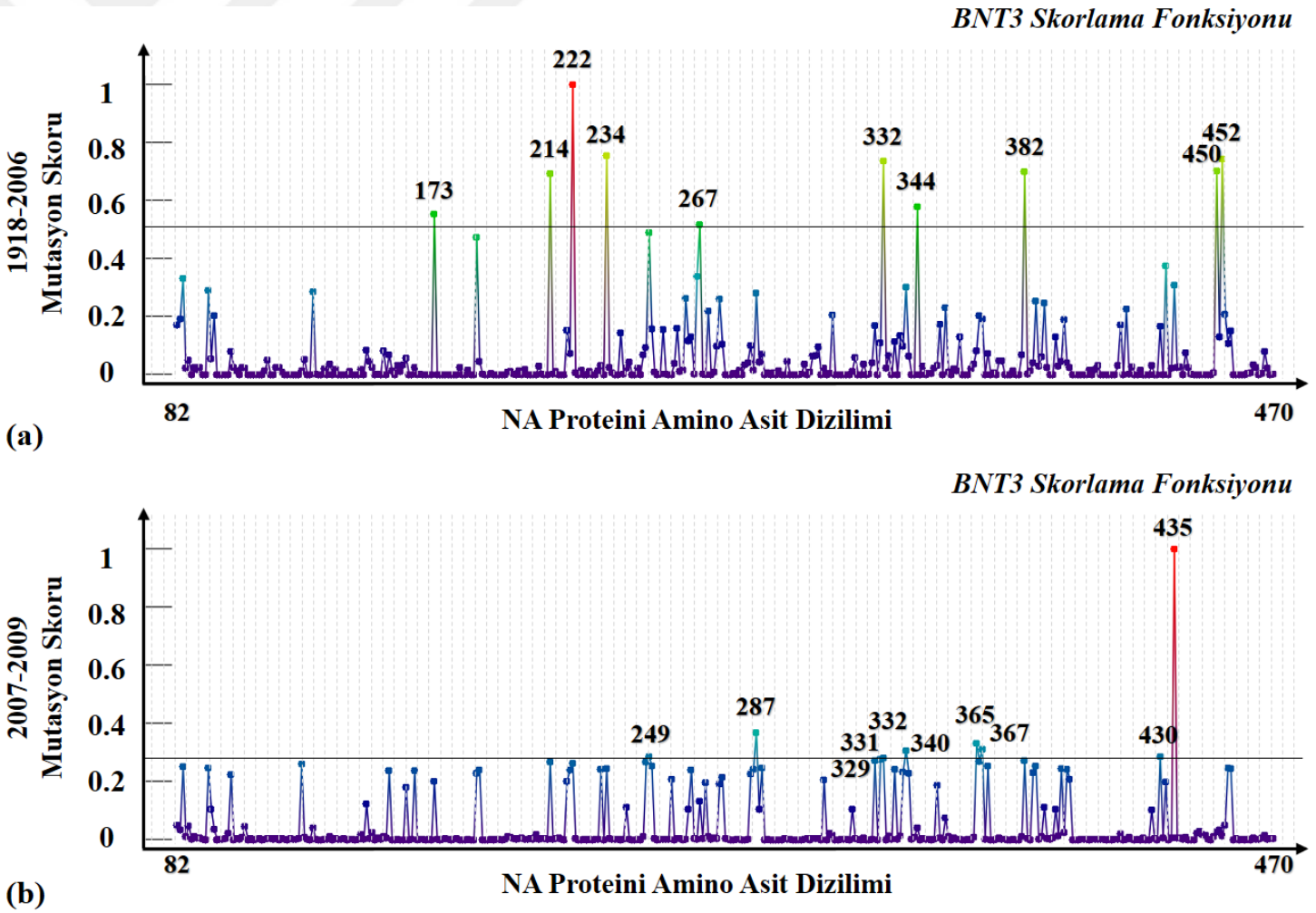

Şekil 3.23: Zaman bilgisi içermeyen mutasyon haritaları. (a) 2009-2015 veri seti mutasyon skorları. (b) 2016-2018 veri seti mutasyon skorları.

Bölüm 2.4.3'te bahsedildiği gibi, eğitim setleri tahmin modeli içerisinde eğitilerek, doğrulama setindeki verilere ulaşılmaya çalışılacaktır. Bir başka deyişle, geçmişteki veriler (Ör. 1918-2006) kullanılarak gelecekte görülebilecek veriler tahmin edilecektir. Bu tahminlerin doğruluğunun anlaşılması için doğrulama seti (Ör: 2007-2009) ile karşılaştırmalar yapılacaktır.

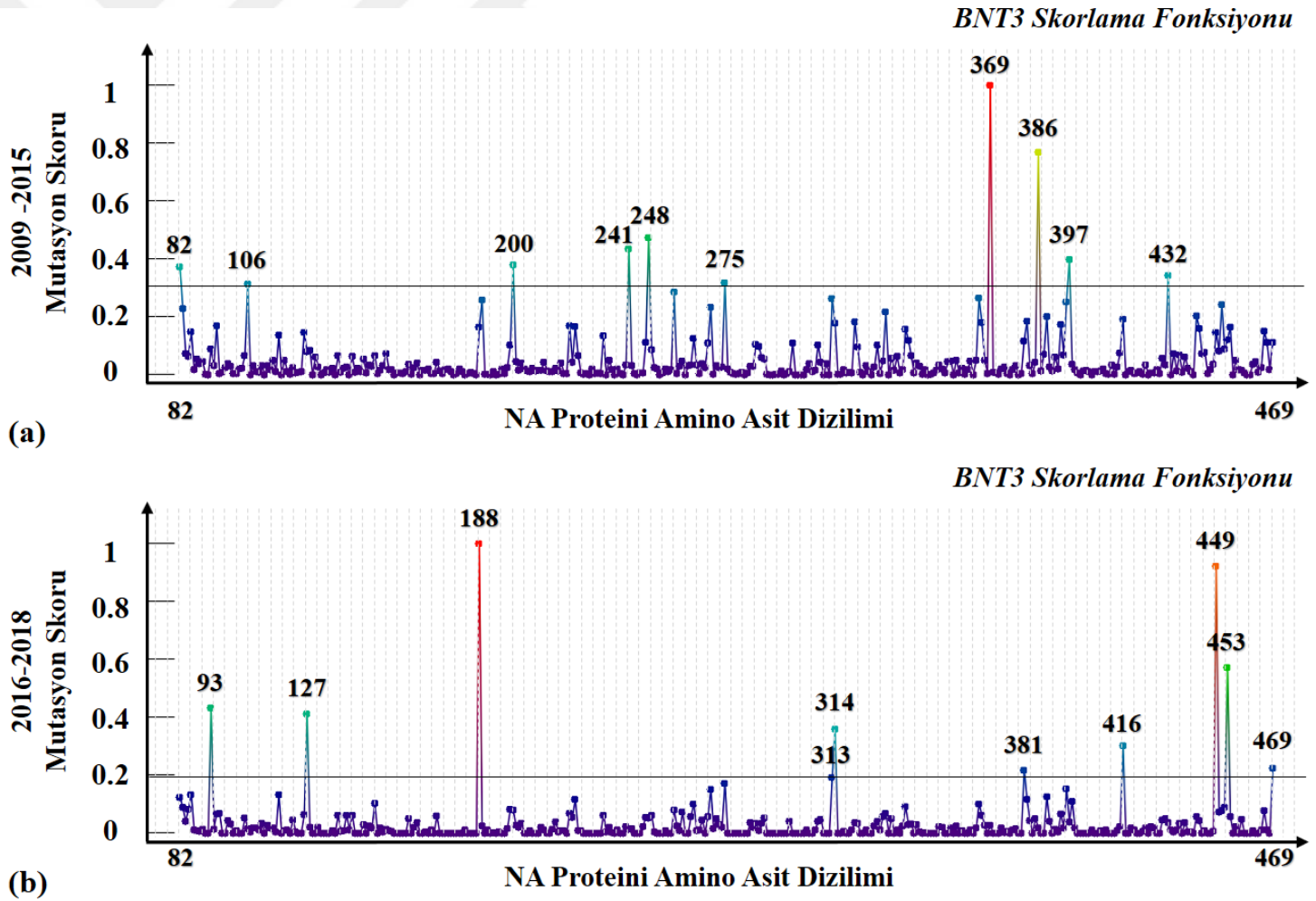
NA proteininin mutasyona karşı hassas ve korunan bölgeleri mutasyon skorları hesaplanarak tespit edilmiştir. Tahmin modelinin algoritması olan rastgele yürüyüş modelinde hangi pozisyonların değişime uğrayacağı mutasyon skorlarına bağlı olarak değişecektir. Toplam vektörü olarak bahsedilen vektör, mutasyon skorlarının art arda toplanmasıyla elde edilmiştir. Bu vektörde, yüksek skora sahip pozisyonlar, toplam değerini diğer düşük değerlere göre daha fazla arttıracağından, rastsal olarak atılan adımın, toplam vektör değeri yüksek olan aralığa denk gelme olasılığı daha yüksektir. Yani mutasyon skor değeri yüksek olan pozisyonların mutasyona uğrama olasılığı daha yüksektir. Eşitlik 2.24'te Toplam vektöründeki $x_1$ ilk pozisyonun mutasyon skorunu belirtmektedir. Toplam vektörüyle ilgili detaylı bilgi için Bölüm 2.4.2'den yararlanılabilir.

Şekil 3.22'de ve Şekil 3.23'te zaman bilgisi kullanılmadan hesaplanan mutasyon haritaları görülmektedir. Şekil 3.22 (a) ve Şekil 3.22 (b) karşılaştırıldığında bazı pozisyonlar iki grupta da mutasyon olasılığı yüksek olarak bulunsa da bazı pozisyonlar için farklılıklar görülmektedir. Bu farklılıklar, tahmin modelinde 1918-2006 verisi kullanılarak 2007-2009 verisinin tahmin edilmesini zorlaştıracaktır. Aynı şekilde Şekil 3.23'teki (a) ve (b) birbirleriyle karşılaştırıldığında da 2009-2015 veri seti ile 2016-2018 veri seti arasında farklılıklar vardır.

