T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KONAK-PATOJEN PROTEİN ETKİLEŞİMİNİN HESAPLAMALI YÖNTEMLER İLE TAHMİNİ

DOKTORA TEZİ

İrfan KÖSESOY

Enstitü Anabilim Dalı	**:**	**BİLGİSAYAR VE BİLİŞİM MÜHENDİSLİĞİ**
Tez Danışmanı	**:**	**Prof. Dr. Cemil ÖZ**
Ortak Danışman	**:**	**Doç. Dr. Murat GÖK**

Ekim 2018

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KONAK-PATOJEN PROTEİN ETKİLEŞİMİNİN HESAPLAMALI YÖNTEMLER İLE TAHMİNİ

DOKTORA TEZİ

İrfan KÖSESOY

Enstitü Anabilim Dalı : **BİLGİSAYAR MÜHENDİSLİĞİ**

Bu tez 19/10/2018 tarihinde aşağıdaki jüri tarafından oybirliği / oyçokluğu ile kabul edilmiştir.

Prof. Dr. Cemil ÖZ **Jüri Başkanı**	**Doç. Dr. Semra BORAN** **Üye**	**Dr. Öğr. Üyesi Seçkin ARI** **Üye**
Doç. Dr. Müfit Çetin **Üye**		**Dr. Öğr. Üyesi Osman Hilmi KOÇAL** **Üye**

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

İrfan KÖSESOY

19.10.2018

TEŞEKKÜR

Doktora eğitimim boyunca değerli bilgi ve deneyimlerinden yararlandığım, her konuda bilgi ve desteğini almaktan çekinmediğim, araştırmanın planlanmasından yazılmasına kadar tüm aşamalarında yardımlarını esirgemeyen, teşvik eden, aynı titizlikte beni yönlendiren değerli danışman hocalarım Prof. Dr. Cemil Öz ve Doç. Dr. Murat Gök'e, çalışmalarım sırasında tavsiyelerinden istifade ettiğim Prof. Dr. Tamer Kahveci'ye teşekkürlerimi sunarım.

Eğitim hayatım boyunca maddi ve manevi desteklerini esirgemeyen başta annem ve babam olmak üzere, tüm aile bireylerime şükranlarımı sunarım.

İÇİNDEKİLER

TEŞEKKÜR i
İÇİNDEKİLER ii
SİMGELER VE KISALTMALAR LİSTESİ iv
ŞEKİLLER LİSTESİ vi
TABLOLAR LİSTESİ vii
ÖZET viii
SUMMARY ix

BÖLÜM 1.
GİRİŞ 1

BÖLÜM 2.
LİTERATÜR ÖZETİ 7

BÖLÜM 3.
MATERYAL VE YÖNTEM 12
3.1. Biyolojik Ağlar ve Çevrimiçi Veri Tabanları 12
3.2. Deneylerde Kullanılan Tahmin Yöntemleri 15
3.2.1. Matris faktorizasyonu 15
3.2.2. Naif bayes 17
3.2.3. Bayes ağları 18
3.2.4. C4.5 19
3.2.5. Rastsal orman 20
3.2.6. k-en yakın komşuluk 21
3.2.7. K* 22
3.3. Öznitelik Çıkarım Yöntemleri 23

3.3.1. Amino asit kompozisyon 24
3.3.2. Amino asit çifti 24
3.3.3. Kompozisyon moment vektörü 25
3.3.4. Bitişik üçlü 26
3.3.5. Kompozisyon, geçiş, dağılım 26
3.3.6. Dipeptit kompozisyon 27
3.3.7. Ortonormal kodlama 28
3.3.8. Taylor venn diyagramı 29
3.3.9. OETMAP 30
3.3.10. Amino asit aşleme modeli 31

BÖLÜM 4.
ÖNERİLEN YÖNTEMLER VE PROSES YAZILIMI 33
4.1. Genişletilmiş Ağ Modeli 33
4.2. Lokasyon Tabanlı Kodlama 36
4.3. PROSES 40
4.3.1. PROSES modülleri 42

BÖLÜM 5.
ARAŞTIRMA BULGULARI 46
5.1. Değerlendirme Metrikleri 46
5.2. Deneysel Çalışma 1 50
5.2.1. Veriseti 50
5.2.2. Veri setlerinin ayrık değerlendirilmesi 52
5.2.3. Çoklu veri seti ile yapılan tahmin değerlendirmesi 57
5.2.4. Özgüllük-duyarlılık grafikleri 59
5.3. Deneysel Çalışma 2 62
5.3.1. Veri seti 63
5.3.2. Bacillus anthracis veri setine ait sonuçlar 64
5.3.3. Yersinia pestis veri setine ait sonuçlar 67

BÖLÜM 6.
TARTIŞMA VE SONUÇ 71

KAYNAKLAR 74
ÖZGEÇMİŞ 82

SİMGELER VE KISALTMALAR LİSTESİ

AAC	: Amino acid composition
AAP	: Amino acid pair
AUC	: Area under curve
BN	: Bayesian Network
CMV	: Composition moment vector
CT	: Conjoint triad
CTD	: Composition, Transition, Distribution
DC	: Dipeptide Composition
DN	: Doğru negatif
DP	: Doğru pozitif
GAM	: Genişletilmiş ağ modeli
HH	: Host-host
Knn	: k-Nearest Neigborhood
LTK	: Lokasyon tabanlı kodlama
MCC	: Matthews correlation coefficient
MF	: Matris fakstorizasyonu
NB	: Naive bayes
OE	: Orthonormal encoding
PHISTO	: Pathogen-host interaction search tool
PKE	: Patojen-konak etkileşimi
PPE	: Protein-protein etkileşimi
PROSES	: Protein sequence based encoding system
RCM	: Residue-couple model
RF	: Random forest
STRING	: Search tool for the retrieval of interacting genes/proteins
SVM	: Support vector machine

TVD	: Taylor's venn diagram
YN	: Yanlış negatif
YP	: Yanlış pozitif

ŞEKİLLER LİSTESİ

Şekil 3.1. String veri tabanından insana ait proteinler için oluşturulmuş örnek etkileşim ağı. 13
Şekil 3.2. Taylor's Venndiyagram 30
Şekil 3.3. k = 3 değeri için örnek bir dizilimin birinci, ikinci ve üçüncü dereceden ranklara göre aminoasit çiftlerinin seçilmesi. 32
Şekil 4.1. Tek tür için genişletilmiş ilişki matrisi 36
Şekil 4.2. Çoklu veri setleri için genişletilen ilişki matrisleri 36
Şekil 4.3. Verilen bir protein dizisi için öznitelik çıkarım örneği 39
Şekil 4.4. PKE tahmininde kullanılacak nihai öz nitelik vektörü çıkarma örneği 40
Şekil 4.5. PROSES modülleri ve modüller arası ilişki diyagramı 43
Şekil 4.6. Protein kodlama modülüne ait akış diyagramı 44
Şekil 4.7. Arama modülüne ait akış diyagramı 45
Şekil 4.8. Dosya dönüştürme modülüne ait akış diyagram 45
Şekil 5.1. Karmaşıklık Matrisi 47
Şekil 5.2. Çapraz doğrulama yöntemi ile veri setinin test ve eğitim verisi olarak ayrılıp sonucun değerlendirilmesi. 50
Şekil 5.3.Matris faktorizasyon tahmin yöntemi için bulunan özgüllük-duyarlılık grafikleri 61
Şekil 5.4. Bacillus veri seti için bulunan MCC sonuçlarının kodlama ve tahmin yöntemlerine göre kıyaslanması 66
Şekil 5.5. Yersinia veri seti için bulunan MCC sonuçlarının kodlama ve tahmin yöntemlerine göre kıyaslanması 69

TABLOLAR LİSTESİ

Tablo 2.1. Deneysel etkileşim yöntemleri, deney ortamları, etkileşim tipi ve ilgili referans tablosu. 8
Tablo 3.1. Literatürde en sık geçen PPE ve PKE veri tabanları 14
Tablo 3.2 Yöntemlere ait öznitelik vektör uzunluğu ve ilgili referans. 23
Tablo 3.3. Amino asitlerin kimyasal özelliklerine göre gruplanması 28
Tablo 3.4. TVD kodlama yönteminde her bir amino aside ait nümerik vektör..... 31
Tablo 4.1. L = 5 değeri için alt diziye ayırma örneği..... 38
Tablo 5.1. Veri setinde olan proteinler ve ağ içindeki etkileşim sayıları..... 51
Tablo 5.2. Bacillus Anthracis veri seti için deneysel sonuçlar 53
Tablo 5.3. Ebola veri seti için deneysel sonuçlar..... 54
Tablo 5.4. Birleştirilmiş veri seti için değerlendirme sonuçları..... 59
Tablo 5.5. LTK yönteminin değerlendirilmesinde kullanılan veri setlerine ait bilgiler..... 64
Tablo 5.6. Bacillus anthracis veri seti için bulunan değerlendirme sonuçları 65
Tablo 5.7. Bacillus veri seti için LTK ile diğer kodlama yöntemlerinin deney sayısı üstünlüğüne göre kıyaslanması 67
Tablo 5.8. Yersinia pestis veri seti için bulunan değerlendirme sonuçları 69
Tablo 5.9. Yersinia veri seti için LTK kodlama ile diğer kodlama yöntemleri arasında yapılan kıyaslamada başarılı olunan deney sayısı..... 70

ÖZET

Anahtar kelimeler: Protein etkileşimleri, patojen-konak etkileşimleri, makine öğrenmesi, hesaplamalı yöntemler

Türler arası patojen-konak protein etkileşimlerin bilinmesi enfeksiyonel hastalıkların teşhis ve tedavisi için geliştirilecek çözüm stratejileri açısından hayati öneme sahiptir. Etkileşim tespitinde kullanılan deneysel yöntemlerin maliyetli olması ve uzun zaman almasından dolayı proteinler arası etkileşimlerin modellendiği hesaplamalı yöntemlerin bu alanda önemli bir yeri vardır. Hesaplamalı yöntemler, tespit süresinin kısaltılması ve maliyetin düşürülmesine ek olarak deneysel yöntemlerle yanlış tespit edilen etkileşimlerin kontrolünde de kullanılmaktadır.

Veri seyrekliği, veri yetersizliği ve doğrulanmış negatif veri setinin olmaması, patojen-konak protein etkileşim tahmini için kullanılan hesaplamalı yöntemlerin ortak problemidir. Bu çalışmada amaç patojen-konak etkileşim tahmin doğruluğunu arttırmak ve veri yetersizliğinden kaynaklanan olumsuzlukları gidermektir. Bu kapsamda genişletilmiş ağ modeli ve lokasyon tabanlı kodlama yöntemleri önerildi. Genişletilmiş ağ modeli "türler arası yeterli etkileşim verisinin olmadığı patojen konak etkileşimleri ile patojen ve konak proteinlere ait tür içi etkileşimlerin entegre edilmesi tahmin doğruluğunu arttırır" hipotezinden esinlenerek geliştirildi. Lokasyon tabanlı kodlama, proteinlerin amino asit diziliminin kodlandığı bir öznitelik çıkarım yöntemidir. Makine öğrenmesi algoritmalarında patojen konak etkileşim tahmininde başarımı etkileyen faktörlerden biri kullanılan özniteliklerdir. Biyolojik veri tabanlarında proteinlere ait en fazla veri amino asit dizilim bilgisidir. Sadece amino asit dizilimini baz alarak geliştirilen güçlü bir öznitelik çıkarım yöntemi, patojen konak etkileşim tahmin doğruluğunu arttıracaktır. Ayrıca amino asit dizilim bilgisinin kullanılması sayesinde bilinen tüm etkileşimler için öznitelik vektörlerinin daha kolay çıkarılması sağlanır.

Tezde protein kodlama ve protein etkileşim tahmini üzerine çalışan araştırmacıların kullanılabileceği, ücretsiz erişilebilen, kullanıcı dostu bir ara yüze sahip web tabanlı PROSES (Protein Sequencebased encoding system) yazılımı geliştirildi. Yazılım özellikle programlama bilgisi olmayan kişiler için faydalıdır. PROSES şu anda Yalova Üniversitesi web sunucusunda yer alan http://proses.yalova.edu.tr adresinde kullanılmaktadır.

PREDICTION OF HOST-PATHOGEN PROTEIN INTERACTIONS BY COMPUTATIONAL METHODS

SUMMARY

Keywords: Protein interactions, pathogen-host interactions, machine learning, computational methods

Knowledge of the pathogen-host protein interactions in the inter species has a vital prospect for a solution strategy to be developed against diagnosis and treatment of infectious diseases. Modeling interactions between proteins has necessitated the development of computational methods in this field, since detection of interactions by experimental methods is both time-consuming and costly. Computational methods are used in decreasing of the detection time and cost; in addition checking of the false detected interactions via experimental methods.

Data scarcity, data inadequacy, and negative data sampling are the common problems of computational methods for used in prediction of pathogen-host protein interaction. In this study, the purpose is that prediction accuracy of the pathogen-host interaction increase and negativeness eliminate because of data inadequacy. Within thisframework, extended network model and location based encoding approaches are proposed. Firstly, the extended network model is created by inspired from the hypothesis of that "integrating the known protein interactions within host and pathogen organisms improve the success of prediction of unknown pathogen-host interactions". Secondly, location based encoding is feature extraction method which is used for encoding of amino acid sequences. One of the important factors is feature which affects success in prediction of pathogen-host interaction within machine learning algorithms. In biological databases, the most data is the information of amino acid sequence regarding proteins. Prediction accuracy of pathogen-host interaction will be increased by that a robust feature extraction method is developed on the basis amino acidsequence. Furthermore, extraction of feature vectors for all the known interactions are provided in easier way by the sake of using the information of amino acid sequence.