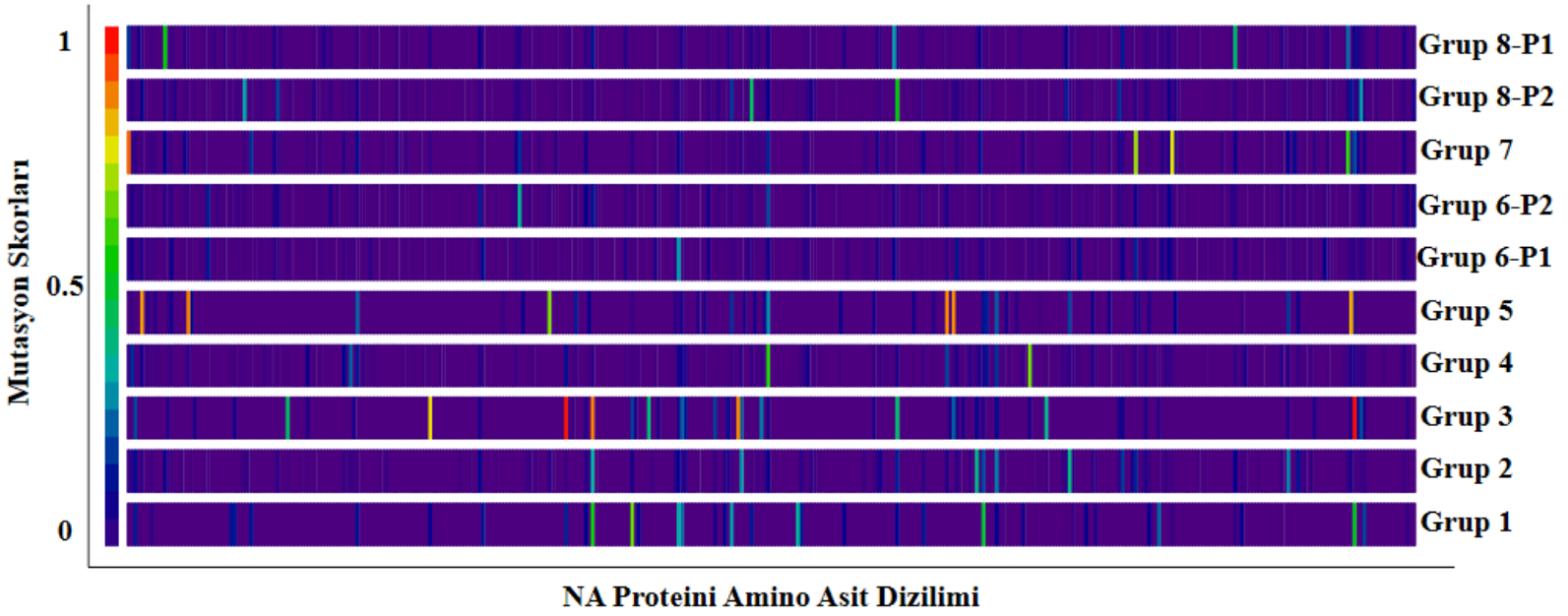
Aynı durum, zaman bilgisi sisteme dahil edildiğinde de görülmektedir (Şekil 3.24-3.25). Buna çözüm olarak ana veri seti filogenetik ağaç yardımıyla 8 gruba bölünerek incelenmiştir. Grupların veri sayısı ve yıl dağılımı bölüm 3.1'de bulunmaktadır. Bölüm 3.1'deki gruplandırmalara göre 8 grup içerisinden Grup 6 ve Grup 8 iki parçaya bölünmüştür. Toplam 10 grup için mutasyon haritaları Şekil 3.26'da görülmektedir. Haritalar karşılaştırıldığında Grup 6-Part 1ve Part 2 kendi içinde, Grup 8-Part1 ve Part 2 kendi içinde en yakın mutasyon skorlarını veren gruplardır.




Şekil 3.24: Zaman bilgisi içeren mutasyon haritaları. (a) 1918-2006 veri seti mutasyon skorları. (b) 2007-2009 veri seti mutasyon skorları.




Şekil 3.25: Zaman bilgisi içeren mutasyon haritaları. (a) 2009-2015 veri seti mutasyon skorları. (b) 2016-2018 veri seti mutasyon skorları.




Şekil 3.26: NA proteini sekans gruplarının mutasyon haritaları.

Bu sonuç gruplar içindeki sekansların birbirlerine daha benzer olduğunu göstermektedir. Bu yeni gruplandırmaya göre tahmin modeli için eğitim ve doğrulama setleri seçilmiştir (Çizelge 3.4).

Çizelge 3.4: Tahmin modelinde kullanılan eğitim ve doğrulama setleri.

| Eğitim Seti | Doğrulama Seti (Hedef) |
|---|---|
| Grup 3 | Grup 4, Grup 5, Grup 8-Part 1, Grup 8-Part 2 |
| Grup 6-Part 2 | Grup 6-Part 1, Grup 7, Grup 8-Part 1, Grup 8-Part 2 |
| Grup 8-Part 2 | Grup 8-Part 1 |

## 3.4 Tahmin Metotlarının Performans Sonuçları

Tahmin modeli için seçilen eğitim gruplarının mutasyon skorları ve amino asit frekansları kullanılarak 5 yıllık tahminler yapılmıştır. Rastgele yürüyüş metodunda atılan her adım, hesaplanan mutasyon hızına göre bir yılı ifade etmektedir. Her yıl için tahmin yapıldıktan sonra kümeleme yöntemi ile seçilen referans sekanslar üzerinde mutasyon hızına bağlı olarak tek pozisyonda değişim olmuştur. Yapılan tahminlerde amaç, hedeflenen doğrulama setine yakın sekansların elde edilmesidir. Hedef ile karşılaştırmalar yapılarak tahminlerin doğruluk analizleri yapılmıştır.

İlk olarak toplam benzerlik skorları hesaplanmıştır. Şekil 3.27-Şekil 3.31'de Y1-Y5 olarak adlandırılan setler, 1. yıl ve 5. yıl arasındaki tahmin setlerini ifade etmektedir. Her bir tahmin seti, hedef (doğrulama seti) ile karşılaştırılarak hedef seti ile olan benzerlik skorları elde edilmiştir. Elde edilen sonuçların, hedef seti içerisinde bulunan sekansların sahip olduğu benzerlik skoruna yakın çıkması beklenmektedir. Bu benzerlik skoruna yaklaşıldığında hedefteki sekanslara yakın sekansların tahmin edildiği görülecektir. Bu nedenle, histogramlarda başlangıç noktası olarak eğitim-hedef karşılaştırması, daha sonra tahmin-hedef karşılaştırmaları ve bitiş noktası olarak hedefin benzerlik skoru verilmiştir.

Toplam benzerlik skorlarında her bir amino asidin birbirine dönüşümü skorlama matrisleri kullanılarak hesaplanmaktadır. Skorlama matrislerinde her amino asidin birbirine dönüşme olasılıkları aynı olmadığından buna ek olarak, sadece değişim olması durumu için pozisyona bağlı amino asit farkları hesaplanmıştır. Toplam benzerlik skorlarında olduğu gibi eğitim-hedef, tahmin-hedef ve hedef karşılaştırmaları yapılmıştır. Amino asit farkının az olması durumu, hedefe yaklaşıldığı anlamına gelmektedir.

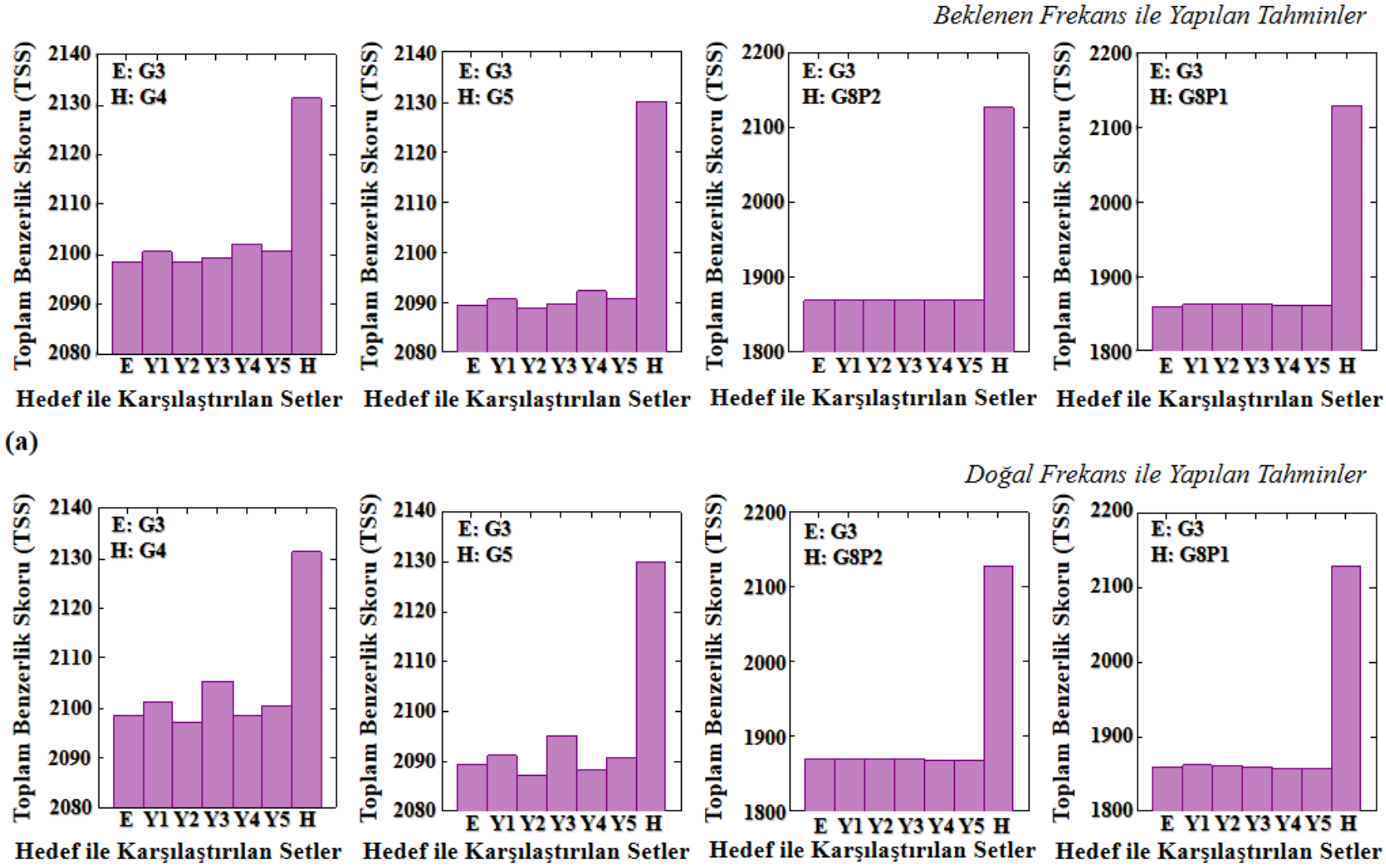
Şekil 3.27, Şekil 3.29 ve Şekil 3.31'deki toplam benzerlik skorlarında yüksek çıkan setler ile Şekil 3.28, Şekil 3.30 ve Şekil 3.31'deki en düşük pozisyon farkına sahip olan setler aynı setlerdir. Bu durum, tahminlerin analizi için kullanılan iki yöntemin verdiği sonuçlar arasında tutarlılık olduğunu göstermektedir.

Eğitim setinin grup 3 olduğu durumda, tahminler grup 4 ve grup 5'e daha yakın sonuçlar verirken, grup 8'e daha uzak sonuçlar vermektedir (Şekil 3.27). Bunun sebebi Şekil 3.9'da görüldüğü gibi grup 1-grup 5'in, 2009 pandemisinden önceki sekansları, grup 6-grup 8'in ise 2009 sonrasındaki sekansları içermesidir. 2009 Pandemisi, NA proteini üzerinde genetik sürüklenmeye sebep olduğu için grup 3 (2009 öncesi) kullanılarak grup 8'in (2009 sonrası) tahmin edilmesi beklenmemektedir. 2009 yılında meydana gelen pandemide domuzda ve insanda görülen influenza virüsleri arasında gen transferi olmuştur. Bu nedenle 2009 yılından önceki sekans bilgileri kullanılarak gerçekleşen bu değişimin tahmin edilmesi bu yöntem ile mümkün değildir.

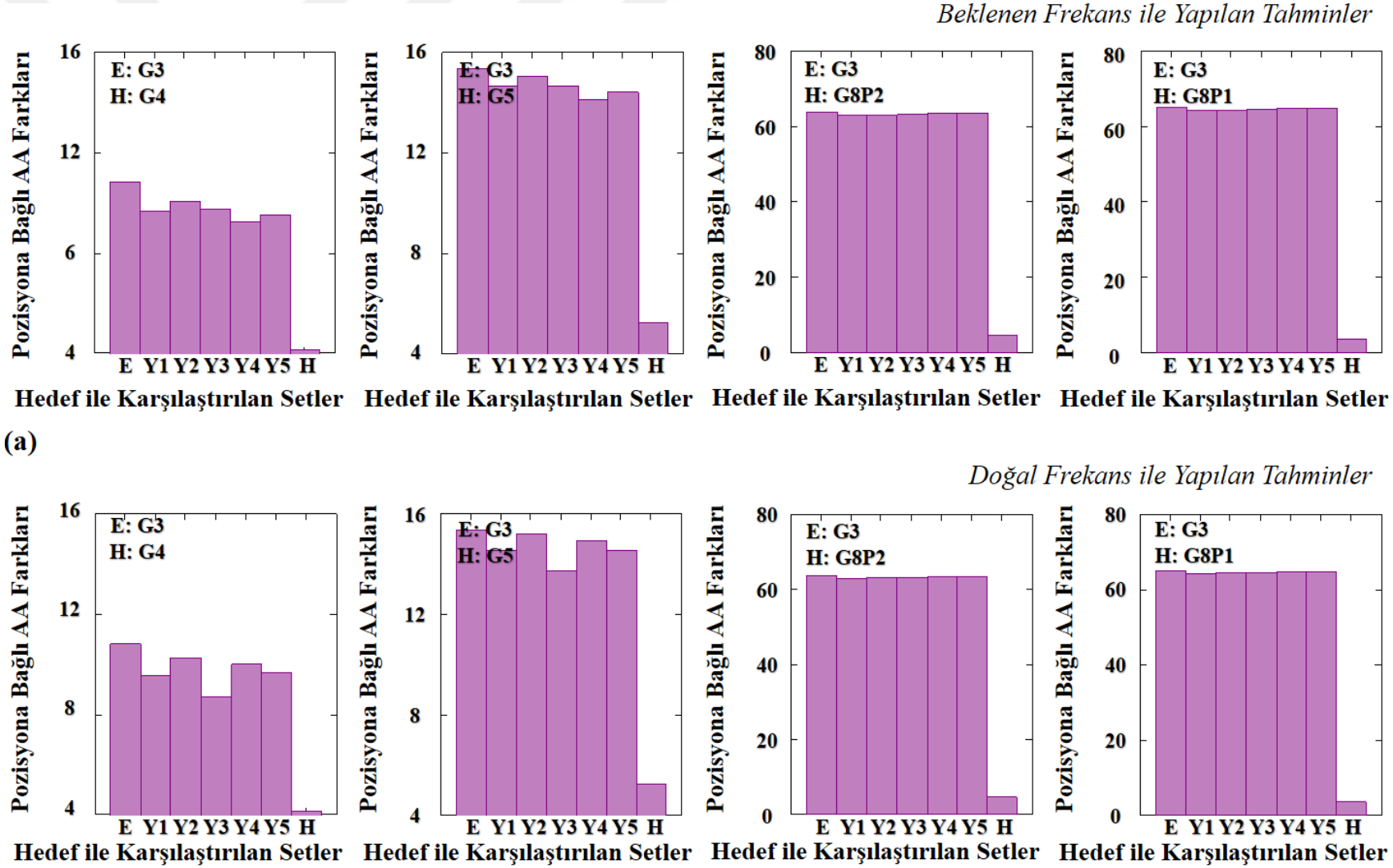
Her grup kendi içinde karşılaştırıldığında, grup 4 ve grup 5 ile toplam benzerlik sonuçlarında dalgalanmalar görülmektedir. Grup 8'in parçalarıyla karşılaştırıldığında ise küçük değer farkları göz ardı edildiği durumda sabit kalmıştır. Tahmin modelinde, beklenen frekanslar kullanılarak tahmin yapıldığında hedef Grup 4 ve Grup 5'e en çok 4. yıl tahmini ile doğal frekanslar kullanıldığında ise 3. yıl tahmini ile yaklaşılmıştır.

Eğitim setinin grup 6 part 2 olduğu durumda, Şekil 3.29'da toplam benzerlik skoru en yüksek değerlere, hedef grup 6-part1 olduğunda ulaşmaktadır. Bu durumun oluşmasındaki etken ana gruplardan biri olan grup 6'nın iki parçaya ayrılarak eğitim ve doğrulama setlerinin oluşturulmasıdır. Diğer hedefler arası karşılaştırmalara bakıldığında hedef ile eğitim seti arasındaki yıl farkı arttıkça, toplam benzerlik skorları azalmaktadır. Beklenen frekanslar kullanılarak tahmin yapıldığında grup 6 part 1 dışındaki hedeflere en çok 3. yıl tahmini ile; doğal frekanslar kullanıldığında ise 1. yıl tahmini ile yaklaşılmıştır. Grup 6 part 1 için iki frekans kullanıldığında da 1. yılda en yüksek benzerlik değerlerine ulaşılmıştır.

Grup 8 part 2 eğitim seti olarak alındığında ise hedef grup 8 part 1 ile toplam benzerlik skoru hesaplandığında tahmin yılı arttıkça değerlerde düşüş görülmektedir (Şekil 3.31). Bu sonuca bağlı olarak iki frekans için de 1. yılda en yüksek benzerlik sonuçlarına ulaşılmıştır.




Şekil 3.27: Grup 3 ile yapılan 5 yıllık tahminlerin ve eğitim (E) setinin, hedef (H) ile hesaplanan toplam benzerlik skorları.




Şekil 3.28: Grup 3 ile yapılan 5 yıllık tahminlerin ve eğitim (E) setinin hedef (H) ile hesaplanan pozisyona bağlı amino asit (AA) farkları.

Beklenen frekans ve doğal frekans olmak üzere kullanılan iki ayrı frekans bilgisi için elde edilen tahmin sonuçları hem toplam benzerlik skoru için hem de pozisyona bağlı amino asit farkı için yakın aralıklarda çıkmıştır.