In this thesis, PROSES (Protein SequencebasedEncodingSystem) which is a user-friendly interface and freely accessible web server, has been designed for researchers, who are working on the field of protein encoding and prediction of protein interaction. The web server is especially useful for those who are not familiar with programming languages. PROSES is currently being used at http://proses.yalova.edu.tr which is storedin the web server of Yalova University.

BÖLÜM 1. GİRİŞ

Proteinler 20 farklı aminoasidin farklı sayı ve sırada moleküler seviyede bir araya gelmesiyle oluşmuş makro moleküllerdir. Hücre büyümesi, üreme, besin alımı, hücreler arası iletişim, gen ekspresyonu gibi yaşamsal faaliyetlerin her adımında görev almaktadır. Biyolojik sistemin işleyişinde görev alan proteinlerin bazıları kendi başlarına bir fonksiyon icra ederken, birçoğu diğer proteinler ile doğrudan veya dolaylı olarak etkileşim içerisindedir. Proteinlerin kendi başlarına veya diğer proteinlerle etkileşime girerek gerçekleştirdikleri fonksiyonun bilinmesi biyolojik işleyişin anlaşılması ve işleyiş sırasında yaşanan sıkıntıların tespit edilmesi ve gerekli önlemlerin alınması açısından önemlidir. Hastalıkların teşhis edilmesi ve tedavi süreçlerinin belirlenmesinde sadece proteinler arası etkileşimlerin bilinmesi yeterli olmamakla birlikte önemli bir yere sahiptir. Hastalık teşhis ve tedavisinde dört aşamalı araştırma ve bilgiye ihtiyaç vardır. Bunlar; ilk aşamada, moleküler seviyede etkileşimlerin tespit edilmesi, ikinci aşamada proteinler arası etkileşimler göz önüne alınarak etkileşim ağlarının (yolaklar-pathways) bilinmesi, üçüncü aşamada hücresel işlemlerin bilinmesi ve son aşamada dokusal seviyede etkilerin tespit edilmesidir. Tezde önerilen yöntemler ile ilk aşamaya katkı sunularak proteinler arası moleküler seviyedeki etkileşimler tahmin edilecektir.

Literatürde protein-protein etkileşimleri (PPE) tür içi ve türler arası etkileşimler olarak incelenmektedir [1]. Tür içi etkileşimler, hücre içindeki proteinlere ait fonksiyonların ve biyolojik işleyişin nasıl kontrol edildiğinin anlaşılması açısında önemlidir [2], [3]. Türler arası etkileşimler genellikle patojen-konak etkileşimleri (PKE) olarak adlandırılır. Bu tür etkileşimlerde patojen canlı, virüs, bakteri, mantar, parazit vb. gibi başka canlılarda hastalıklara sebep olan organizmalara denmektedir. Patojen organizmalar konak denilen başka canlılar üzerine herhangi bir yolla yerleşip kendi proteinlerini konak canlının hücre çekirdeğinden içeri bırakmaktadır. Konak

hücrenin çekirdeğine yerleşen patojen proteinler buradaki birtakım proteinlerle moleküler seviyede etkileşip, konak proteinlerin yapısını bozmaktadır. Yapısı bozulan konak hücredeki proteinler yapmaları gereken fonksiyonları yerine getirememekte, dolayısıyla bu durum konak canlıda biyolojik işleyişin aksamasına sebep olmaktadır. Patojenlerin başka bir canlı üzerine yerleşip çoğalmasıyla oluşan hastalıklara bulaşıcı, enfeksiyonel hastalık veya salgın denmektedir. Salgınlar, insanlık tarihi boyunca çok büyük kitlesel ölümlere sebep olmuştur. Özellikle eski çağlarda salgınların sebep olduğu patojenlerin bilinmemesinden dolayı hastalığın toplum içinde yayılmasını engelleyecek tedbirler de alınamamıştır.

Tarihte yaşanan salgınlar içerisinde en çok kayıp verilenlerden biri kara vebadır. 1330'larda yersinia pestis adı verilen bakterinin sebep olduğu veba salgınının Doğu Asya veya Orta Asya'nın bir bölgesinde ortaya çıktığı tahmin edilmektedir. Savaşa giden ordular, sıçanlar ve pireler aracılığıyla dünyanın farklı bölgelerine yayılmıştır. Asya, Avrupa ve Kuzey Afrika'da hızla yayılmıştır. Kara veba, Avrasya nüfusunun dörtte birinden fazlasının canına mal olmuştur. Dünya genelinde 75 ila 200 milyon arasında insanın ölümüne sebep olduğu tahmin edilmektedir [4].

Tarihte büyük kayıplara sebep olan patojenlerden biri de çiçek virüsüdür. 1520 yılında Küba'dan Meksika'ya giden bir İspanyol filosunda yer alan köleler aracılığıyla Meksika'nın Cempoallan kasabasında yayılmaya başlamış, altı ay gibi kısa bir sürede ülkenin tümünü sarmıştır. 1520 yılında 22 milyon olan Meksika nüfusu salgın süresince sekiz milyon insanın hayatına mal olmuştur [5]. Bu salgından yaklaşık iki yüzyıl sonra, 1778 senesinde, çiçek virüsü ile beraber tifo virüsü de İngiliz denizciler aracılığıyla Havvaii adalarına bulaşmıştır. Hastalıktan önce adalarda yarım milyon insan yaşamaktaydı. Beş sene gibi bir süre için salgın 400 binden fazla kişinin ölümüne sebep olmuştur [6].

Salgınlar 20. Yüzyıla gelene kadar dünyanın farklı bölgelerinde benzer kitlesel ölümlere sebep olmaya devam etmiştir. Birinci dünya savaşı sırasında Kuzey Fransa'daki askerler arasında güçlü bir grip türü olan "İspanyol gribi" yayılmaya başlamıştır. 20. Yüzyılda ulaşım ağının da gelişmesiyle birlikte salgınların yayılma

hızıda artmıştır. Öyle ki İspanyol gribi birkaç ay içinde dünya nüfusunun üçte birine bulaşmıştı. Virüs Hindistan nüfusunun %5'inin (15 milyon insan) Tahiti adası nüfusunun yüzde 14'ünün, Samoa adası nüfusunun %20'sinin ölümüne sebep olmuştur. Salgın dünya genelinde yaklaşık 50 ila 100 milyon arasında insanın ölümüne neden olmuştur.

20. Yüzyıla gelindiğinde çiçek, grip, tifo, veba gibi salgın hastalıkların tedavisinde önemli başarılar elde edilmiştir. Bu salgınlara karşı geliştirilen çeşitli aşı, anti bakteriyel ve diğer medikal alt yapı sayesinde tarihte büyük ölümlere sebep olan salgınların bazıları neredeyse yok denecek kadar azalmıştır. Örneğin çiçek hastalığıyla küresel çapta yapılan mücadele sonucu, 1979 yılında dünya sağlık örgütünün yaptığı açıklamada hastalığın neredeyse bittiği ifade edilmiştir. 21. Yüzyılda enfeksiyonel hastalıkların bazılarında önemli mesafeler kat edilmesine, hatta birçok salgının ortadan kaldırılmasına rağmen, halen patojenlerin sebep olduğu Ebola, HIV, Influenza, SARS, E. coli gibi hastalıklar her yıl milyonlarca insanın sağlığını kötü yönde etkilemekte ve ölümlere sebep olmaktadır. Sadece 2013 yılında salgın hastalıklardan dolayı 9,2 milyon kişi hayatını kaybetmiştir. Bu sayı o sene içerisinde gerçekleşen ölümlerin tümünün % 17'sine karşılık gelmektedir [7].

Salgınlar insan sağlığını ve yaşamını tehdit etmenin yanında ekonomik olarak da büyük maliyetlere sebep olmaktadır. Hastalıklara karşı tedavi stratejileri geliştirerek maddi ve manevi kayıpların önüne geçmek için enfeksiyon mekanizmasının anlaşılması önemlidir. Patojen ve konak organizmalara ait proteinler arası fiziksel etkileşimlerin tespiti, patojenlerin konak canlıda sebep olduğu enfeksiyonel hastalık mekanizmasının anlaşılması açısından ilk ve en önemli aşamadır. Hastalıklara sebep olan etkileşimlerin tespiti ile tedavi yöntemlerinin (aşı, antibiyotik vb.) geliştirilmesi sağlanacak ve hastalığın yayılmasını engellemede daha etkili çözümler bulunacaktır.

Hem tür içi hem türler arası protein etkileşim tespitinde kullanılan yöntemler deneysel (in vivo, in vitro) ve hesaplamalı (in silico) olarak iki ana başlık altında toplanmaktadır. Deneysel yöntemler küçük ölçekli ve geniş ölçekli yöntemler olarak ayrılır. Genetik, biyokimyasal ve biyofiziksel özelliklere bakılarak yapılan tespitler

küçük ölçekli yöntemler olarak adlandırılmaktadır. Küçük ölçekli yöntemlerde tek deney ile bir protein çiftine ait etkileşim incelenmektedir [8]. Son yıllarda binlerce protein çiftinin tek seferde tespit edildiği geniş ölçekli yöntemler geliştirilmiştir [9]. Yeast two hybrid systems, mass spectromety, protein chip gibi yöntemler geniş ölçekli deneysel tespit yöntemleridir. Deneysel yöntemler etkileşim tespitinde zaman alan, pahalı yöntemlerdir dolayısıyla bu yolla bulunan etkileşimler olası etkileşim çiftlerinin çok azını kapsamaktadır. Örneğin insanda yaklaşık 100 000 olan protein sayısı, 1000 farklı proteini olan bir organizma ile çaprazlandığında olası tüm etkileşimlerin kontrolü için 10^8 deney gerektirmektedir. Deneysel yöntemlerin uygulama zorluğu, proteinler arası etkileşimlerin modellenmeye çalışılarak etkileşimlerin tahmin edilmeye çalışıldığı hesaplamalı yöntemlerin geliştirilmesi ihtiyacını doğurmuştur. Bu yaklaşım biyomoleküler ve medikal bilimler ile matematiksel hesaplamalar ve mühendislik disiplininin bir araya getirildiği disiplinler arası bir araştırma alanıdır [10], [11]. Deneysel olarak doğrulanmış etkileşim verilerinden yola çıkarak bilinmeyen etkileşimler hesaplamalı yöntemlerle tahmin edilmektedir [12]. Bu yöntemlerde protein çiftlerine ait protein yapı bilgisi, domain, gen komşuluğu, filo genetik profil, gen ekspresyonu ve literatür tarama bilgisi gibi öznitelikler tek başlarına veya kendi aralarında kombine edilerek etkileşim tahmininde kullanılmaktadır [13].

Hesaplamalı yöntemler proteinler arası etkileşim tespitinde, tespit süresinin kısaltılması ve maliyetin düşürülmesi dışında deneysel yöntemlerle yanlış tespit edilen etkileşimlerin kontrolünde de kullanılmaktadır.

Tezde amaç makine öğrenmesi tabanlı hesaplamalı yöntemler kullanarak konak ve patojen organizmalara ait proteinlerin etkileşim tahmininde, literatürde geçen yöntemlere göre daha doğru sonuçlar elde etmektir. Bu kapsamda danışmanlı öğrenmeyi esas alan makine öğrenmesi algoritmaları kullanılarak proteinler arası etkileşim tahmininde doğruluğu arttırmak üzere çalışmalar yapılmıştır. Tahmin doğruluğunun arttırılması amacıyla danışmanlı öğrenmenin farklı adımlarına uygulanabilecek genişletilmiş ağ modeli ve lokasyon tabanlı öznitelik kodlama olarak adlandırılan iki farklı yöntem önerilmiştir. Genişletilmiş ağ modeli, konak

patojen arası etkileşim ağlarına tür içi ağların eklenmesi (genişletilmesi) ile tahmin doğruluğunun artacağı hipotezi üzerine geliştirilmiştir. Bu yöntemde yeterli verinin olmadığı türler arası etkileşim ağları, tür içi ağlar kullanarak genişletilmiştir. Lokasyon tabanlı öznitelik kodlama ile de, öğrenme sürecinde proteinlerin ayırt edilebilirliğini arttırmak, dolayısıyla daha doğru tahminlerde bulunmak amacıyla dizilim tabanlı yeni bir öznitelik vektör çıkarım yöntemi önerilmiştir. Yapılan deneyler sonrası tezde önerilen her iki yöntemin de tahmin doğruluğunu arttırdığı görülmüştür.

Tezgiriş bölümü ile birlikte toplamda altı bölümden oluşmaktadır.

İkinci bölümde protein etkileşim tespitinde kullanılan hesaplamalı ve deneysel yöntemlere ait literatür özeti verilmiştir. Deneysel yöntemlerin neler olduğu kısaca açıklanmış ve literatürde geçen çalışmalara ait referanslar verilmiştir. Hesaplamalı yöntemler farklı başlıklar altında incelenmektedir. Bu başlıklar kısaca açıklanmış ve literatürde yapılan önemli çalışmalara değinilmiştir.

Üçüncü bölümde deneysel çalışmalarda kullanılan veri setlerinin erişildiği veri tabanları, tezde geçen protein kodlama ve etkileşim tahmin yöntemleri açıklanmıştır. Hesaplamalı yöntemlerin geliştirilmesi için kullanılan biyolojik ağlara ait verilere ulaşılacak çevrimiçi veri tabanları tanıtılmıştır. Yöntemler başlığı altında deneylerde kullanılan tahmin metotları ve öznitelik kodlama yöntemleri açıklanmıştır.

Dördüncü bölümde, tezde önerilen genişletilmiş ağ modeli ve lokasyon tabanlı öznitelik kodlama yöntemleri açıklanmıştır. Önerilen yöntemler haricinde tez kapsamında geliştirilen, etkileşim tahmini öncesi gerekli ön işlemlerin yapılmasında ve protein kodlama, protein dizi sorgulama gibi işlemlerin gerekli olduğu diğer çalışmalarda kullanılabilecek PROSES yazılımı tanıtılmıştır.