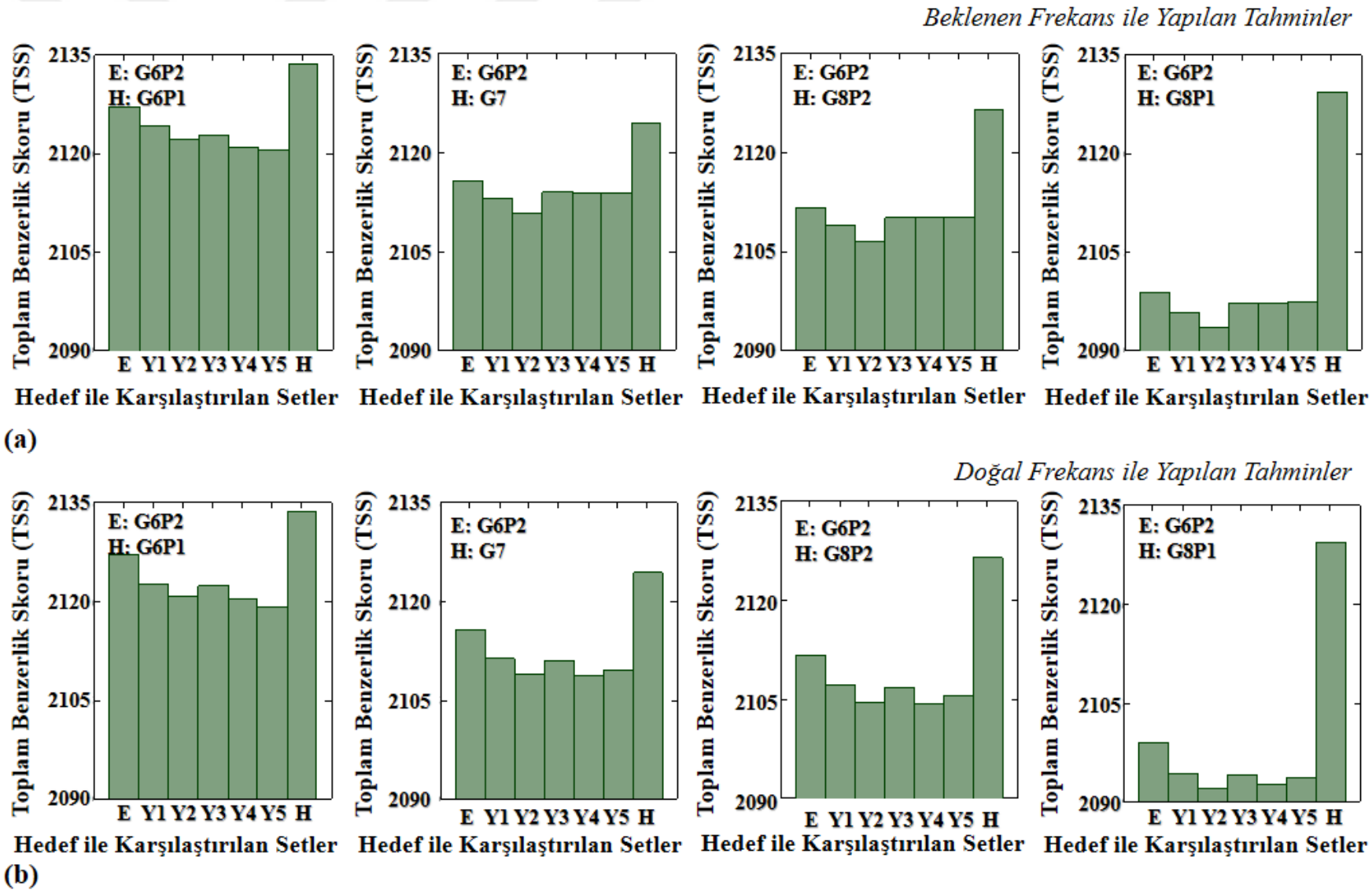
Genel olarak tüm sonuçlarda hedefe tam olarak yaklaşılamamıştır. Hedef içerisinde bulunan sekansların hepsinin tahmin edilebilmesi durumunda anca, hedeflerin (H) kolonu olarak verilen histogram değerlerine ulaşılabilir. Gerçekte tüm hepsinin tahmin edilebilmesi pek mümkün değildir. Bu nedenle tahmin edilen sekanslardan hedef ile eşleşen ya da en yakın eşleşmeye sahip sekansların varlığı araştırılmıştır. Toplam benzerlik skoru en yüksek olan yıla ait tahmin setleri seçilerek hedef ile karşılaştırılmıştır. Karşılaştırma sonucu, en yakın benzerliğe sahip sekanslar EK 3'teki çizelgelerde sunulmuştur. Çizelgelerdeki sekansların hedef ile olan farkları ve sayıları Çizelge 3.5'de bulunmaktadır.

Bu çalışmada 5 yıla kadar tahminler yapılmıştır. Eğitim ve doğrulama seti yıl aralığı fazla olan setler için 5 yıldan fazla tahmin yapılması gerekir. Ayrıca, mutasyon hızı hesaplamasında alınan referans deneysel ortamda ölçülen bir veri olduğundan, gerçek mutasyon hızını ifade etmiyor olabilir. Bu sebepten ötürü, mutasyon hızı NA amino asit dizilimi başına bir mutasyon değil, daha fazla olması gerekiyor olabilir.

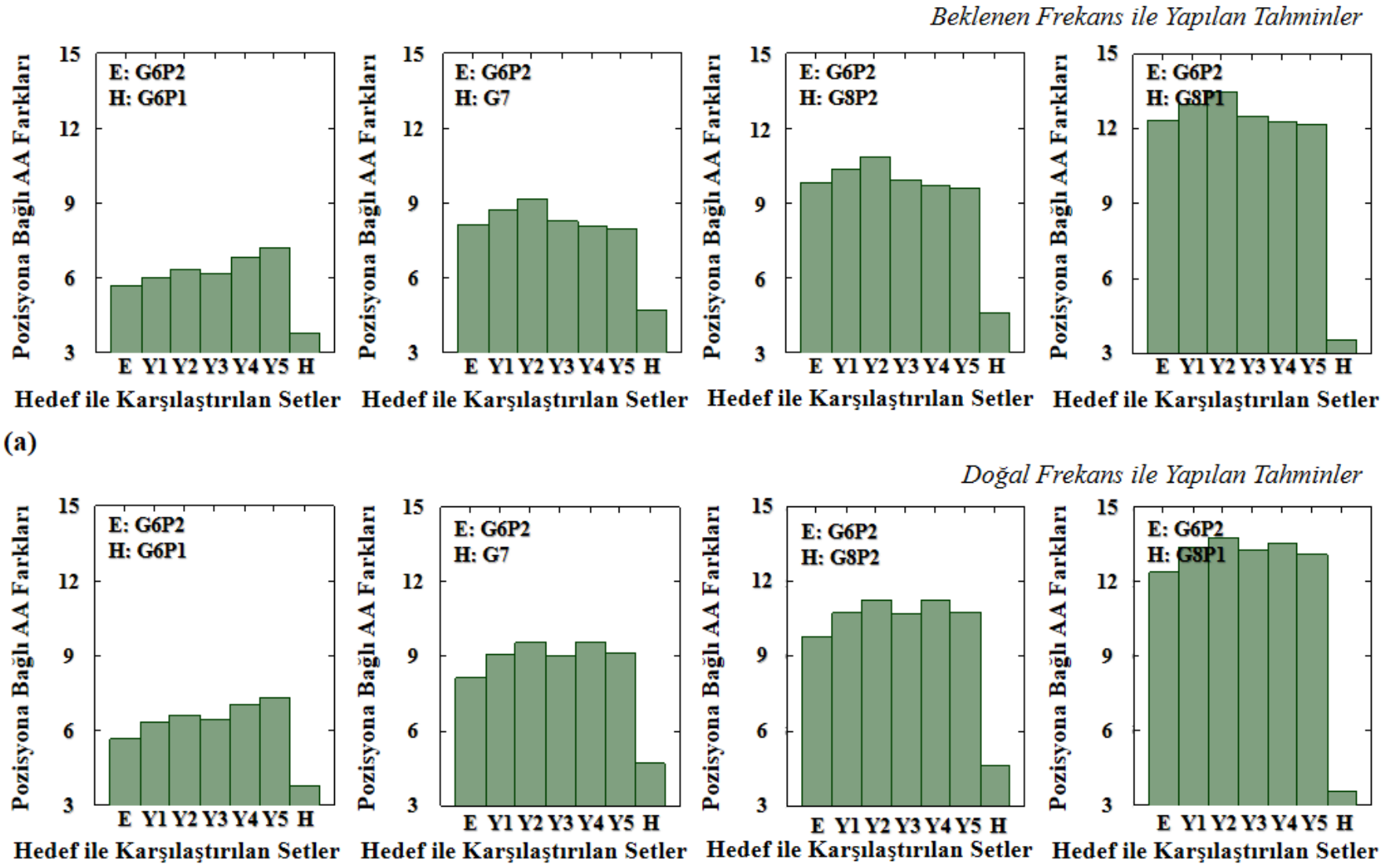
Bu tez kapsamında oluşturulmuş, tahmin modeli üzerinde sonuçlar kısmında sunulan geliştirme yöntemleri kullanılarak sistem iyileştirilebilir ve hedefe daha yakın sonuçlar elde edilebilir.

Çizelge 3.5: Tahmin doğruluğu olan sekansların bilgileri

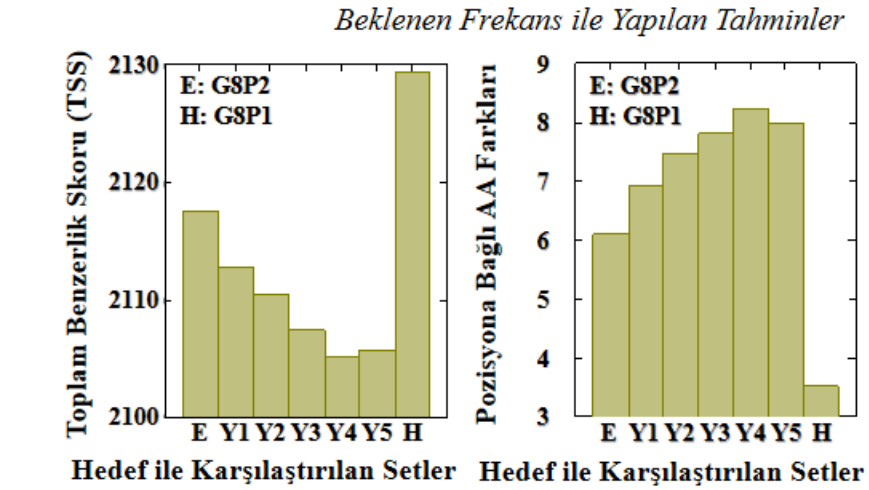
| | Eğitim G3 | | | | | |
|---|---|---|---|---|---|---|
| | Beklenen Frekans | | | Doğal Frekans | | |
| | Sekans Numarası | Minimum Fark | Sekans Sayısı | Sekans Numarası | Minimum Fark | Sekans Sayısı |
| Hedef G4 | 1 | 2 | 1 | 4 | 2 | 1 |
| Hedef G5 | 2, 3 | 7 | 2 | 5 | 7 | 1 |
| Hedef G8P2 | - | 57 | 19 | - | 57 | 16 |
| Hedef G8P1 | - | 60 | 20 | - | 60 | 15 |



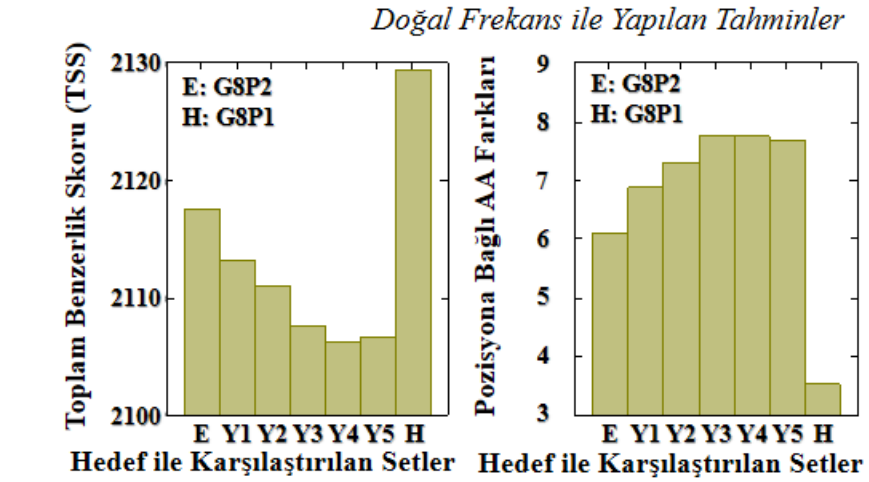

Şekil 3.29: Grup 6 Part 2 ile yapılan 5 yıllık tahminlerin ve eğitim (E) setinin, hedef (H) ile hesaplanan toplam benzerlik skorları.




Şekil 3.30: Grup 6 Part 2 ile yapılan 5 yıllık tahminlerin ve eğitim (E) setinin hedef (H) ile hesaplanan pozisyona bağlı amino asit (AA) farkları.




(a)




(b)

Şekil 3.31: Grup 8 Part 2 ile yapılan 5 yıllık tahminlerin ve eğitim (E) setinin, hedef (H) ile hesaplanan toplam benzerlik skorları ve pozisyona bağlı amino asit (AA) farkları.

Çizelge 3.5 - Devamı: Tahmin doğruluğu yüksek olan sekansların bilgileri

| | Eğitim G6P2 | | | | | |
|---|---|---|---|---|---|---|
| | Beklenen Frekans | | | Doğal Frekans | | |
| | Sekans Numarası | Minimum Fark | Sekans Sayısı | Sekans Numarası | Minimum Fark | Sekans Sayısı |
| Hedef G6P1 | 1 | 0 | 1 | 3,4 | 0 | 2 |
| Hedef G7 | 2 | 1 | 1 | 5 | 1 | 1 |
| Hedef G8P2 | 2 | 1 | 1 | 6, 7 | 2 | 2 |
| Hedef G8P1 | - | 7 | 30 | 6, 7, 8 | 7 | 3 |
| | Eğitim G8P2 | | | | | |
| | Beklenen Frekans | | | Doğal Frekans | | |
| | Sekans Numarası | Minimum Fark | Sekans Sayısı | Sekans Numarası | Minimum Fark | Sekans Sayısı |
| Hedef G8P1 | 1 | 2 | 9 | 2 | 2 | 11 |

# 4. SONUÇ VE ÖNERİLER

Biyolojik verilerin incelenip, içlerinde var olan bilginin anlaşılabilir ve kullanılabilir bir şekilde çıkarıldığı biyoenformatik-tabanlı çalışmalar, günümüzde genomiks proteomiks gibi birçok alanda kullanılmaktadır. Bu analizler sayesinde, gen dizilimlerinde hastalıklara sebep olan bölgelerin belirlenmesi ve olası varyasyonların anlaşılması önemli çalışma alanları içerisindedir.

Bu tez kapsamında, influenza virüslerinin sahip olduğu yüzey proteinlerinden biri olan NA proteininde meydana gelebilecek mutasyonların tahmini için bir model oluşturulmuştur. H1N1 virüsündeki NA proteininin evrimsel değişimi tahmin edilmeye çalışılmıştır. Oluşturulan tahmin modeli için öncelikle, NA proteinleri için var olan veri bankalarından sekans bilgileri (amino asit dizilimleri) derlenmiş, yapılan öncül araştırmalar ile eğitim ve doğrulama setleri oluşturulmuş, daha sonra da hem bu setler içerisindeki amino asit frekansları hem de NA proteinindeki her amino asit bölgesinin mutasyona uğrama olasılıkları (mutasyon skorları) hesaplanmıştır. Mutasyon skorlarının hesaplanması için hem farklı hesaplama teknikleri hem de bu teknikler içerisinde kullanılan farklı skorlama matrisleri (BLOSUM62, PAM120, GONNET ve PET91) kullanılmış ve bu matrislerin etkisi Birim matris ve rastsal oluşturulan matris ile karşılaştırılmıştır. Karşılaştırmalar sonucunda skorlama matrisleri içerisindeki verilerin (amino asitler arasındaki ilişkilerin) hangi yöntem kullanılırsa kullanılsın çok önemli olduğu gösterilmiştir. Kullanılan yöntemler içerisinde ise iki farklı ana grup oluşmuş, bir grup mutasyon bölgelerini (Karlin, Hogervorst, BNT2 ve BNT3) bulmada, diğer grup (Sander ve Valdar) ise korunan bölgeleri bulmada iyi sonuçlar vermiştir. Bu metotlar kendi aralarında karşılaştırıldıklarında aynı skorlama matrisi kullanıldığı durumlarda BNT3 dışında benzer sonuçlar vermektedirler.

Mutasyon skorlarına ek olarak, literatürdeki deneysel çalışmalardan elde edilen influenza virüsünün mutasyon hızı bilgisinden yararlanılarak NA proteininin mutasyon hızı mutasyon/yıl olarak hesaplanmıştır. Sonuç olarak 388-9 amino asitlik dizilime sahip NA proteininin yılda yaklaşık olarak bir mutasyona uğradığı

bulunmuştur. Son olarak amino asitler arası dönüşüm için amino asitlerin hem kodon tablosundan hesaplanan beklenen frekansları hem de NA proteinleri içerisindeki frekans dağılımları (doğal frekans) kullanılmıştır.

Tüm bu analizlerden elde edilen sonuçlar rastgele yürüyüş algoritması içerisinde bir araya getirilerek, eğitim setlerinden, zamanla oluşan mutasyonlar sonucu karşılaştığımız NA amino asit dizilimleri modellenmiştir. Geliştirdiğimiz mutasyon skorları, literatürde karşılaşılan ve virüse antiviral direnç kazandıran mutasyon bölgelerini büyük oranda tespit edebilmektedir. Model sonucu elde edilen sekanslar ile doğrulama setleri içerisindeki sekanslar karşılaştırılarak modellerin doğruluk derecelerinin hesaplanması için çeşitli yöntemler denenmiştir. Tahmin edilen setlerin belirli yıl aralıklarında doğrulama setine yakınsadığı gözlenmekle beraber tam olarak aynı sekanslara ulaşılamamıştır. Bu veri setleri içerisinde tahmin sonuçlarının tam olarak aynı sonuçları vermesi beklenmemekle birlikte, tahmin modelinin iyileştirilerek daha yüksek doğruluk oranlarına sahip sekansların elde edilmesi için modelin geliştirilmesi çalışmaları devam etmektedir. Burada temel problem, veri kalitesi ve verilerin düzgün bir şekilde ayrıştırılıp kullanılabilir (nerede ve ne tür mutasyonların olacağını belirleyen) özelliklerinin çıkarılmasıdır. Modelin geliştirilmesi için yapılmakta olan ve yapılması planlanan çalışmalar ileriki paragraflarda sunulmuştur.

Bu tez içerisinde, skorlama matrisleri ve skorlama fonksiyonları arasındaki ilişkiler incelenmiş ve tahmin modelinde tek bir matris ve tek bir fonksiyon seçilerek hesaplamalar yapılmıştır. İleride farklı kombinasyonlar oluşturularak tahminler arası ilişki incelenerek sistem optimize edilebilir.

Yakın sonuçlar gösteren skorlama fonksiyonlarından BNT3 fonksiyonu seçilerek evrimsel değişim tahminleri yapılmıştır. Tek bir fonksiyon yerine, yüksek korelasyona sahip fonksiyonların bir kombinasyonu oluşturularak, farklı ağırlıklarda her biri sisteme dahil edilerek fonksiyonlar arası etkinin tahmin sonuçlarına etkisi incelenebilir. Bu sayede, BNT3 fonksiyonu sonuçlarıyla gözden kaçırılan mutasyon bölgeleri, bu kombinasyon ile saptanabilir.

Tahmin modelinde, skorlama matrisleri içerisinden yaygın olarak kullanılan BLOSUM62 matrisi kullanılmıştır. Bu matris oluşturulurken kullanılan protein sekansları ile NA protein sekanslarındaki amino asit dağılımları tamamen aynı değildir. Bu durumda amino asitlerin birbirine dönüşme olasılıklarını gösteren BLOSUM62 matrisi yerine, NA proteinine özel skorlama matrisleri oluşturulabilir.