Dördüncü bölümde önerilen yöntemlerin doğruluğu için yapılan deney sonuçları verilmiştir. Deneylerde kullanılan değerlendirme metrikleri açıklanmıştır. Genişletilmiş ağ modeli ile ilgili deney sonuçları “deneysel çalışma 1”, lokasyon

tabanlı kodlama yöntemi ile ilgili deney sonuçları ise “deneysel çalışma 2” başlığı altında detaylı olarak yorumlanmıştır.

Tezin son bölümünde önerilen yöntemler ve geliştirilen yazılım hakkında değerlendirmeler yapılıp sonuçlar özetlenmiştir.

BÖLÜM 2. LİTERATÜR ÖZETİ

Proteinler arası etkileşimlerin tespiti ve etkileşim sonucu oluşan ağlar ile ilgili birçok çalışma yapılmıştır. Bu çalışmaların büyük çoğunluğu tür içi etkileşim ağları ile ilgili olup konak-patojen etkileşimleri kapsayan türler arası etkileşimler üzerine yapılan çalışmalar daha azdır [14].

Proteinler arası etkileşim tespitinde kullanılan yöntemler, deneysel ve hesaplamalı olarak iki ana başlık altında toplanmaktadır. Deneysel yöntemler, fiziksel etkileşimler ve fonksiyonel yakınlığı tespit eden yöntemler olarak ayrılır. Fiziksel etkileşim tespit yöntemleri de karmaşık (complex) ve ikili (binary) tanımlama olarak ayrılmaktadır. Bu yöntemler ayrıca canlı üzerinde (in vivo) yapılan ve canlı dışında (in vitro) yapılan tespitler olarak da farklılık göstermektedir [15]. Tablo 2.1.'de farklı deneysel etkileşim tespit yöntemi, deney ortamı, etkileşim tipi ve detaylı bilgiye ulaşılacak referanslar verilmiştir.

Deneysel yöntemler, tek bir deneyde tespit edilen etkileşim sayısına göre küçük ölçekli ve geniş ölçekli yöntemler olarak ikiye ayrılmaktadır. Küçük ölçekli etkileşim tespit yöntemlerinde her bir deneyde bir protein çiftine ait etkileşim durumu test edilmektedir [16]. Son yıllarda geliştirilen Y2H, affinity purification, Mass spectrometry, DNA ve protein microarrays gibi geniş ölçekli tespit yöntemleri sayesinde aynı anda binlerce protein çifti arasındaki etkileşim durumunu test etmek mümkün hale gelmiştir.

Tür içi ve türler arası etkileşimlerin deneysel yöntemlerle tespit edilmesi uzun zaman almakta ve yüksek maliyet gerektirmektedir. Ayrıca farklı deneysel yöntemlerle tespit edilen etkileşimlerde yanlış tespitler (false positive, false negative) olabilmektedir [17].

Deneysel yöntemlerin bu tür dezavantajlarından dolayı son yıllarda etkileşim tespitinde hesaplamalı yöntemler önem kazanmıştır. Deneysel yöntemlerle bulunan protein etkileşimleri VirHostNet [18], PHI-base [19], PHIDIAS[20], HPIDB[21], STRING[22] gibi veri tabanlarında paylaşılmaktadır. PPE ve PKE verilerinin paylaşıldığı bu kaynaklar kullanılarak etkileşim tahmininde kullanılmak üzere hesaplamalı modeller geliştirilmektedir. Literatürde hesaplamalı yöntemler genel olarak makine öğrenmesi, homoloji, yapısal, domain-motif tabanlı yöntemler olarak dört ana başlık altında kategorize edilmektedir [14], [23]. Tahmin performansını arttırmak amacıyla bu yöntemler kombine edilerek de kullanılmaktır. PPE ve PKE tahmininde veri eksikliği, özniteliklerin çıkarılamaması ve doğrulanmış negatif veri, hesaplamalı yöntemlerin tümünde karşılaşılan en önemli üç problemdir [1].

Tablo 2.1. Deneysel etkileşim yöntemleri, deney ortamları, etkileşim tipi ve ilgili referans tablosu.

Deneysel Yöntem	Deney Ortamı	Etkileşim Tipi	Referans
Y2H	Canlı	Fiziksel (ikili)	[24], [25]
Affinity purification–MS	Yapay	Fiziksel (karmaşık)	[26]
DNA microarrays/Gene coexpression	Yapay	Fonksiyonel Yakınlık	[27]
Protein microarrays	Yapay	Fiziksel (karmaşık)	[28], [29]
Synthetic lethality	Canlı	Fonksiyonel Yakınlık	[30], [31]
Phage display	Yapay	Fiziksel (karmaşık)	[32]
X-ray crystallography, NMR spectroscopy	Yapay	Fiziksel (karmaşık)	[33]
Fluorescence resonance energy transfer	Canlı	Fiziksel (ikili)	[34]
Surface plasmon resonance	Yapay	Fiziksel (karmaşık)	[35]
Atomic force microscopy	Yapay	Fiziksel (ikili)	[36]
Electron microscopy	Yapay	Fiziksel (karmaşık)	[37]

Proteinlerin yapısal özellikleri, yerel parçaların sıralanışı, üç boyutlu biçimleri ve atomların üç boyutlu uzaydaki konumlarına bakılarak belirlenir. Bu özellikler göz önüne alındığında proteinler, yapısal olarak birincil, ikincil, üçüncül ve dördüncül olarak incelenmektedir. Örneğin proteine ait ikincil yapı, bir biyopolimerin hidrojen bağı yapılarına bakarak tanımlanırken, üçüncül yapılar atomik düzeydeki konumlar ile ilgilidir. Bu tür yapısal özellikler PPE tahmininde kullanılmaktadır. Cai ve arkadaşları [17], proteinlerin ikincil yapılarından yola çıkarak SVM tabanlı bir model ile etkileşim tahmini yapmış ve %88 başarı elde etmiştir. Benzer şekilde Yu ve arkadaşları [38] proteinlerin ikincil yapılarından yola çıkarak helix ve düzensiz yapıların etkileşim bölgelerinin tespitinde kullanılabileceğini göstermişlerdir. Ancak

proteinlere ait yapısal bilgilerin sınırlı olması ve tespit edilen yeni protein çiftlerinin gün geçtikçe artması, modelin uygulanmasını zorlaştırmaktadır.

Etkileşim tahmininde kullanılan nümerik yöntemler ile sıkça kullanılan bilgilerden biri de proteinlerin kökensel olarak yakınlığını gösteren homolog bölgelerdir. Homolog bölgeler proteinler arasında farklılıklar olmasına rağmen aradaki yapısal ve fonksiyonel benzerlik hakkında bilgi vermektedir. Proteinlerin homoloji bilgisi PPE tahmininde de kullanılmıştır. Zhao ve arkadaşları [39], etkileşim tahmininde skor matrisleri ve oto kovaryans değerlerini kullanarak tür içi etkileşim tahminin de %90.71 doğruluğa ulaşmıştır. Benzer şekilde Liu ve arkadaşları [40] amino asitlerin hidropati profilinden yola çıkarak proteinlere ait yeni bir öznitelik vektörü önermiştir. Bu yöntemle yapılan tahminlerde protein dizilimleri arasında düşük benzerliğin olması (dolayısıyla homolog bölgelerin az olması) PPE tahminini zorlaştırmaktadır.

Proteinlere ait domain ve motif bilgisi de PPE tahmininde kullanılan bir diğer yöntemdir. Domain, protein dizilimi içerisinde dizinin geri kalanından bağımsız olarak kendi başına bir fonksiyon gerçekleştirebilen alt parçalara denmektedir. Bu bilgiyi kullanan çalışmalarda proteinlere ait domain bilgisinden yola çıkarak etkileşim tahmini yapılmaktadır. Dyer ve arkadaşları [41], etkileşim tahmininde en az bir domain içeren konak ve patojen proteinlerin etkileşim ve domain bilgisini kullanarak istatistiksel tabanlı bir algoritma önerdiler. Aralarında etkileşim olduğu bilinen proteinlerin domain bilgilerinden bayes tabanlı bir tahmin modeli geliştirdiler. Bu alanda yapılan çalışmalar genellikle tek bir organizma üzerinde uygulanıp başarılı sonuçlar elde edilmiştir. Domain ve motif tabanlı yöntemlerde karşılaşın en önemli problem veri yetersizliğidir.

Literatürde türler arası etkileşim ağlarına ait özellikler de PKE tahmin probleminde kullanılmıştır. Protein ağlarına ait derece (degree), merkezilik (centrality), kümeleme katsayısı (clustring coefficient) gibi özelliklerden yola çıkarak PKE tahmininde bulunulmuştur. Dyer ve arkadaşları [42] patojen ve konak proteinlere ait derece ve merkezilik (betweenness centrality) özelliklerini tahmin için kullanmıştır. Taştan ve arkadaşları [43], Nouretdinov ve arkadaşları [44] çalışmalarında derece, kümeleme

katsayısı ve merkezilik özelliklerini kullanarak tahminde bulunmuşlardır. Bu yöntemin PKE probleminde kullanılabilmesi için organizmalara ait protien ağlarına ihtiyaç vardır.

Makine öğrenmesi tabanlı, danışmanlı ve yarı danışmanlı metotlar tür içi ve türler arası protein etkileşim tahmini problemine başarıyla uygulanmıştır [45], [46]. Bu yöntemler sınıflandırma yapmak için negatif ve pozitif olarak etiketlenmiş veriye ihtiyaç duymaktadır. Pozitif veriler deneysel olarak tespit edilmesine rağmen proteinler arası etkileşimin olmadığına dair kanıtlanmış veriler mevcut değildir. Bu sebeple yapılan çalışmalarda negatif verilerin oluşturulması önemli bir problemdir. Daha önce makine öğrenmesi tabanlı çözüm öneren çalışmalarda bu probleme farklı çözümler önerilmiştir. Bunlardan biri negatif veri seti ihtiyacının ortadan kaldırıldığı veri madenciliği teknikleri ile sadece pozitif veri setleri kullanılarak etkileşim tahminleriyapmaktır [47]–[50]. Ancak negatif veri setinin göz ardı edildiği bu tür yöntemlerde model pozitif sınıf lehine öğrenmede bulunmakta, dolayısıyla yanlış pozitif oranının artmasına sebep olmaktadır [51].

Makine öğrenmesi tabanlı yöntemleri kullanan çoğu çalışmada, negatif veri setleri olası tüm etkileşim uzayı içinden rastgele seçilmektedir [14]. Rastgele seçim yapılan çalışmalarda negatif verinin pozitif verilere oranı farklılık göstermektedir. [52], [53]'te bu oran 1/100 olarak belirlenmiştir. [51]'de pozitif, negatif sınıflar eşit sayıda alınmış ancak negatif sınıfların oluşturulmasında sabselüler ortak lokalize çiftler (subcellular co-localized pairs) ayrı tutulmuştur. Bu şekilde oluşturulan negatif sınıfların rastgele oluşturulana göre daha iyi performans sağladığı görülmüştür. [54]'de yapılan çalışmada negatif, pozitif veriler farklı oranlarda rastgele seçilmesinin model performansını nasıl etkilediği incelenmiştir. Yapılan gözlemlerde, pozitif verinin negatif verilere oranının sonuçları değiştirdiği ancak doğruluk üzerinde çok büyük bir etkisinin olmadığı görülmüştür.

Makine öğrenmesi tabanlı yöntemlerde karşılaşılan diğer bir problem veri yetersizliğidir (data scarcity). Deneysel çalışmaların sınırlı olduğu patojen sistemlerde bu sorunun üstesinden gelmek için birden fazla türe ait etkileşimlerin

birlikte kullanılabileceği matematiksel modeller geliştirilmiş. Bu yöntemlerde farklı veri setleri üzerinde eş zamanlı öğrenme yapılarak tahmin performansını arttırmak hedeflenmiştir. PKE tahmininde performansı arttırmak için farklı organizmalara ait veri setlerinin kombine edildiği çalışma sayısı azdır. [14], [55]'te kısmen etiketlenmiş veri setinden PKE tahmini yapmak üzere yarı danışmalı çoklu-görev (multitask) yöntemi önerilmiştir. Bu yöntemde temel fikir danışmalı bir sınıflandırıcının yanında düzenleme parametresine sahip yarı danışmalı bir yöntemi yardımcı olarak kullanıp çoklu-görev bir öğrenme yapmaktır. Xi ve arkadaşları [56] ortak matris faktorizasyonu (collactive matrix factorization) yaklaşımını kullanarak birden fazla etkileşim ağına ait ilişki matrisini eş zamanlı faktorize etmişlerdir. Bu işlem sırasında faktörlerin ortak kullandığı parametreler sayesinde veri setleri arası bilgi paylaşımı yapılmaktadır. Kshirsagar ve arkadaşlarının [52] yaptığı çalışmada matris tamamlama temelli bir yöntem ile farklı PKE verileri üzerinden eş zamanlı öğrenme yapılmıştır. Bu çalışmada türler arası etkileşimlere has benzerlik matrisleri kullanılarak farklı etkileşimler bir arada değerlendirilmiştir.

Literatürde hesaplamalı yöntemlerle PPE ve PKE tahmini üzerine yapılan çalışmalara bakıldığında tüm yöntemlerin veri yetersizliği (buna bağlı olarak özniteliklerin çıkarılamaması) ve doğrulanmış negatif veri setinin olmaması gibi ortak problemlerinin olduğu görülmektedir. PKE tahmininde doğruluğu arttırmak amacıyla önerilecek yeni modelin bu tür problemleri göz önünde bulundurması gerekmektedir. Bu çalışmada önerilen genişletilmiş ağ modelinde yetersiz olduğu düşünülen türler arası etkileşim ağlarına ağ içerisinde yer alan proteinlere ait tür içi etkileşimler de dâhil edilerek veri yetersizliği problemi aşılmaya çalışılmıştır. Tez de önerilen bir diğer yöntem olan lokasyon tabanlı öznitelik çıkarımı ile proteinlere ait veri sıkıntısının yaşanmadığı birincil yapılar kullanılarak PKE tahmini yapılmıştır.

BÖLÜM 3. MATERYAL VE YÖNTEM

Bu bölümde önerilen yöntemlerin başarı değerlendirmesinde kullanılan veri setlerinin temini, kullanılan makine öğrenmesi tabanlı tahmin yöntemleri ve protein kodlama yöntemleri açıklanmıştır. Alt başlıklarda ilk olarak tezde önerilen yöntemlerin test edilmesi için gerekli veri setlerinin temin edileceği çevrimiçi veri tabanları ve bu veri tabanlarında hangi verilerin test için uygun olduğu açıklanmıştır. Daha sonra etkileşim tahmininde kullanılan makine öğrenmesi tabanlı etkileşim tahmin yöntemleri anlatılmıştır. Son olarak tahmin yöntemine verilecek öznitelik vektörlerinin oluşturulması için kullanılacak kodlama yöntemleri anlatılmıştır.

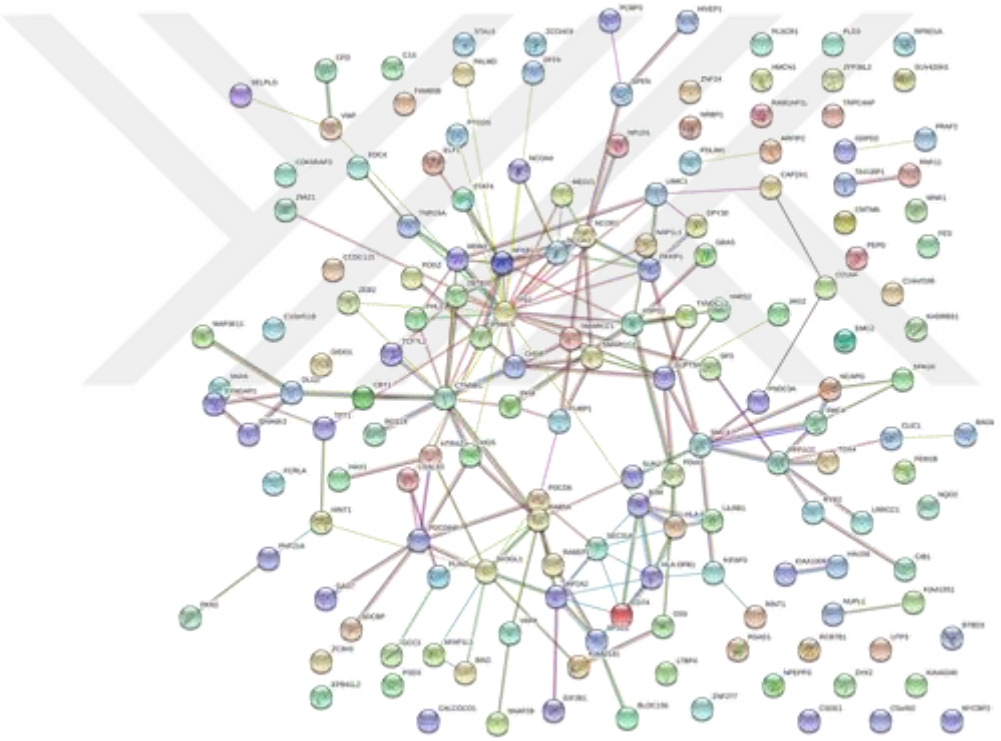
3.1. Biyolojik Ağlar ve Çevrimiçi Veri Tabanları

Teknolojideki ilerleme ile beraber farklı organizmalara ait PPE verilerinin kullanıma açıldığı çok sayıda veri tabanı bulunmaktadır. Son yıllarda proteinlerin ve PPE etkileşimlerinin yer aldığı 100 den fazla çevrimiçi veri tabanı bilimsel yayınlarda tanıtılmış ve araştırmacıların kullanımına sunulmuştur [57]. Protein etkileşimlerinin yer aldığı bu veri tabanlarından bir kısmı sadece deneysel olarak doğrulanmış etkileşimleri, bir kısmı hesaplamalı yöntemlerle bulunan etkileşim tahminlerini, bir kısmı da her iki yöntemle bulunan verileri paylaşmaktadır.

Veri tabanlarında paylaşılan etkileşim verileri bütün olarak düşünüldüğünde bu veriler ile bir etkileşim ağı oluşturulabilmektedir. Etkileşim ağlarının analiz edilmesi ile organizmaların kendi içinde ve farklı türler arasında oluşan karmaşık yapı hakkında önemli bilgiler elde edilmektedir.

PPE verilerinin analiz edilmesi ve etkileşim ağlarının görselleştirilmesi için de son yıllarda çok sayıda yazılım aracı geliştirilmiştir. Bu araçlar sayesinde etkileşim

verilerine ait ağlar çıkarılmakta ve ağlara ait derece (degree), merkezilik (centrality), kümeleme katsayısı (clustring coefficient) gibi özellikler verilmektedir. Yapılan sorgulamalarda proteinler arası etkileşimlerin tespit edildiği yöntemler belirlenerek de sorgulamalar yapılabilmektedir. Şekil 3.1.'de insana ait örnek 200 protein arasındaki etkileşimler, STRING veri tabanından yapılan sorgulama ile görselleştirilmiştir. Sisteme proteinlerin UniProt kimlikleri verilmiş ve bu etkileşimlere ait yönsüz bir graf oluşturulmuştur. Şekilde nodlar arası kenarlar, tespit edilen yönteme göre farklı renklerle gösterilmiştir. Grafın tümü için ve yöntemlere göre kümelenmiş haliyle çıkarılan istatistiksel sonuçlar da çıktı olarak üretilmektedir.



Şekil 3.1. String veri tabanından insana ait proteinler için oluşturulmuş örnek etkileşim ağı.

Tablo 3.1. Literatürde en sık geçen PPE ve PKE veri tabanları

Veritabanı	URL	Referans
VirHostNet	http://virhostnet.prabi.fr/	[18]
PHI-base	http://www.phi-base.org/	[19]
PHIDIAS	http://www.phidias.us/	[20]
HPIDB	http://hpidb.igbb.msstate.edu/index.html	[21]
BIND	http://binddb.org/	[58]
DIP	http://dip.doe-mbi.ucla.edu	[59]
MINT	http://mint.bio.uniroma2.it/mint	[60]
BioGrid	http://www.thebiogrid.org	[61]
STRING	http://string-db.org/	[22]
HPRD	http://www.hprd.org/	[62]
IntAct	http://www.ebi.ac.uk/intact/	[63]
PDBsum	www.ebi.ac.uk/pdbsum	[64]
ProPrint	http://crdd.osdd.net:8081/ProPrint/	[65]
MIPS	http://mips.gsf.de	[66]
PDZBase	http://abc.med.cornell.edu/pdzbase	[67]
iRefIndex	http://irefindex.org	[68]
KEGG	http://www.genome.ad.jp/kegg/	[69]
PHISTO	http://www.phisto.org	[70]

Tablo 3.1.'de PPE ve PKE etkileşiminde en çok kullanılan veri tabanları ve ilgili referanslar verilmiştir. Veri tabanlarında, paylaşılan etkileşimlerin hangi organizmaya ait olduğu, etkileşim tipinin nasıl belirlendiği ve güncel etkileşim sayıları web sayfalarında paylaşılmaktadır. Araştırmacılar bu verileri XML, SIF, txt, xls, xlsx, v.b. dosya formatlarında kullanmaktadır. Veriler veri tabanlarından bu formatların birinde veya birkaçından indirilebilmektedir. Veri tabanlarında aramalar, protein adı veya protein kimliği gibi tekil anahtar kelimelere göre yapılmaktadır. Bu çalışmada kullanılan veriler STRING (Search Tool for the Retrieval of Interacting Genes/Proteins), PHISTO (Pathogen host interaction search tool) ve UniProt veri tabanlarından indirildi. STRING veri tabanı çok sayıda organizmaya ait aralarında fiziksel ve anlamlı fonksiyonel ilişki olan proteinler arası etkileşimleri tutar. Çalışmada genişletilmiş ağ modelinde önerilen tür içi etkileşimler bu veri tabanından elde edildi. Türler arası etkileşimler ise PHISTO veri tabanından alındı. PHISTO aslında etkileşim verilerinin bir laboratuvardan veya araştırma merkezinden alınıp paylaşıldığı bir veri tabanı değildir. PHISTO, birçok organizmaya ait proteinlerin insan proteini ile yaptığı patojen konak etkileşimlerinin paylaşıldığı dokuz farklı veri tabanından (IRefIndex, MPIDB, APID, Reactome, STRING, BIND, MINT, IntAct, DIP) sorgulanıp kullanıcı dostu bir ara yüz ile bir araya getirildiği bir araçtır. STRING ve PHISTO yazılımlarında veriler sorgulandıktan sonra etkileşimlere ait görsel ve istatistiksel bilgilere ulaşmak da mümkündür. Şekil 3.1.'de STRING veri

tabanından alınan etkileşim verilerine ait ağ görseli verilmiştir. Çalışmada makine öğrenmesi tabanlı algoritmalar ile yapılan tahminlerde proteinlerin amino asitlere ait dizi bilgileri kullanıldı.

Tür içi ve türler arası etkileşim ağları indirildikten sonra ağ içinde geçen her bir proteine ait amino asit dizi bilgisi UniProt veri tabanından indirildi. Protein etkileşim tahmininde kullanılan en önemli veri tabanları Tablo 3.1.'de verilmiştir.

3.2. Deneylerde Kullanılan Tahmin Yöntemleri

Bu bölümde, önerilen yöntemlerin başarı değerlendirmesinde kullanılan tahmin yöntemleri açıklanmıştır. Yöntemlerin başarı değerlendirmesinde matris faktorizasyonu, karar ağaçları, istatistiksel ve örnek tabanlı sınıflandırıcılar gibi literatürde farklı problemlerin çözümünde sıkça kullanılan makine öğrenmesi tabanlı tahmin metotları kullanılmıştır. Kullanılan sınıflandırıcıların bir kısmı daha önceki çalışmalarda PKE tahmininde kullanılmışken bir kısmı ilk defa bu çalışmada test edilmiştir. Kullanılan tahmin metodunun PKE problemine uygun olup olmadığı deneysel çalışmalardan elde edilen sonuçlara göre değerlendirilmiştir.

3.2.1. Matris faktorizasyonu

Matris faktorizasyonu (Matrix factorization), danışmalı öğrenme başlığı altında yer alan bir makine öğrenmesi yöntemidir. Daha çok puanlama tahmininde matris tamamlama amacıyla kullanılmaktadır. Kullanıcıların bazı ürünlere yaptıkları tercih puanından yola çıkarak diğer ürünlere verebileceği puanlar tahmin edilmektedir. Bu metot proteinler arası etkileşim tahmininde ilişki matrisi üzerinden, puanlama tahminine benzer şekilde, bilinmeyen etkileşimlerin tahmin edilmesi mantığı ile çalışır. [52]'de proteinlere ait öznitelik vektörlerini haritalayan faktorize edilmiş matrisler yardımıyla konak patojen etkileşimi tahmin edilmiştir. Bu yayında [71]'den alınan faktorizasyon modeli genişletilerek birden fazla türe ait verilerin eş zamanlı değerlendirilmesine olanak sağlanmıştır.

PKE etkileşim verileri, türlerin her biri bir tarafta olmak üzere ikili bir graf şeklinde gösterilir. G_t, υ ve ς tipinde nodları birbirine bağlayan ikili bir graf olsun m_t ve n_t sırasıyla υ ve ς nod türlerine ait nod sayıları olsun. $M \in \mathbb{R}^{mt \times nt}$ matrisi G_t grafı içerisindeki etkileşimleri gösteren bir ilişki matrisi olsun. Graf içerisindeki tüm kenarlar Ω kümesinde tanımlı olsun. υ tipinde nodların öznitelik uzayı X ve ς tipindeki nodların öznitelik uzayı Y olsun. Öznitelik vektör uzunlukları eşit ve d_t olduğunu varsayalım. υ tipindeki nodların her birine ait öznitelik vektörü $x_i \in X$ ve ς tipindeki nodlara ait öznitelik vektörü $y_i \in Y$ olmaktadır. Matris tamamlama probleminde amaç, M matrisinde nodlar arasındaki ilişkiyi tanımlayan bir $f:X \times Y \rightarrow \mathbb{R}$ fonksiyonunu öğrenmektir. f fonksiyonunun $X \times Y$ uzayı üzerinde bilineer olduğu kabul edilir ve aşağıdaki formda yazılır.

$$f(x_i, y_j) = x_i^T H y_i = x_i^T U V^T y_j \tag{3.1}$$

Denklem 3.1'de $H \in \mathbb{R}^{dt \times dt}$ matrisi X ve Y öznitelik uzayını haritalamaktadır. Bu modelde H matrisinin $U \in \mathbb{R}^{dt \times k}$ ve $V \in \mathbb{R}^{dt \times k}$ boyutlarında olan matrislerin çarpımı şeklinde yazılabileceği kabul edilir. $H = UV^T$ denkleminde bulunan U ve V matrisleri iki öznitelik uzayını haritalamakta kullanılmaktadır. Burada amaç eğitim veri setini kullanılarak optimum U ve V matrislerini bulmaktır. Denklem 3.2'de matrislerin faktörize edilmesinde kullanılan amaç fonksiyon verilmiştir. Bu denklem, Ω kümesindeki her bir eleman göz önüne alınarak yapılan tahminin ne kadar iyi olduğunu gösteren bir uyum (data fitting) terimi ve döngüsel adımları H matrisi için kontrol eden bir düzenleme (regularization) teriminden oluşmaktadır.

$$L(U,V) = \sum_{(i,j)\in\Omega} c_{i,j}\, \ell\left(M_{ij}, x_i^T U V^T y_j\right) + \lambda(\|U\|_F^2 + \|V\|_F^2)$$
$$\ell(a,b) = (a-b)^2 \tag{3.2}$$

Veri uydurma terimi, kareli hata, lojistik-kayıp gibi herhangi bir kayıp fonksiyonu olabilir. Kayıp fonksiyonu çözülecek problemin hassasiyetine ve tahmin edilecek değişkenin doğasına göre seçilir [71].