Mutasyon haritaları NA proteininin çoğu bölgesinin korunduğunu göstermektedir. Eğer bir pozisyon tamamen korunuyorsa mutasyon değeri 0'dır. Ancak çok az değişime uğrayan bölgelerin aldığı skor 0'dan farklı olacaktır. Bu az değişime uğrama olasılığı olan bölgelerin tahmin sisteminde gürültüye sebep olmaması için mutasyon haritasında bir eşik değeri belirlenerek o eşik değeri altındaki skorların mutasyona uğrama olasılıkları 0 kabul edilebilir. Bu sayede tahmin modelinde, mutasyona uğrama olasılığı yüksek olan bölgelerde değişim daha net olarak gözlemlenecektir.

Bir proteinde gözlemlenen değişim tek pozisyonda ise bu değişim etrafındaki amino asitleri etkilemektedir. Bu çalışmada her pozisyondaki değişim etrafındaki amino asitlerden bağımsız olarak gerçekleşmektedir. Amino asit değişimleri tek pozisyonda olabileceği gibi bölgesel olarak birden fazla amino asidin değişimi olarak da gözlenebilmektedir. Protein üzerinde amino asit değişimlerine komşu pozisyonların etkisinin tespit edilmesi ile gerçeğe daha yakın bir tahmin modeli oluşturulabilecektir. Tahmin edilen sekanslar için yapı tahmini yapılarak var olan ilaçların bağlanma etkinliği incelenecektir. Bu sayede tehlikeli sekanslar tespit edilebilecektir.

Oluşturulan bu model sayesinde, gelecekte karşılaşılma olasılığı bulunan mutasyonların önceden tahmin edilebilmesi için gerekli bilgi birikimine katkıda bulunulmuştur. Tahmin edilen sekanslar üzerinden yapılacak yapısal analizler ile öncelikle var olan ilaçların bu sekanslar üzerindeki etkileri belirlenebilecek ve/veya daha etkili yeni ilaç moleküllerinin tasarlanması çalışmalarına katkı sağlanabilecektir.

## KAYNAKLAR


**Abed, Y., Bouhy, X., L'Huillier, A. G., Rhéaume, C., Pizzorno, A., Retamal, M., Boivin, G. vd.** (2016). The E119D neuraminidase mutation identified in a multidrug-resistant influenza A(H1N1)pdm09 isolate severely alters viral fitness in vitro and in animal models. *Antiviral Research*, *132*, 6–12.

**Abed, Y., Goyette, N., & Boivin, G.** (2004). A reverse genetics study of resistance to neuraminidase inhibitors in an influenza A/H1N1 virus. *Antivir Ther*, *9*(4), 577–581.

**Amaro, R. E., Li, W. W., Walker, R. C., Bush, R. M., Votapka, L., & Swift, R. V.** (2011). Mechanism of 150-cavity formation in influenza neuraminidase. *Nature Communications*, *2*(1), 387–388.

**Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N.** (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, *44*, 344–350.

**Baek, Y. H., Song, M.-S., Lee, E.-Y., Kim, Y., Kim, E.-H., Park, S.-J., Choi, Y.-K. vd.** (2015). Profiling and Characterization of Influenza Virus N1 Strains Potentially Resistant to Multiple Neuraminidase Inhibitors. *Journal of Virology*, *89*(1), 287–299.

**Bar-Joseph, Z., Gifford, D. K., & Jaakkola, T. S.** (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, *17*.

**Benner, S. A., Cohen, M. A., & Gonnet, G. H.** (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Engineering*, *7*, 1323–1332.

**Berg, J. M., Tymoczko, J. L., & Stryer, L.** (2002). *Biochemistry* (5th baskı). W. H. Freeman and Company.

**Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Bourne, P. E. vd.** (2000). The Protein Data Bank, *28*(1), 235–242.

**Bloom, J. D., Gong, L. I., & Baltimore, D.** (2010). Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance, *328*(5983), 1272–1275.

**Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Schneider, M. vd.** (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *31*(1), 365–370.

**Boivin, G. (2013).** Detection and management of antiviral resistance for influenza viruses. *Influenza and other Respiratory Viruses*, *7*(SUPPL.3), 18–23.

**Bolken, T. C., & Hruby, D. E.** (2008). Discovery and Development of Antiviral Drugs for Biodefense: Experience of a Small Biotechnology Company. *NIH, Antiviral Res.*, *77*(1), 1301–1315.

**Bromberg, Y., & Rost, B.** (2007). SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, *35*(11), 3823–

3835.
**Casella, G., Fienberg, S., & Olkin, I.** (2011). *Probability and Statistics for Machine Learning*. Springer.
**Combe, M., & Sanjuán, R.** (2014). Variability in the mutation rates of RNA viruses. *Future Virology*, *9*(6), 605–615.
**Compans, R. W., & Oldstone, M. B. A.** (Ed.). (2014). *Influenza Pathogenesis and Control - Volume 1*. Springer.
**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. *Atlas of protein sequence and structure*.
**de Clercq, E.** (2012). Milestones in the discovery of antiviral agents: nucleosides and nucleotides. *Acta Pharmaceutica Sinica B*, *2*(6), 535–548.
**Durbin, R., Eddy, S., Krogh, A., & Mitchison, G.** (1998). *Biological sequence analysis Probabilistic models of proteins and nucleic acids*. Cambridge University Press.
**Edwards, D., Stajich, J., & Hansen, D.** (Ed.). (2009). *Bioinformatics Tools and Applications*. Springer.
**Englund, J. A.** (2002). Antiviral therapy of influenza. *Seminars in Pediatric Infectious Diseases*, *13*(2), 120–128.
**Fodor, A. A., & Aldrich, R. W.** (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function and Genetics*, *56*(2), 211–221.
**Gamblin, S. J., Haire, L. F., Russell, R. J., Stevens, D. J., Xiao, B., Ha, Y., Skehel, J. J. vd.** (2004). The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science*, *303*, 1838–1842.
**Garman, E., & Laver, G.** (2005). The structure, function, and inhibition of influenza virus neuraminidase. *Viral Membrane Proteins: Structure, Function, and Drug Design*, 247–267.
**Gibbs, A. J., Calisher, C. H., & Garcia-Arenal, F.** (1996). Evolutionary Virology: Molecular Basis of Virus Evolution. *Science*, *273*(August).
**Gobel, U., Sander, C., Schneider, R., & Valencia, A.** (1994). Correlated Mutations and Residue Contacts in Proteins. *Proteins: Structure, Function and Genetics*, *18*, 309–317.
**Gregory, V., Daniels, R. S., Siqueira, M. M., Hurt, A. C., Huang, W., Lackenby, A., Besselaar, T. G. vd.** (2017). Global update on the susceptibility of human influenza viruses to neuraminidase inhibitors, 2015–2016. *Antiviral Research*, *146*, 12–20.
**Gupta, M. K.** (2015). Mutations in surface protein of swine flu: A major problem for H1N1 inhibitor. *Asian Journal of Biomedical and Pharmaceutical Sciences*, *5*(50), 1–9.
**Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Neher, R. A. vd.** (2018). Nextstrain: real-time tracking of pathogen evolution. *Oxford Acamedic*.
**Hayden, F. G., & De Jong, M. D.** (2011). Emerging influenza antiviral resistance threats. *Journal of Infectious Diseases*, *203*(1), 6–10.
**Hecht, M., Bromberg, Y., & Rost, B.** (2013). News from the Protein Mutability Landscape. *Journal of Molecular Biology*, *425*(21), 1–12.
**Henikoff, S., & Henikoff, J. G.** (1992). Amino acid substitution matrices from protein

blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

**Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schaerfe, C. P. I., Springer, M., Sander, C., & Marks, D. S.** (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, *35*(2), 128–135.

**Huang, L., Cao, Y., Zhou, J., Qin, K., Zhu, W., Zhu, Y., Shu, Y. vd.** (2014). A conformational restriction in the influenza a virus neuraminidase binding site by R152 results in a combinational effect of I222T and H274Y on oseltamivir resistance. *Antimicrobial Agents and Chemotherapy*, *58*(3), 1639–1645.

**Isin, B., Doruker, P., & Bahar, I.** (2002). Functional motions of influenza virus hemagglutinin: A structure-based analytical approach. *Biophysical Journal*, *82*(2), 569–581.

**Jagadesh, A., Salam, A. A. A., Mudgal, P. P., & Arunkumar, G.** (2016). Influenza virus neuraminidase (NA): a target for antivirals and vaccines. *Archives of Virology*, *161*(8), 2087–2094.

**Jefferson, T., Jones, M., Doshi, P., & Del Mar, C.** (2009). Neuraminidase inhibitors for preventing and treating influenza in healthy adults: systematic review and meta-analysis. *Bmj*, *339*, 1–8.

**Jones, D. T., Taylor, W. R., & Thornton, J. M.** (1992). The rapid generation of mutation data matricies from protein sequences. *Computer Applications in the Biosciences*, *8*(3), 275–282.

**Kawaoka, Y., & Neumann, G.** (2012). *Influenza Virus Methods and Protocols*. Methods in Molecular Biology, Springer Protocols.

**Keith, J. M.** (Ed.). (2008). *Bioinformatics Volume I Data, Sequeunce Analysis and Evolution*. Humana Press, Springer.

**Klebe, G.** (2013). *Drug Design Methodology, Concepts, and Mode-of-Action*. Springer Reference.

**Knobler, S., Mack, A., Mahmoud, A., & Lemon, S.** (2005). The Story of Influenza". The Threat of Pandemic Influenza: Are We Ready? *National Academies Press*, 60–61.

**Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., & Ben-Tal, N.** (2005). ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research*, *33*, 299–302.

**Lauring, A. S., Frydman, J., & Andino, R.** (2013). The role of mutational robustness in RNA virus evolution. *Nature Reviews Microbiology*, *11*(5), 327–336.

**Lipman, D. J., Wilbur, W. J., Smith, T. F., & Waterman, M. S.** (1984). On the statistical significance of nucleic acid similarities. *Nucleic acids research*, *12*, 215–226.

**Ma, X., Meng, H., & Lai, L.** (2016). Motions of Allosteric and Orthosteric Ligand-Binding Sites in Proteins are Highly Correlated. *Journal of Chemical Information and Modeling*, *56*(9), 1725–1733.