Denklem 3.2'de PKE problemi için daha hızlı yakınsadığı ve adım boyutunun daha hassas olduğu düşünülerek karesel hata fonksiyonu kullanılmıştır. Denklemde geçen λ, kayıp fonksiyonu ve düzenleme terimi arasında bir karar parametresi olarak kullanılmaktadır. $c_{i,j}$, Ω kümesi içindeki (i,j) çifti arasındaki hata oranını belirlemeye imkân sağlayan bir ağırlık katsayısıdır. Formülde öğrenme işlemi, $H = UV^T$ denkleminde optimum U ve V faktörlerinin bulunması, yani $||U||_F^2 + ||V||_F^2$ teriminin minimize edilmesi anlamına gelir.

3.2.2. Naif bayes

Naif bayes (Naive Bayes), bayes tabanlı kuralları esas alan istatistiksel bir sınıflandırma algoritmasıdır. Bayes teorisine göre X hipotezinin doğru verilmesi durumunda Y olayının gerçekleşme ihtimali denklem 3.3'te verilmiştir. Aşağıdaki denklemde verilen X hipotezi sınıflandırma için verilen öznitelik vektörüdür ve birden fazla öznitelikten oluşur. Y olayı ise X özniteliğinin yer alabileceği olası tüm sınıfların birini temsil eder. Eğitim tamamlanıp modelin oluşturulması sonrasında X öznitelik vektörünün verilmesi halinde hangi olayın gerçekleşeceği (örneğin hangi sınıfa atanacağı) tahmin edilir.

$$P(Y \mid X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{3.3}$$

Naif bayes (NB) öznitelik vektöründeki tüm değişkenlerin koşullu bağımsız olduğunu varsayar. Y'nin verilmesi durumunda n adet özelliğin X için tahmin değeri aşağıdaki gibi olmaktadır:

$$P(X_1, \ldots, X_n \mid Y) = \prod_{i=1}^{n} P(X_i|Y) \tag{3.4}$$

Genel olarak Y'nin herhangi bir ayrık değerli değişken olduğunu varsayarsak, (X_1, … ,X_n) öznitelikleri ayrık ya da gerçek değerli özniteliklerdir. Gerçek değerli

öznitelikler için şartlı olasılık, ortalamanın μ ve standart sapmanın σ olduğu, denklem 3.5'teki olasılık dağılım fonksiyonuna göre hesaplanır.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.5}$$

Sınıflandırılması gereken her X örneği için olası Y değerleri üzerinde olasılık dağılımı çıkaracak bir sınıflandırıcı tanımlanır. Bayes kuralına göre k. olası değer için Y'nin olasılığını veren ifade şöyledir:

$$P\left(Y = y_k \,|\, X_1, \ldots, X_n\right) = \frac{P(Y = y_k) \prod_i P\left(X_i \,|\, Y = y_k\right)}{\sum_j P\left(Y = y_j\right) \prod_i P\left(X_i \,|\, Y = y_k\right)} \tag{3.6}$$

Denklem 3.6, NB sınıflandırıcısının kullandığı temel denklemdir. Yeni bir $X_{yeni}(X_1, \ldots, X_n)$ örneği verildiğinde olası Y değerini bulmak için NB şu kuralı kullanmaktadır:

$$y \leftarrow \underset{y_k}{argmax} \frac{P(Y = y_k) \prod_i P\left(X_i \,|\, Y = y_k\right)}{\sum_j P\left(Y = y_j\right) \prod_i P\left(X_i \,|\, Y = y_k\right)} \tag{3.7}$$

3.2.3. Bayes ağları

Bayes ağları(Bayesian Networks), rastgele değişkenler arasındaki istatistiksel ilişkilerin tanımlandığı graf tabanlı bir modeldir. Bayes ağları (BN), V = (A_1, A_2, A_3, …, A_N) şeklinde değişken dizisi ve yönlü kenarlardan (E) oluşmaktadır. Bu ağlar rastgele değişkenler üzerinde birleşik olasılık dağılımını tanımlamayı sağlayan, G = (V, E) olarak gösterilebilen döngüsel olmayan yönlü graflardır [72]. Verilen bir G grafındaki her bir değişken kendi atasından koşullu bağımsız olmaktadır [73]. Bir bayes ağında, P(V) değişkenlerine ait birleşik dağılım, ağ içinde belirtilen tüm koşulların çarpımına eşittir. P(V) aşağıdaki şekilde tanımlanmaktadır:

$$P(A_1, A_2, A_3, \ldots, A_N) = \prod_{i=1}^{N} P(A_i | Pa_i) \tag{3.8}$$

Yukarıdaki denklemde $P(A_1, A_2, A_3, \ldots, A_N)$, V değişkenlerinin herhangi bir kombinasyonuna ait olasılıktır. $P(A_i|Pa_i)$ ise Pa_i'nin verilmesi durumunda A_i olayının gerçekleşme ihtimalidir. Bu denklemdeki her bir değişkene ait koşullu dağılımın, maksimum benzerlik (maximum likelihood) tahmini ile öğrenilebilen parametrik bir formu vardır.

3.2.4. C4.5

C4.5, karar ağacı tabanlı bir sınıflandırma çeşididir. Karar ağaçları düğüm kenar ve yapraklardan oluşmaktadır. Karar ağacının düğümlerinde eğitim verisinin öznitelikleri, yapraklarında ise karar sonucu ulaşılan sınıflar yer almaktadır. Ağaç üzerindeki dallanmalar kenarlar üzerinde yer alan spesifik bir değere veya aralığa göre gerçekleşir. Karar ağaçlarının oluşturulmasında genel prensip verilen eğitim verisinden yola çıkarak belirlenen bir ölçüm fonksiyonu ile tüm öznitelikler arasından veriyi en iyi ayıran özniteliği seçmektir. En iyi ayıran öznitelik belirlendikten sonra veri seti tekrar sınıflara göre gruplanır. Tüm gruplar aynı sınıfa dâhil olana kadar işlemler rekürsif olarak tekrar eder. Veri setindeki tüm örneklerin aynı sınıfa dâhil olması ile işlem sonlandırılır. Karar ağacı ailesine dâhil olan algoritmalar arasındaki fark kullandıkları ölçüm fonksiyonuna göre farklılık göstermektedir. C4.5, Quinlan [74] tarafından karar ağaçlarına dayalı geliştirilen ve daha önce gene kendisinin geliştirmiş olduğu ID3 [75] algoritmasının güncellenmiş halidir. C4.5 öznitelikleri ayrıştırılmasında kullandığı bilgi kazanım oranı (information gain ratio) ile ID3 algoritmasından ayrılmaktadır. C4.5 algoritmasının matematiksel ifadesi aşağıda açıklanmıştır.

C eğitim verisindeki sınıf sayısı, S j. sınıf ve T eğitim verisi olsun. S sınıfına ait entropi değeri denklem 3.9'a göre hesaplanır.

$$Entropy(S) = -\sum_{j=1}^{C} p(S,j) \times \log p(S,j) \tag{3.9}$$

Entropi değeri bulunduktan sonra T eğitim verisine ait bilgi kazancı denklem 3.10'a göre bulunur.

$$Gain(S,T) = Entropy(S) - \sum_{v \in Values(T_S)} \frac{|T_{S,v}|}{|T_S|} Entropy(S_v) \quad (3.10)$$

Denklem 3.10'da geçen T_s eğitim verisinde S özniteliğine ait alt veri setidir. $T_{s,v}$eğitim verisinde S özniteliğinin v değerine sahip olduğu örneklere ait alt veri setidir. Bu değerler bulunduktan sonra S özniteliğine ait kazanç oranı (GainRatio) denklem 3.11'deki gibi hesaplanır.

$$GainRatio(S,T) = \frac{Gain(S,T)}{SplitInfo(S,T)} \quad (3.11)$$

Denklem 3.11'de S özniteliğinin T eğitim verisinden ayrılma bilgisini ifade eden SplitInfo(S,T) değeri denklem 3.12'deki gibi hesaplanır.

$$\mathrm{SplitInfo(S,T)} = -\sum_{\mathrm{v \in Values(T_S)}} \frac{|\mathrm{T_{S,v}}|}{|\mathrm{T_S}|} \times \log \frac{|\mathrm{T_{S,v}}|}{|\mathrm{T_S}|} \quad (3.12)$$

3.2.5. Rastsal orman

Karar ağaçları özetle yüksek entropiye sahip özniteliklerin ebeveyn olarak belirlenmesi mantığıyla oluşturulur. Bu durum sınıflandırıcının eğitim verisine fazla odaklanmasına (overfit) ve test sırasında önemli bazı öznitelikleri kaçırmasına sebep olabilmektedir. Dolayısıyla test aşamasında kaçırılan özniteliklerden dolayı düşük doğruluk sonuçları elde edilebilir. Rastsal orman (Random Forest) sınıflandırma yönteminde eğitim verisi, rastgele K alt kümeye ayrılır. Bu alt kümelerin her biri için farklı karar ağaçları oluşturularak çeşitliliğin artması sağlanır. Bu yaklaşım ile yüksek oy modeli kullanılarak test örneği en fazla oyun alındığı sınıfa atanır. Bu sayede birden fazla karar ağacı kombine edilerek tek bir model haline getirilir.

Rastsal orman (RF) algoritmasının geliştirmeye çalıştığı çözümlerden biri aralarında korelasyonun olmadığı veya düşük olduğu karar ağaçları üretmektir. Bu problem alt veri setleri ve bu veri setlerine ait ağacın oluşturulmasında farklı bir yol izlenerek çözülmüştür. Aşağıda RF metodunun çalışmasına ait algoritma adımlar verilmiştir.

1. Veri setinde tekrarlı olabilecek şekilde N boyutunda örnek al.
2. Tekrarlanmayacak şekilde rastgele öz nitelikler seç
3. Seçilen öz nitelikler ile gini indeksini kullanarak verilere göre karar ağacı oluştur.
4. Adım 2'yi tüm öznitelikler kullanılana kadar tekrar et.
5. 1-4 arası adımları istenilen sayıda tekrar et.

Yukarıdaki algoritmada birinci adımda oluşturulan alt veri setinde tekrar eden örnekler olabilmektedir. Bu şekilde farklı karar ağaçları oluşturmak mümkün olmaktadır. İkinci adımda öznitelikler rastgele seçilir ve bunlar seçim listesinde yeniden yer almaz. Bu adımda genellikle tek adımda iki veya üç öznitelik birlikte seçilir. Bu seçimlere göre karar ağacı şekillenir. 1-4 arası adımlar uygulandığında her döngü sonrası yeni bir karar ağacı oluşur. Beşinci adımda algoritmanın sonlandırılması kullanıcının vereceği K parametresine göre değişmektedir. Literatürde bu sayının kaç olması gerektiği ile ilgili çalışmalar yapılmıştır. Ancak olması gereken karar ağacı sayısı problemde kullanılan veri setine göre değişmektedir.

3.2.6. k-en yakın komşuluk

k-en yakın komşuluk (k-Nearest Neighborhood), örnek tabanlı bir sınıflandırıcı türüdür. Danışmanlı öğrenme algoritmaları sınıfına dâhildir. Sınıfı bilinmeyen her hangi bir örneği en yakın k komşusuna bakarak sınıflandıran bir algoritmadır. Bu algoritmada sınıflandırma doğruluğu seçilen k değerine duyarlıdır. Bu sebeple test sırasında seçilen k komşu sayısı sınıflandırıcı performansını doğrudan etkilemektedir. Yöntemin kolay uygulanabilir olması ve lineer olarak sınıflandırılması, farklı dağılıma sahip veri setlerinde başarılı sonuçlar elde etmesi en önemli avantajlarıdır.

Yöntemin eğitim fazının olmaması dolayısıyla tüm işlemlerin test aşamasında yapılması bir hesaplama maliyeti getirmesi en önemli dezavantajıdır. Ayrıca sınıflandırma sonrası doğruluğun yüksek olması için de geniş bir veri setine ihtiyaç duymaktadır.

Çalışmada kullandığımız özniteliklerin tümü sayısal değerlerden oluşmaktadır. Bu sebeple iki öznitelik arasındaki benzerliğin bir metrik ile tanımlanması gerekmektedir. Uzaklık metrikleri probleme göre çebişev, Öklid, kare (manhattan), hamming gibi metriklerden biri olabilir. Örneğin eğitim verisinin öznitelik vektörü X, test verisinin öznitelik vektörü X' ve vektör uzunluğu n olsun. Eğitim ve test vektörleri arası d Öklid uzaklığı denklem 3.13'teki gibi hesaplanır.

$$d = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2} \tag{3.13}$$

Sınıfın belirlenmesi için verilen bir X' örneği ile veri setindeki her bir örnek arasındaki d uzaklığı hesaplanır. Eğitim verisindeki her örnek için bulunan uzaklıklar sıralanır. Bu adımdan sonra kullanıcıların belirlediği 'k' komşuluk değerine göre en yakın örnekler alınır. Verilen X' örneğinin atanacağı sınıf, komşular arasında en fazla benzerliğin olduğu sınıfa atanır. Problemde her bir sınıf A kümesinin bir elemanı olsun. Buna göre örneğin her bir sınıfa ait olma şartlı olasılığı denklem 3.14'e göre bulunur.

$$P(y = j|X' = x) = \frac{1}{k}\sum_{i \in A} I(y^i == j) \tag{3.14}$$

Yukarıdaki denklemde geçen I fonksiyonu, aldığı parametrelerin eşit olması durumunda 1, diğer durumlarda 0 döndürmektedir. Sonuç olarak X' örneği olasılığı en yüksek olan sınıfa atanmaktadır.

3.2.7. K*

K*, örnek tabanlı bir sınıflandırıcıdır. Örnek tabanlı metotlar her hangi bir veriyi, önceden sınıflandırılmış eğitim veri setindeki örneklerle karşılaştırarak sınıflandırmaktadır [76]. Bu sınıflandırıcı, örnekler arası benzerliğin belirlenmesinde bazı uzaklık ve benzerlik fonksiyonlarından yararlanır. K* algoritması entropi temelli bir uzaklık fonksiyonu kullanarak diğer örnek tabanlı sınıflandırıcılardan ayrılır [77].

3.3. Öznitelik Çıkarım Yöntemleri

Bu bölümde literatürde sıkça kullanılan protein kodlama yöntemleri anlatılmıştır. Anlatılan yöntemlerin bir kısmı yapılan deneylerde önerilen yöntemlerin performansını kıyaslamak için kullanılmıştır. Deneylerde kullanılmayan kodlama yöntemleri PROSES yazılımı içinde yer almakta olup biyoinformatik alanında çalışan araştırmacıların kullanımına sunulmuştur. Bu yöntemlerden amino asit kompozisyon (AAC), Amino asit çifti (AAP), Bitişik Üçlü (CT), Komposizyon-Geçiş-Dağılım ve dipeptit kompozisyon (DC) yöntemleri protein dizi uzunluğundan bağımsız olarak eşit uzunlukta öznitelik vektörleri üretmektedir. Kompozisyon moment vektör (CMV) yönteminde, M kaçıncı dereceden moment alınacağını belirten parametre olmak üzere, öznitelik vektör boyutu 20×M olmaktadır. Residue-couple model (RCM) yönteminde, R öznitelik vektörününrank parametresi olmak üzere, öznitelik vektör boyutu R×400 olmaktadır. Vektör boyutu verilen bir parametre ile değişen yöntemlerde, sabit uzunlukta öznitelikler üretmek mümkündür. Geri kalan Ortonormal kodlama (OE), OETMAP, Taylor Venn Diyagram (TVD) yöntemlerinde ise öznitelik vektör boyutu protein dizi uzunluğu ile orantılı olmaktadır. Her bir yöntemin oluşturduğu öznitelik vektör boyutu ve ilgili referans, Tablo 3.2.'de verilmiştir. Her bir metoda ait detaylı bilgi ilgili başlık altında anlatılmıştır.

Tablo 3.2 Yöntemlere ait öznitelik vektör uzunluğu ve ilgili referans.

Metot	Uzunluk	Referans	Metot	Uzunluk	Referans
AAC	20	[78]	CMV	20×M	[79]
AAP	400	[80]	DC	400	[78]
CT	343	[3]	OE	N×20	[81]
CTDC	21	[82]	OETMAP	N×30	[83]
CTDD	105	[82]	RCM	R×400	[84]
CTDT	21	[82]	TVD	N×10	[85]

3.3.1. Amino asit kompozisyon

Amino asit kompozisyon (Amino Acid Composition) kodlama yöntemiyle 20 farklı amino asidin protein dizilimindeki frekanslarına bakılarak öznitelik vektörü çıkarılır. Protein dizilimi içerisinde her bir amino asidin tekrar sayısı hesaplanıp toplam dizi uzunluğuna bölünerek öznitelik vektörü içerisindeki nümerik değer hesaplanır. Amino asit dizi uzunluğundan bağımsız olarak 1×20 boyutunda bir öznitelik vektörü oluşur. Dizi uzunluğu N ve öznitelik değeri hesaplanmak istenen i. aminoasidin tekrar sayısına n_i dersek bu proteine ait öznitelik vektörü denklem 3.15'e göre oluşturulur.

$$F_{AAC} = \left[\frac{n_1}{N}, \frac{n_2}{N}, \ldots, \frac{n_{20}}{N}\right] \tag{3.15}$$

Denklem 3.15'te bulunan öznitelik vektörü sadece proteinde yer alan amino asitlerin kompozisyon bilgisini içermektedir.

3.3.2. Amino asit çifti

Amino asit çifti (Amino Acid Pair) Chen ve ark. [80] tarafından geliştirilmiştir. Bu yöntemde bir dipeptidin protein diziliminde geçme sıklığı diğer dipeptitlerinki ile kıyaslanır. Örneğin, MTAEEMK dizilimine sahip bir protein MT, TA, AE, EE, EM, MK şeklinde dipeptitlere ayrılır. Olası tüm dipeptitler göz önüne alındığında aminoasit dizi uzunluğundan bağımsız olarak 20 farklı aminoasit için 20×20 boyutunda öznitelik vektörü oluşmaktadır. Bir dipeptitin dizi içerisinde geçme sıklığına f^{+}_{AAP}, diğer tüm dipeptitlerin sayısına f^{-}_{AAP} dersek, bu dipeptit için AAP değeri denklem 3.16'ya göre hesaplanır.

$$R = log\left(\frac{f^{+}_{AAP}}{f^{-}_{AAP}}\right) \tag{3.16}$$

Her bir dipeptit için R değeri hesaplandıktan sonra bulunan değerler denklem 3.17'deki gibi [-1,+1] aralığına normalize edilir. Normalizasyonun amacı öğrenme sürecinde herhangi bir özniteliğin diğer özniteliği baskılamasına engel olmaktır.

Denklem 3.17'de geçen min ve max değişkenleri sırasıyla dizilimdeki en düşük ve en yüksek R değerine sahip dipeptitlere karşılık gelmektedir.

$$R_{AAP} = 2\left(\frac{R_{AAP} - min}{max - min}\right) - 1 \tag{3.17}$$

3.3.3. Kompozisyon moment vektörü

Proteinlerin birincil yapısına (amino asit dizilimi) bakarak amino asitlerin frekans ve konum bilgileri çıkarılabilir. AAC ve AAP kodlama yöntemleri frekans bilgisini kullanırken konum bilgisini göz ardı etmektedir. Kompozisyon moment vektörü (Composition Moment Vector) yönteminde ise proteinlerin birincil yapısından hem konum hem frekans bilgisi kullanılarak öznitelik vektörü çıkarılmaktadır.

P herhangi bir protein, A_i bu proteinin aminoasit dizisi içerisinde geçen i. amino asidi olsun. Tüm amino asitlerin A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y şeklinde sıralandığını ve her birinin sırasıyla (x_1,x_2,…,x_{20}) değişkenlerine karşılık geldiğini düşünelim. N boyutundaki bir aminoasit dizisine sahip bir proteinin, i = 1,2,…20 ve k≥0 için k. dereceden moment vektörü denklem 3.18'e göre hesaplanır. Denklemde geçen $n_{i,j}$ i. amino asidin j. lokasyonuna, K_i ise i. aminoasidin dizi içerisindeki toplam sayısına karşılık gelir.

$$x_i^{(k)} = \frac{1}{N(N-1)\dots(N-k)} \sum_{j=1}^{K_i} n_{i,j}^k \tag{3.18}$$

k = 0 için oluşan Kompozisyon moment vektörü (CMV), sadece frekans bilgisi olup AAC kodlama ile aynı sonucu vermektedir. K'nın maksimum değeri için P proteinine ait moment matrisi denklem 3.19'a göre hesaplanır.

$$A_{K+1} = \begin{bmatrix} x_1^{(0)} & x_2^{(0)} & \dots & x_{20}^{(0)} \\ x_1^{(1)} & x_2^{(1)} & \dots & x_{20}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(K)} & x_2^{(K)} & \dots & x_{20}^{(K)} \end{bmatrix} \tag{3.19}$$

Herhangi iki protein için oluşturulan moment matrisinin aynı olması bu iki proteinin aynı olduğunu göstermektedir. CMV kodlamaya göre oluşturulan öznitelik vektör boyutu K parametresine göre değişkenlik gösterip $K \times 20$ boyutunda olmaktadır. N dizi uzunluğuna sahip bir protein için verilecek maksimum K değeri N-1 olmak zorundadır.

3.3.4. Bitişik üçlü

Bitişik Üçlü (Conjoint Triad) kodlama yönteminde proteinlere ait öznitelik vektörleri amino asitlerin frekanslarına ve kimyasal özelliklerine bakılarak çıkarılır. Amino asitlerin elektrostatik ve hidrofobik özellikleri göz önüne alınarak belirlenen yedi farklı sınıftan birine atanmaktadır. Daha sonra proteinin birincil yapısındaki her bir amino asit kendisinden sonra gelen iki amino asit ile birlikte üçlü birimlere ayrılır. Yapılan bu sınıflama ve gruplamaya göre, proteinlerin aminoasit dizi uzunluğundan bağımsız olarak toplamda 343 (7^3) farklı üçlü grup oluşabilir. Herhangi bir P proteinine ait öznitelik vektörü tüm üçlü gruplara ait frekansların hesaplanıp normalize edilmesi ile bulunur. Aminoasit dizisi içinde geçen i. üçlü gruba ait tekrar sayısına f_i dersek, bu proteine ait R öznitelik vektöründeki r_i değeri denklem 3.20'ye göre hesaplanır.

$$r_i = (f_i - \min\{f_1, f_2, \dots, f_{343}\})/(\max\{f_1, f_2, \dots, f_{343}\}) \quad (3.20)$$

3.3.5. Kompozisyon, geçiş, dağılım

Kompozisyon, Geçiş ve dağılım (Composition, Transition, Distribution) öznitelikleri proteinlerin amino asitlere ait kimyasal ve yapısal özelliklerini göstermektedirler. Bu özelliklerin her biri ayrı ayrı veya birleştirilerek tek bir öznitelik vektörü olarak kullanılabilir [86]. Öznitelik vektörü iki adımda oluşturulmaktadır.

İlk adımda protein dizisi amino asitlerin yedi farklı kimyasal özelliğine (hydrophobicity, normalized Van der Waals volume, polarity, polarizibility, charge, secondary structures, solvent accessibility) göre sınıflandırılır. Yedi farklı özelliğin her biri Tomii ve Kanehisa [82], [87] tarafından belirlenen üç farklı gruptan birine atanır. Örneğin polarity özelliği; pozitif, negatif ve neutral olarak, secundary structure özelliği; helix, strand ve coil olarak gruplanır. Bu yönteme göre yapılan gruplama ve sınıflandırma detayları Tablo 3.3.'te verilmiştir.

Kompozisyon yönteminde (CTDC), kimyasal özelliklerine göre etiketlenen amino asitlerin frekansları hesaplanıp normalize edilir. Bu yöntemde öznitelik vektör uzunluğu 21 (7×3) olmaktadır.

Geçiş yönteminde (CTDT), öznitelik vektörü, etiketlenmiş her bir aminoasidin kendisinden sonra gelen amino asidin sınıf ve grubuna geçişlerinin sayılıp normalize edilmesi ile bulunur. Olası tüm geçiş sayısı 21 olup bu aynı zamanda öznitelik vektör boyutudur.

Dağılım yönteminde (CTDD), öznitelik vektörü, belirli bir gruba ait aminoasidin ilk, %25, %50, %75 ve %100'ünün geçtiği lokasyonlar belirlenerek bu gruba ait dağılımların normalize edilmesiyle bulunur. Her bir grup için beş farklı lokasyon olasılığı olduğundan 21 grup için 105 (21×5) farklı öznitelik değeri çıkarılır.

3.3.6. Dipeptit kompozisyon

Bu yöntemde amino asit dizisi ikili gruplar (dipeptide) halinde ayrılır. Örneğin, MTAEEMK dizilimine sahip bir protein MT, TA, AE, EE, EM, MK olacak şekilde dipeptitlerine ayrılır. 20 farklı aminoasit için oluşturulabilecek dipeptide sayısı 400 (20×20) olmaktadır. Herhangi bir amino asit dizisinde dipeptitler bulunup frekansları hesaplandıktan sonra olası tüm dipeptitlere bölünerek öznitelik hesaplanır. Buyöntemde öznitelik vektör boyutu 400 olmaktadır. F_i i. dipeptide ait öznitelik ve

bu dipeptidin verilen dizi içerisinde geçen tekrar sayı d_i olsun. Öznitelik değerleridenklem 3.21'e göre hesaplanır.

$$F_i = \frac{d_i}{400} \tag{3.21}$$

Tablo 3.3. Amino asitlerin kimyasal özelliklerine göre gruplanması

Sınıf	Grup-1	Grup-2	Grup-3
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R,K,E,D,Q,N	G, A, S,T,P,H,Y	C,L,V,I,M,F,W
Normalized Van der Waalsvolume	Volume range 0–2.78	Volume range 2.95–94.0	Volume range 4.03–8.08
	G,A,S,T,P,D	N,V,E,Q,I,L	M,H,K,F,R,Y,W
Polarity	Polarityvalue 4.9–6.2	Polarityvalue 8.0–9.2	Polarityvalue 10.4–13.0
	L,I,F,W,C,M,V,Y	P,A,T,G,S	H,Q,R,K,N,E,D
Polarizibility	Polarizabilityvalue 0–1.08	Polarizabilityvalue 0.128–120.186	Polarizabilityvalue 0.219–0.409
	G,A,S,D,T	C,P,N,V,E,Q,I,L	K,M,H,F,R,Y,W
Charge	Positive	Neutral	Negative
	K,R	A,N,C,Q,G,H,I,L,M,F,P,S,T,W,Y,V	D,E
Secondarystructures	Helix	Strand	Coil
	EALMQKRH	VIYCWFT	G,N,P,S,D
Solventaccessibility	Buried	Exposed	Intermediate
	A,L,F,C,G,I,V,W	P,K,Q,E,N,D	M,P,S,T,H,Y