**Mathura, V. S., & Kangueane, P.** (Ed.). (2009). *Bioinformatics: A Concept-Based Introduction*. Springer.

**Mckimm-Breschkin, J. L.** (2013). Influenza neuraminidase inhibitors: Antiviral action and mechanisms of resistance. *Influenza and other Respiratory Viruses*, *7*(1 SUPPL.1), 25–36.

**Mihajlovic, M. L., & Mitrasinovic, P. M.** (2008). Another look at the molecular

mechanism of the resistance of H5N1 influenza A virus neuraminidase (NA) to oseltamivir (OTV). *Biophysical Chemistry*, *136*(2–3), 152–158.

**Miller, M. P., & Kumar, S.** (2001). Understanding human disease mutations through the use of interspecific genetic variation. *Human Molecular Genetics*, *10*(21), 2319–2328.

**Monto, A. S., McKimm-Breschkin, J. L., Macken, C., Hampson, A. W., Hay, A., Klimov, A., Zambon, M. vd.** (2006). Detection of influenza viruses resistant to neuraminidase inhibitors in global surveillance during the first 3 years of their use. *Antimicrobial Agents and Chemotherapy*, *50*(7), 2395–2402.

**Needleman, S. B., & Wunsch, C. D.** (1970). A general method applicable to search for similarities in amino acid sequence of 2 proteins. *J. Mol. Biol.*, *48*, 443–453.

**Neher, R. A., & Bedford, T.** (2015). nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, *31*(21), 3546–3548.

**Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A., & Shraiman, B. I.** (2016). Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences*, *113*(12), E1701–E1709.

**Ng, P. C., & Henikoff, S.** (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Research*, *11*, 863–874.

**Nisn.** (2010). An Overview of Antiviral Drug Resistance Data presented at Options for the Control of Influenza VII. *Nisn*, (September).

**Okomo-Adhiambo, M., Nguyen, H. T., Sleeman, K., Sheu, T. G., Deyde, V. M., Garten, R. J., Gubareva, L. V. vd.** (2010). Host cell selection of influenza neuraminidase variants: Implications for drug resistance monitoring in A(H1N1) viruses. *Antiviral Research*, *85*(2), 381–388.

**Oren, E. E., Tamerler, C., Sahin, D., Hnilova, M., Ozgur, U., Seker, S., Samudrala, R. vd.** (2007). Sequence analysis A novel knowledge-based approach to design inorganic-binding peptides, *23*(21), 2816–2822.

**Pearson, W. R., & Lipman, D. J.** (1988). Improved tools for biological sequence comparison, *85*, 2444–2448.

**Petrova, V. N., & Russell, C. A.** (2018). The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, *16*(1), 47–60.

**Pizzorno, A., Abed, Y., Rhéaume, C., Bouhy, X., & Boivin, G.** (2013). Evaluation of recombinant 2009 pandemic influenza A (H1N1) viruses harboring zanamivir resistance mutations in mice and ferrets. *Antimicrobial Agents and Chemotherapy*, *57*(4), 1784–1789.

**Pizzorno, A., Bouhy, X., Abed, Y., & Boivin, G.** (2011). Generation and Characterization of Recombinant Influenza A(H1N1) Viruses Resistant to Neuraminidase Inhibitors. *The Journal of Infectious Diseases*, *203*(6), 25–31.

**Plant, E. P., & Ye, Z.** (2013). Gene Constellation of Influenza Vaccine Seed Viruses. *Veterinary antibiotics in the enviroment*, 135–152.

**Rappuoli, R., & Giudice, G. Del.** (2011). *Influenza Vaccines for the Future* (2nd editio). Springer.

**Sanjuán, R., & Domingo-Calap, P.** (2016). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, *73*(23), 4433–4448.

**Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R.** (2010).

Viral Mutation Rates. *Journal of Virology*, *84*(19), 9733–9748.

**Schuchat, A., Tappero, J., & Blandford, J.** (2014). Global health and the US Centers for Disease Control and Prevention. *The Lancet*, *384*, 98–101.

**Shi, Y., Wu, Y., Zhang, W., Qi, J., & Gao, G. F.** (2014). Enabling the "host jump": Structural determinants of receptor-binding specificity in influenza A viruses. *Nature Reviews Microbiology*, *12*, 822–831.

**Shtyrya, Y. A., Mochalova, L. V, & Bovin, N. V.** (2009). Influenza virus neuraminidase: structure and function. *Acta naturae*, *1*(2), 26–32.

**Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C.** (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*, 452–457.

**Takashita, E., Meijer, A., Lackenby, A., Gubareva, L., Rebelo-De-Andrade, H., Besselaar, T., Tashiro, M. vd.** (2015). Global update on the susceptibility of human influenza viruses to neuraminidase inhibitors, 2013-2014. *Antiviral Research*, *117*, 27–38.

**Thompson, J. D., Higgins, D. G., & Gibson, T. J.** (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.

**Thusberg, J., & Vihinen, M.** (2009). Pathogenic or not? and if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Human Mutation*, *30*(5), 703–714.

**Tu, V., Abed, Y., Barbeau, X., Carbonneau, J., Fage, C., Lagüe, P., & Boivin, G.** (2017). The I427T neuraminidase (NA) substitution, located outside the NA active site of an influenza A(H1N1)pdm09 variant with reduced susceptibility to NA inhibitors, alters NA properties and impairs viral fitness. *Antiviral Research*, *137*, 6–13.

**Valdar, W. S. J.** (2002). Scoring residue conservation. *Proteins: Structure, Function and Genetics*, *48*(2), 227–241.

**van der Meer, J. Y., Biewenga, L., & Poelarends, G. J.** (2016). The Generation and Exploitation of Protein Mutability Landscapes for Enzyme Engineering. *ChemBioChem*, *17*(19), 1792–1799.

**Visher, E., Whitefield, S. E., McCrone, J. T., Fitzsimmons, W., & Lauring, A. S.** (2016). The Mutational Robustness of Influenza A Virus. *PLoS Pathogens*, *12*(8), 1–25.

**von Itzstein, M.** (2012). *Influenza Virus Sialidase-A Drug Discovery Target*. (M. J. Parnham & J. Bruinvels, Ed.). Springer.

**Wargo, A. R., & Kurath, G.** (2012). Viral fitness: Definitions, measurement, and current insights. *Current Opinion in Virology*, *2*(5), 538–545.

**Wilson, C. A., Kreychman, J., & Gerstein, M.** (2000). Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology*, *297*, 233–249.

**Wu, N. C., Young, A. P., Dandekar, S., Wijersuriya, H., Al-Mawsawi, L. Q., Wu, T.-T., & Sun, R.** (2013). Systematic Identification of H274Y Compensatory Mutations in Influenza A Virus Neuraminidase by High-Throughput Screening. *Journal of Virology*, *87*(2), 1193–1199.

**Xu, R., Ekiert, D., Krause, J., Hai, R., Crowe, J., & Wilson, I.** (2010). Structural Basis of Preexisting Immunity to the 2009 H1N1 Pandemic Influenza

Virus, *328*(5976), 357–360.

**Xu, Y., Xu, D., & Liang, J.** (Ed.). (2007). *Computational Methods for Protein Structure Prediction and Modeling Volume 1: Basic Characterization*. Springer.

**Yang, Z., & Rannala, B.** (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, *13*(5), 303–314.

**Yongkiettrakul, S., Nivitchanyong, T., Pannengpetch, S., Wanitchang, A., Jongkaewwattana, A., & Srimanote, P.** (2013). Neuraminidase amino acids 149 and 347 determine the infectivity and oseltamivir sensitivity of pandemic influenza A/H1N1 (2009) and avian influenza A/H5N1. *Virus Research*, *175*(2), 128–133.

**Url-1** https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/what-genomics alındığı tarih: 03.05.2018

**Url-2** https://www.ebi.ac.uk/uniprot/TrEMBLstats alındığı tarih: 07.05.2018

**Url-3** www.fludb.org alındığı tarih: 08.05.2018

**Url-4** https://www.fda.gov/AboutFDA/WhatWeDo/History/ucm304485.htm alındığı tarih: 11.05.2018

**Url-5** http://www.titck.gov.tr/Denetim/IlacDenetim alındığı tarih: 11.05.2018

**Url-6** https://www.fda.gov/ForPatients/Approvals/Drugs/default.htm alındığı tarih: 11.05.2018

**Url-7** https://www.cdc.gov/flu/about/viruses/change.htm alındığı tarih: 11.05.2018

**Url-8** https://www.cdc.gov/media/releases/2017/p1213-flu-death-estimate.html alındığı tarih: 11.05.2018

**Url-9** https://jgi.doe.gov/proving-codon-genetic-code-flexibility/ alındığı tarih: 20.05.18

**EKLER**

EK 1: Skorlama Matrisleri

EK 2: Amino Asit Tablosu

EK 3: Tahmin Doğruluğu Yüksek Olan Sekanslar

**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 | -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 |

**EK 2**

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 | |

Çizelge Ek 6: Grup 3 ile yapılan tahminlerden hedeflere en yakın olanlar

1. SVTLAGNSSLCSISGWAIYTKDNSIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRALMSCPLGEAPSP
   YNSKFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKYNGIITGTIKSWKKQILRTQESECVCMNGSCFTIMTDGPSNGAASYKIF
   KIEKGKVTKSIELNAPNFHYEECSCYPDTGTVMCVCRDNWHGSNRPWVSFNQNLDYQIGYICSGVFGDNPRPKDGEGSCNPVTV
   DGADGVKGFSYKYGNGVWIGRTKSNRLRKGFEMIWDPNGWTNTDSDFSVKQDVVAITDWSGYSGSFVQHPELTGLDCIRPCFW
   VELVRGLPRENTTIWTSGSSISFCGVNSDTANWSWPDGAELPFTIDK
2. SVTLAGNSSLCSISGWAIYTKDNSIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRALMSCPLGEAPSP
   YNSKFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKYNGIITETIKSWKKKILRTQESECVCMHGSCFTIMTDGPSNGAASYKIF
   KIEKGKVTKSIELNAPNFHYEECSCYPDTGTVMCVCRDNWHGSNRPWVSFNQNLDYQIGYICSGVFGDNPRPKDGEGSCNPVTV
   DGADGVKGFSYKYGNGVWIGRTKSNRLRKGFEMIWDPNGWTNTDSDFSVKQDVVAITDWSGYSGSFVQHPELTGLDCIRPCFW
   VELVRGLPRENTTIWTSGSSISFCGVNSETANWSWPDGAELPFTIDK
3. SVTLAGNSSLCSISGWAIYTKDNSIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRALMSCPLGEAPSP
   YNSKFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKYNGIITSTIKSWKKQILRTQESECVCMNGSCFTIMTDGPSNGAASYKIF
   KIEKGKVTKSIELNAPNFHYEECSCYPDTGTVMCVCRDNWHGSNRPWVSFNQNLDYQPGYICSGVFGDNPRPKDGEGSCNPVTV
   DGADGVKGFSYKYGNGVWIGRTKSNRLRKGFEMIWDPNGWTNTDSDFSVKQDVVAITDWSGYSGSFVQHPELTGLDCIRPCFW
   VELVRGLPRENTTIWTSGSSISFCGVNSDTANWSWPDGAELPFTIDK
4. SVTLAGNSSLCSISGWAIYTKDNSIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRALMSCPLGEAPSP
   YNSKFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKYNGIITGTIKSWKKQILRTQESECVCMNGSCFTIMTDGPSNGAASYKIF
   KIEKGKVTKSIELNAPNFHYEECSCYPDTGTVMCVCRDNWHGSNRPWVSFNQNLDYQIGYICSGVFGDNPRPKDGEGSCNPVTV
   DGADGVKGFSYKYGNGVWIGRTKSNRLRKGFEMIWDPNGWTNTDSDFSVKQDVVAITDWSGYSGSFVQHPELTGLDCIRPCFW
   VELVRGLPRENTTIWTSGSSISFCGVNSDTANWSWPDGAELPFTIDK
5. SVTLAGNSSLCSISGWAIYTKDNSIRIGSKGDVFVIREPFISCSHLECKTFFLTQGALLNDKHSNGTVKDRSPYRALMSCPLGEAPSP
   YNSKFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKYNGIITGTIKSWKKQILRTQESECVCMNGSCFTIMTDGPSNGAASYKIF
   KIEKGKVTKSIELNAPNFHYEECSCYPDTGTVMCVCRDNWHGSNRPWVSFNQNLDYQIGYICSGVFGDNPRPKDGEGSCNPVTV
   DGADGVKGFSYKYGNGVWIGRTKSNRLRKGFEMIWDPNGWTNTDSDFSVKQDVVAITDWSGYSGSFVQHPELTGLDCIRPCFW
   VELVRGLPRENTTIWTSGSSISFCGVNSDTANWSWPDGAELPFTIDK

Çizelge Ek 7: Grup 6 Part 2 ile yapılan tahminlerden hedefe en yakın olanlar

1. SVKLAGNSSLCPVSGWAIYSKDNSIRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
   NSRFESVAWSASACHDGINWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTVMTDGPSDGQASYKIFRI
   EKGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGA
   NGVKGFSFKYGNGVWIGRTKSISSRSGFEMIWDPNGWTGTDNNFSIKQDIVGINEWSGYSGSFVQHPELTGLDCIRPCFWVELIRG
   RPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
2. SVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
   NSRFESVAWSASACHDGINWLTIGISGPDSGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIMTDGPSDGQASYKIFRIE
   KGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGAN
   GVKGFSFKYGNGVWIGRTKSISSRKGFEMIWDPNGWTGTDNNFSIKQDIVGINEWSGYSGSFVQHPELTGLDCIRPCFWVELIRGR
   PKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
3. SVKLAGNSSLCPVSGWAIYSKDNSIRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
   NSRFESVAWSASACHDGINWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTVMTDGPSDGQASYKIFRI
   EKGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGA
   NGVKGFSFKYGNGVWIGRTKSISSRTGFEMIWDPNGWTGTDNNFSIKQDIVGINEWSGYSGSFVQHPELTGLDCIRPCFWVELIRG
   RPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
4. SVKLAGNSSLCPVSGWAIYSKDNSIRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
   NSRFESVAWSASACHDGINWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTVMTDGPSDGQASYKIFRI
   EKGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGA
   NGVKGFSFKYGNGVWIGRTKSISSRSGFEMIWDPNGWTGTDNNFSIKQDIVGINEWSGYSGSFVQHPELTGLDCIRPCFWVELIRG
   RPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
5. SVKLAGNSSLCPVSGWAIYSKDNSIRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
   NSRFESVAWSASACHDGINWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIMTDGPSDGQASYKIFRI
   EKGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGA
   NGVKGFSFKYGNGVWIGRTKSISSRKGFEMIWDPNGWTGTDNNFSIKQDIVGINEWSGYSGSFVQHPELTGLDCIRPCFWVELIRG
   RPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK

6. SVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
NSRFESVAWSASACHDGINWLTIGISGPDSGAVAVLKYNGIITDTIKSWRNDILRTQESECACVNGSCFTIMTDGPSDGQASYKIFRIE
KGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGAN
GVKGFSFKYGNGVWIGRTKSISSRKGFEMIWDPNGWTGTDNNFSIKQDIVGINEWTGYSGSFVQHPELTGLDCIRPCFWVELIRGR
PEENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
7. SVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
NSRFESVAWSASACHDGINWLTIGISGPDSGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIMTDGPSDGQASYKIFRIE
KGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGAN
GVKGFSFKYGNGVWIGRTKSISSRKGFEMIWDPNGWTGTDNNFSIKQDIVGINEWTGYSGSFVQHPELTGLDCIRPCFWVELIRGR
PKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
8. SVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
NSRFESVAWSASACHDGINWLTIGISGPDSGAVAVLKYNGIITDTIKSWRNDILRTQESECACVNGSCFTIMTDGPSDGQASYKIFRIE
KGKIVKSVEMNAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQIGYICSGIFGDNPRPNDKTGSCGPVSSNGAN
GVKGFSFKYGNGVWIGRTKSISSRKGFEMIWDPNGWTGTDNKFSIKQDIVGINEWTGYSGSFVQHPELTGLDCIRPCFWVELIRGR
PKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK

Çizelge Ek 8: Grup 8 Part 2 ile yapılan tahminlerden hedefe en yakın olanlar

1. SVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
NSRFESVAWSASACHDGINWLTIGISGPDSGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIMTDGPSDGQASYKIFRIE
KGKIIKSVEMKAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQMGYICSGVFGDNPRPNDKTGSCGPVSSNGA
NGVKGFSFKYGNGVWIGRTKSISSRKGFEMVWDPNGWTGTDNKFSIKQDIVGKNEWSGYSGSFVQHPELTGLDCIRPCFWVELIR
GRPEENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK
2. SVKLAGNSSLCPVSGWAIYSKDNSVRIGSKGDVFVIREPFISCSPLECRTFFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPY
NSRFESVAWSASACHDGINWLTIGISGPDSGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIMTDGPSDGQASYKIFRIE
KGKIIKSVEMKAPNYHYEECSCYPDSSEITCVCRDNWHGSNRPWVSFNQNLEYQMGYICSGVFGDNPRPNDKTGSCGPVSSSGAN
GVKGFSFKYGNGVWIGRTKSISSRKGFEMVWDPNGWTGTDNKFSIKQDIVGINEWSGYSGSFVQHPELTGLDCIRPCFWVELIRG
RPEENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFTIDK

## ÖZGEÇMİŞ


**Ad-Soyad** : Elif CANDAŞ

**Uyruğu** : T.C.

**Doğum Tarihi ve Yeri** : 18.03.1993/ İSTANBUL

**E-posta** : e.candas@etu.edu.tr / candaselif@gmail.com


### ÖĞRENİM DURUMU:

- **Yüksek Lisans :** 2019, TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Biyomedikal Mühendisliği (Tam Burslu, 3.71/4.00)
- **Lisans :** 2016, TOBB Ekonomi ve Teknoloji Üniversitesi, Mühendislik Fakültesi, Biyomedikal Mühendisliği (Yarı Burslu, 3.36/4.00)

  2016, TOBB Ekonomi ve Teknoloji Üniversitesi, Mühendislik Fakültesi, Elektrik ve Elektronik Mühendisliği (Yan Dal, 2.67/4.00)

### MESLEKİ DENEYİM VE ÖDÜLLER:

| Yıl | Yer | Görev |
|---|---|---|
| 2016-2019 | TOBB Ekonomi ve Teknoloji Üniversitesi | Tam Burslu Yüksek Lisans Öğrencisi |

**YABANCI DİL:** İngilizce

### TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

**Candas E.**, Oren E.E., 2017. Calculation of Protein Mutability Landscape and Thereon Forecasting Evolutionary Pathways: Neuraminidase of H1N1 Virus as a Case Study, 5th International BAU Drug Design Congress, October 19-21, Istanbul, Turkey.

Gokce G., **Candas E**., Aydin N. S., Oren E. E., 2016. Forecasting Antiviral Drug Resistance Development Among Influenza Viruses, XXV International Materials Research Congress, Symposium A.2. Bionanodesign, August 14-19, Cancun, Mexico.

**DİĞER YAYINLAR, SUNUMLAR VE PATENTLER:**

Erdogan H., Babur E., Yilmaz M., **Candas E.**, Goerdesel M., Dede Y., Oren E.E., Demirel G., Ozturk M. & Demirel G., 2015. Morphological versatility in self-assembly of Val-Ala and Ala-Val dipeptides, *Langmuir*, 31, 7337-7345.

**Candas E.**, Gokce G., Demir B., Demirel G., Oren E. E., 2015. Modeling of Morphological Versatility in Self-Assembly of Val-Ala and Ala-Val Dipeptides 2015 MRS Fall Meeting & Exhibit, Symposium WW- Modeling and Theory-Driven Design of Soft Materials, November 29 – December 4, Boston, MA, USA.