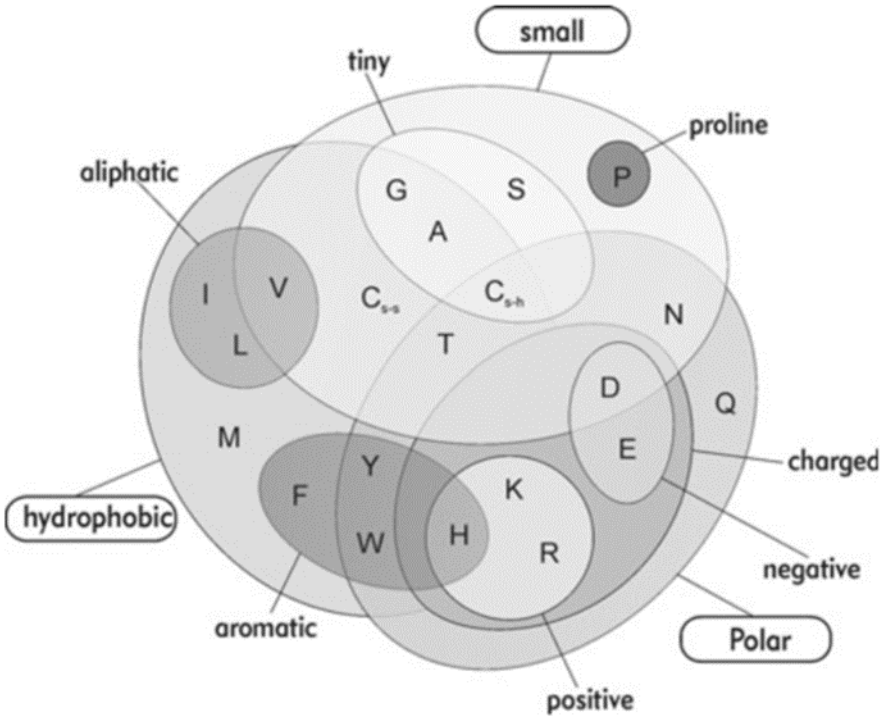
3.3.7. Ortonormal kodlama

Ortonormal kodlaman(Orthonormal Encoding) yönteminde 20 farklı aminoasidin her biri 1×20 boyutunda bir vektör ile ifade edilir. Bu vektör P= {A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V} kümesindeki amino asit dizilimine göre oluşturulur. Özniteliği oluşturulacak amino asidin P kümesindeki lokasyon değeri '1' geri kalan değerler '0' olacak şekilde atama yapılır. Örneğin P kümesinde lokasyonu bir olan 'A' amino asidine ait vektör F_A = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] olmaktadır. F_i, P kümesindeki i. aminoaside ait vektör olmak

üzere, MTAEEMK dizilimine sahip bir proteine ait öznitelik vektörü F=[F_M, F_T, F_A, F_E, F_E, F_M, F_K] olur. N uzunluğunda bir diziye sahip proteine ait F öznitelik vektörü N×20 boyutunda olmaktadır.

3.3.8. Taylor venn diyagramı

Taylor venn diyagramı (TVD), amino asitlerin fizikokimyasal özellikleri arasındaki ilişkilerden yola çıkarak proteinleri kodlayan bir yöntemdir. Bu yöntemde ilk olarak amino asitlerin sahip oldukları fizikokimyasal özellikler belirlenir. Şekil 3.2.'de 10 farklı fizikokimyasal özellik diyagramlar halinde gösterilip amino asitler bu diyagramlara yerleştirilmiştir. Diyagramların birbiriyle kesişme ve kapsama durumları olduğundan dolayı bir amino asit birden fazla fizikokimyasal özelliğe sahip olabilmektedir. Amino asitlerin sahip oldukları özellikleri nümerik formatta göstermek amacıyla 10×1 boyutunda vektör tanımlanır. Bu vektörün her bir elemanı bir fizikokimyasal özelliğe karşılık gelir. Kodlanmak istenen aminoasidin sahip olduğu özelliklerin vektörde karşılık geldiği yerler 1, geri kalan değerler 0 olarak atanır. Bu mantıktan hareketle Şekil 3.2.'te verilen diyagrama göre her bir aminoasit için 10×1 boyutunda vektör elde edilir. Her bir amino asidin kimyasal özelliklerini gösteren nümerik değerler Tablo 3.4.'te verilmiştir.



Şekil 3.2. Taylor Venn diyagram [88]

Proteinler, TVD yöntemine göre kodlanırken dizilimde yer alan her bir amino asit Tablo 3.4.'teki nümerik vektöre bakılarak kodlanır. Örneğin 'W' amino asidi F_W = [1 0 0 1 0 0 0 0 0 0] vektörü ile nümerik olarak tanımlanır. N uzunluğunda bir diziye sahip proteine ait F öznitelik vektörü $N \times 10$ boyutunda olmaktadır.

3.3.9. OETMAP

OETMAP, ortonormal kodlama ve TVD yöntemlerinin birlikte kullanıldığı hibrit bir yöntemdir. L uzunluğunda bir P proteinine ait i. amino asidin 20 bitlik OE vektörü $\{\vec{a}_i\}$ ve 10 bitlik TVD vektörü $\{\vec{b}_i\}$ olsun. OETMAP yönteminde bu aminoaside ait vektör $\{\vec{x}_i\}$ olsun. Bu vektör aşağıda görüldüğü gibi OE ve TVD vektörlerinin birleştirilmesi ile elde edilir.

$$\vec{x}_i=(\{\vec{a}_i||\{\vec{b}_i\}) \tag{3.22}$$

L uzunluğundaki P proteininin OETMAP ile kodlanmasıyla elde edilen $\vec{X}$ öznitelik vektörü Lx30 boyutunda olmaktadır.

Örnek verecek olursak ‘ALDFEQEM’ dizilimine sahip bir proteinde ‘D’ amino asidi için OETMAP vektörü aşağıdaki gibi bulunur.

$$\{\vec{a}_3 = [0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]\}$$
$$\{\vec{b}_3 = [0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0]\}$$
$$\{\vec{x}_3 = [0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0]\}$$

Yukarıda yapılan işlem tüm amino asitler için yapılıp bulunan vektörler hesaplanarak birleştirilir ve öznitelik vektörü çıkarılır.

Tablo 3.4. TVD kodlama yönteminde her bir amino aside ait nümerik vektör.

Özellik	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Hydrophobic	1	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0	1	1	1	1
Positive	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Negative	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Polar	0	1	1	1	0	1	1	0	1	0	0	1	0	0	0	1	1	1	1	0
Charged	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
Small	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1
Tiny	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Aliphatic	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
Aromatic	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
Proline	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

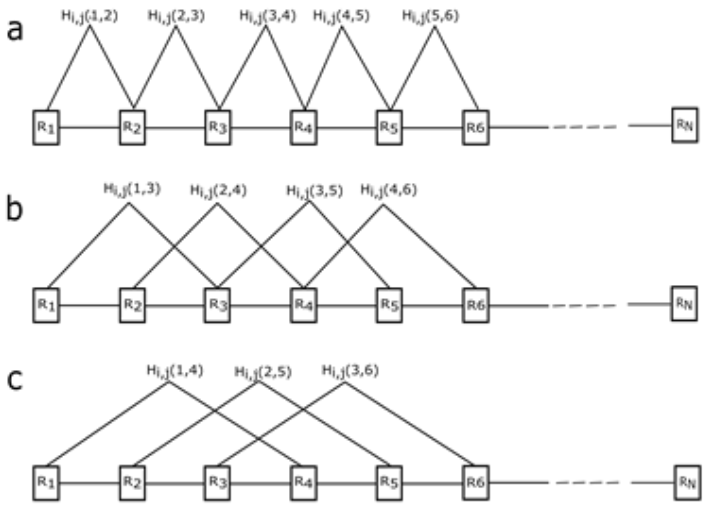
3.3.10. Amino asit aşleme modeli

Amino asit eşleme modelinde (Residue-Couple Model) amino asit dizilimine bakarak farklı desenlere sahip çiftler (dipeptide) oluşturulur. Dizilimden çıkarılan çiftler olası tüm çiftlerin yer aldığı vektör içerisinde 1, geri kalanlar ise 0 olacak şekilde kodlanır. Aminoasit dizi uzunluğundan bağımsız olarak bir proteinde 400 (20×20) aminoasit çifti oluşturulabilir. Bu yöntemde farklı desende dipeptitler oluşturmak için çiftlerin seçildiği lokasyonlar yöntem uygulanmadan önce seçilen k parametresine göre değişmektedir. Proteine ait dizilimde ilk amino asidin seçildiği konuma n dersek, ikinci aminoasit n+k konumundan seçilmektedir. Amino asit çiftlerinin dizilim içinden seçilip kodlanmasına ait formül denklem 3.23’te verilmiştir.

$$X_{i,j}^{(k)} = \frac{1}{(N-k)} \sum_{n=1}^{N-k} H_{i,j}(n, n+k)$$
$$k < N, i = 1,2, \ldots ,20 \; vej = 1,2, \ldots 20$$
$$X_{i,j}^{(1)} = \frac{1}{(N-1)} \sum_{n=1}^{n-1} H_{i,j}(n, n+1) \qquad (3.23)$$
$$X_{i,j}^{(2)} = \frac{1}{(N-2)} \sum_{n=1}^{n-2} H_{i,j}(n, n+2)$$

Denklem 3.23'te n. konumdaki amino asidin i. , n+k konumundaki aminoasidin j. olması durumunda $H_{i,j}(n,n+k) = 1$ diğer durumda $H_{i,j}(n,n+k) = 0$ olur. Şekil 3.3.'de k = 3 değeri için sırasıyla dizilimin birinci, ikinci ve üçüncü dereceden ranklarının nasıl oluşturulduğu görsel olarak açıklanmıştır.

RCM metodunda her bir rank için 400×1 boyutunda öznitelik vektörü oluşturulur. Kodlanmak istenen proteinden kaçıncı dereceye kadar rank alınacağına k parametresi ile karar verilir. Oluşan öznitelik vektörü 400×k boyutunda olur.



Şekil 3.3. k = 3 değeri için örnek bir dizilimin birinci, ikinci ve üçüncü dereceden ranklara göre amino asit çiftlerinin seçilmesi.

BÖLÜM 4. ÖNERİLEN YÖNTEMLER VE PROSES YAZILIMI

Bu bölümde, patojen ve konak organizmalar arasında gerçekleşen protein etkileşimlerinin tahmin doğruluğunu arttırmak amacıyla önerilen yöntemler açıklanmıştır. Yöntemlere ek olarak, PPE probleminin, amino asit dizilim tabanlı öznitelik vektörleri kullanılarak, makine öğrenmesi algoritmalar ile çözümünde gerekli ön işlemlerin yapılacağı web tabanlı PROSES yazılımı geliştirilmiştir. PROSES yazılımı ile yapılacak işlemler ve yazılıma ait modüller bu bölümde açıklanmıştır.

4.1. Genişletilmiş Ağ Modeli

Genişletilmiş ağ modeli, PKE tahmininde karşılaşılan veri yetersizliği problemi dikkate alınarak geliştirilmiştir. Tür içi ve türler arası protein etkileşimlerinin deneysel yöntemlerle belirlenmesi yüksek maliyet ve uzun zaman gerektirmektedir. Deneysel yöntemlerle belirlenen ve veri tabanlarında paylaşılan etkileşim verilerinin az olması hesaplamalı modellerin geliştirilmesi açısından dezavantajdır. İki türe ait tüm proteinlere ait olası tüm etkileşim sayısı, türlere ait protein sayılarının çarpımına eşittir. Veri tabanlarında türler arası bilinen etkileşimler, olabilecek tüm etkileşimlerin çok azını oluşturmaktadır. Bilinen etkileşimlerin kodlanması ile eğitilen bir model, olası tüm etkileşimlerin çok az bir bölümünü kullanmaktadır. Türler arası etkileşim havuzu modelin öğrendiğinden çok daha fazla çeşitlilik barındırmaktadır. Bu sebeple danışmalı öğrenmeye dayalı algoritmalar türler arası etkileşim tahmininde istenilen başarıya ulaşmamaktadır.

Genişletilmiş ağ modeli, türler arası etkileşim ağına, ağda yer alan proteinlerin tür içinde yaptığı etkileşimlerin dâhil edilmesi ile eğitilen danışmalı modelin tahmin doğruluğunu arttıracağı hipotezi temel alınarak geliştirilmiştir. Bu doğrultuda konak-

patojen etkileşime ek olarak her iki türe ait proteinlerin kendi içlerindeki etkileşim ağı da öğrenme sürecine katılmıştır. Böylece öğrenme modeli çeşitliliğin daha fazla olduğu bir veri seti kullanarak tahmin yapacaktır.

Proteinler arası etkileşimler ilişki matrisleri kullanılarak sayısallaştırılır. İlişki matrisinin boyutunu organizmaların sahip olduğu protein sayısı belirler. Türlere ait proteinler ayrı ayrı satır ve sütunlara yerleştirilir. Aralarında etkileşim olan proteinler, satır ve sütunların kesiştiği yerde 1 olarak kodlanır. Bilinen tüm etkileşimler 1 olarak kodlandıktan sonra geri kalan kısımlar 0 olur. İlişki matrisinde 0 olan yerler bilinmeyen etkileşimleri gösterip, 1'lere, yani bilinen etkileşimlere oranla hayli fazladır. Bu durum veri seyrekliği (data scarcity) olarak ifade edilir.

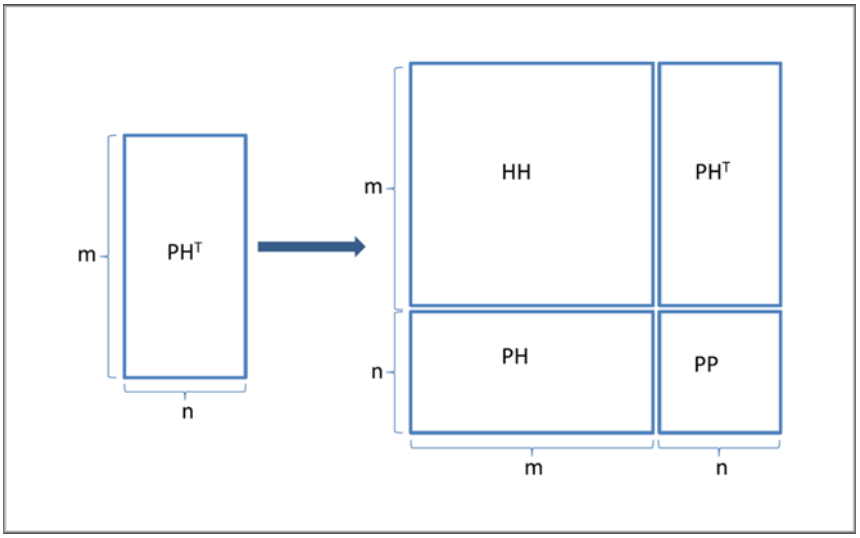
Genişletilmiş ağ modelini matematiksel olarak ifade edelim. Konak organizmaya ait proteinlerin sayısı m ve bu proteinlerin kümesi $X = \{x_1, x_2, \dots, x_m\}$ olsun. Patojen organizmaya ait proteinlerin sayısı n ve bu proteinlerin kümesi de $Y = \{y_1, y_2, \dots, y_n\}$ olsun. Konak-patojen proteinler arası etkileşimleri gösteren ilişki matrisi $M \in \mathbb{R}^{m \times n}$ olsun. M ilişki matrisine göre öğrenme sürecinde kullanılacak $\{(x_i,y_i)\}$ pozitif etkileşimlerin veri seti olsun. M ilişki matrisine, patojen ve konak organizmalara ait proteinlerin eklenmesi ile oluşan yeni ilişki matrisi (k = m+n olmak üzere), $M_{yeni} \in \mathbb{R}^{k \times k}$ boyutlarında genişletilmektedir. Veri setindeki etkileşimlerin yer aldığı Ω_{yeni} veri seti şöyle olmaktadır:

$$\begin{aligned} \Omega_1 &= \{(x_i, y_j)\}, patojen - konak(PH) etkileşimler \\ \Omega_2 &= \{(x_i, x_j)\}, i \neq j, konak - konak(HH) etkileşimler \\ \Omega_3 &= \{(y_i, y_j)\}, i \neq j, patojen - patojen(PP) etkileşimler \\ \Omega_{new} &= \Omega_1 \cup \Omega_2 \cup \Omega_3 \end{aligned} \tag{4.1}$$

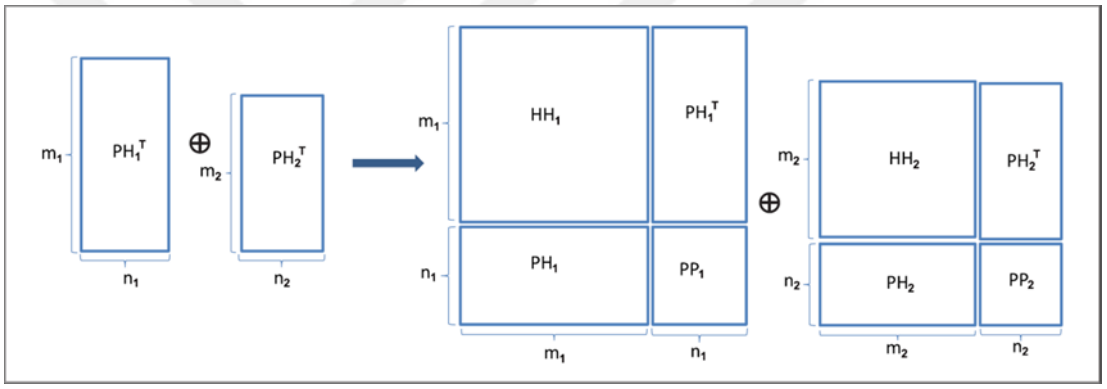
Konak-patojen ilişki matrisinin (PH) tür içi etkileşimleri de içine alacak şekilde nasıl genişletildiği Şekil 4.1.'de görülmektedir. Şekilde konak organizmaya ait tür içi etkileşimler 'HH', patojen organizmaya ait tür içi etkileşimler 'PP' kare matrislerinde yer almaktadır. Bu matrislerin türler arası etkileşim matrisi 'PH' ile birleştirilmesi sonucu M_{yeni} ilişki matrisi oluşmaktadır. Yeni oluşan ilişki matrisi simetrik bir kare

matristir. Bu matris protein sayısı fazla olan organizmalar için devasa boyutlara ulaşabilir. Örneğin, insana ait 70.000, bacillus anthracis bakterisine ait 5000 civarında protein vardır. Bu iki tür için GAM yöntemiyle oluşturulacak yeni ilişki matrisi 75000×75000 boyutunda olacaktır. Bu matrisin PH alanında (70000×5000) bilinen etkileşim sayısı, diğer bir deyişle 1'lerin veya veri setinde pozitif etiketli örneklerin sayısı 5000 civarındadır. İnsan proteinlerinin kendi içindeki bilinen etkileşim sayısı (HH alanı) diğer alanlarda yer alan bilinen etkileşimlerden çok daha fazladır. Eğitim verisinde, yeni ilişki matrisindeki tüm pozitiflerin yer alması tahmin modelinin sayıca fazla olan tür içi etkileşim lehine öğrenme yapmasına sebep olacak ve patojen-konak etkileşim tahmininde yetersiz kalacaktır. Bu nedenle, GAM yönteminde danışmalı öğrenme yöntemleri ile kullanmak üzere hazırlanacak veri setinde oluşan yeni ilişki matrisinde yer alan bilinen etkileşimlerin tümü kullanılmaz. Veri seti oluşturulurken ilk olarak patojen-konak etkileşimler baz alınır. Her bir organizmanın ağ içinde yer alan proteinleri kullanılarak tür içi etkileşimler için yapılan sorgulamada veriler güvenirliğine göre sıralanır. Bu işlem STRING veri tabanında etkileşim güvenirliğini gösteren bir indekse göre yapılmaktadır. Veri seti oluşturulurken tür içi pozitif örnek sayısı, türler arası bilinen pozitif örnek sayısına yakın olacak şekilde seçilir. Böylece $|\Omega_1| \approx |\Omega_2| + |\Omega_3|$ olur.

GAM yönteminde önemli noktalardan biri de negatif veri setinin nasıl oluşturulduğudur. Biyolojik veri tabanlarında doğrulanmış negatif etkileşim verisi olmadığından dolayı negatif etiketli veriler bilinmeyen etkileşimler arasından rastgele seçilmektedir. İlişki matrisi üzerinden söyleyecek olursak negatif örnekler 0 olan alanlardan rastgele seçilmektedir. GAM yöntemine göre negatif veri seti oluşturulurken sadece PH bölgesinde yer alan yani türler arası bilinmeyen etkileşim verileri arasından rastgele seçim yapılır. Literatür çalışmalarında genellikle negatif örnek sayısı pozitif örnek sayısından fazla olacak şekilde veri seti oluşturulur. Pozitif etiketli örneklerin negatif etiketli örneklere oranı 1/1, 1/5, 1/10 arasında değişmektedir.



Şekil 4.1. Tek tür için genişletilmiş ilişki matrisi



Şekil 4.2. Çoklu veri setleri için genişletilen ilişki matrisleri

4.2. Lokasyon Tabanlı Kodlama

Hesaplamalı yöntemlerde veri seyrekliği veya yetersizliği (data scarcity) dışında, veriden kaynaklı bir diğer sorun yöntemin kullandığı protein bilgisine erişememe durumudur. Hesaplamalı yöntemler proteinlere ait farklı bilgileri kullanarak tahminde bulunmaktadır. Proteinlere ait bu bilgiler hesaplamalı yöntemler için etkileşim verisi dışında ikinci bir kısıt olarak karşımıza çıkmaktadır. Örneğin proteinlerin ikincil yapısını kullanan hesaplamalı yöntemler için etkileşim ağında yer alan tüm proteinlere ait ikincil yapıya erişmek zordur. Bu nedenle etkileşimin bilinmesi yeterli olmayacak, aynı zamanda etkileşimde yer alan proteinlerin ikincil yapısına ihtiyaç duyulacaktır. İkincil yapılar bilinmediği takdirde ağdaki etkileşimler

kullanılmayacak ve veri yetersizliği problemi daha da derinleşecektir. Biyolojik veri tabanlarında proteinlere ait en geniş bilginin amino asit dizilimleri olduğu görülmektedir. Dolayısıyla amino asit verisinden yola çıkarak geliştirilecek doğruluğu yüksek bir tahmin metodunun büyük önemi vardır.

Lokasyon tabanlı kodlama (LTK), amino asit dizilimlerinin kullanıldığı bir öznitelik çıkarım yöntemidir. Danışmalı makine öğrenmesi algoritmaları ile kullanılacak veri setindeki tüm örneklerin eşit uzunlukta öznitelik vektörlerine sahip olması gerekmektedir. PKE tahmini için kullanılacak kodlama yönteminin farklı uzunluktaki amino asit dizisini, sabit uzunlukta öznitelik vektörüne dönüştürmesi gerekmektedir. Bu sebeple LTK öncelikle sabit uzunlukta öznitelik vektörü oluşturacak şekilde geliştirildi.

Literatürde sabit uzunlukta öznitelik vektörü oluşturan diğer yöntemler genellikle amino asitlerin dizilim içerisindeki frekans veya kimyasal yapı bilgisini kullanmaktadır. LTK yöntemi ise amino asitlerin dizi içerisindeki konumlarının bir proteinin karakteristiğini ortaya koyabileceği hipotezi üzerine geliştirilmiştir. Yöntemin uygulaması üç aşamaya ayrılmıştır. Birinci adımda amino asit dizisi önceden belirlenmesi gereken bir parametreye göre alt dizilere ayrılır. İkinci adımda her bir dizi nümerik olarak kodlanır. Bu aşamada amino asitlerin indeks faktörü göz önüne alınarak aynı işlem alt dizilerin tersine de uygulanır. Son adımda patojen ve konak proteinlerine ait öznitelik vektörleri birleştirilerek PKE tahmininde kullanılacak nihai öznitelik vektörü oluşturulur. Her bir adıma ait detaylı açıklama aşağıda verilmiştir.

- Adım 1. Aminoasit dizisinin alt dizilere ayrılması

Bu adımda ilk olarak proteine ait dizilimin kaç alt diziye ayrılacağına karar verilir. Yöntem, amino asit dizisine uygulanmadan önce alt dizi sayısı parametre olarak verilmelidir. Burada alt dizi sayısı L değişkeni ile ifade ediliyor olsun. Alt dizisi bulunmak istenen amino asit dizi uzunluğu N, ve her bir alt dizi indeksi i olsun. Amino asit dizisinin alt dizilere ayrılacağı lokasyonları belirten d_i değerleri denklem

4.2'ye göre hesaplanır. İşlem sonrası dizinin kesilme yerlerini gösteren virgüllü sayılar bir alt tamsayıya yuvarlanır.

$$d_i = \left\lfloor i * \frac{N}{L} \right\rfloor \tag{4.2}$$

Tablo 4.1.'de protein dizi uzunluğu 49 olan bir amino asit dizisinin L = 5 değeri için alt dizilere ayırma örneği görülmektedir. Alt dizilerin sonuncusu proteinin amino asit dizilimiyle aynıdır.

Tablo 4.1. L = 5 değeri için alt diziye ayırma örneği

	Dizi	**Uzunluk**
S_n	VQDLMETDLYKLLKSQQLSNDHICYFLYQILRGLKYIHSANVLHRDLKP	49
S_1	VQDLMETDL	9
S_2	VQDLMETDLYKLLKSQQLS	19
S_3	VQDLMETDLYKLLKSQQLSNDHICYFLY	29
S_4	VQDLMETDLYKLLKSQQLSNDHICYFLYQILRGLKYIHS	39
S_5	VQDLMETDLYKLLKSQQLSNDHICYFLYQILRGLKYIHSANVLHRDLKP	49

– Adım 2. Amino asit dizisine ait öznitelik vektörünün çıkarılması

Bu adımda protein diziliminde yer alan 20 farklı amino asidin her biri için, dizilimde yer alan amino asitlerin lokasyonuna göre nümerik bir değer hesaplanmaktadır. Yapılan işlemler sonrası bulunan öznitelik vektörü F, dizi içerisindeki her hangi bir n_i amino asidinin dizi içerisindeki indis değerleri kümesi c_j ve amino asit dizi uzunluğu N olsun. Bu değişkenlere göre n_i amino asidinin öznitelik vektöründeki nümerik değeri denklem 4.3'e göre hesaplanır.

$$F(n_i) = \sum_{j \in C_1} \frac{j}{N} \tag{4.3}$$

Yukarıda verilen denkleme göre her amino asit için nümerik bir değer hesaplanır. Bu yönteme göre protein diziliminden bağımsız olarak 1×20 boyutunda bir öznitelik vektörü oluşur. Şekil 4.3.'te uzunluğu 26 olan örnek bir protein için öz nitelik vektörünün nasıl oluşturulduğu gösterilmiştir. Denklem 4.3'ün daha iyi anlaşılması

için Şekil 4.3.'te verilen dizilime göre 'L' amino asidine ait özniteliği hesaplayalım. Dizilimde 'L' amino asidine ait lokasyonlar {2,6,7,15,20}, tekrar sayısı j=4 ve dizi uzunluğu 26 olmaktadır. Denklem 4.4'te L amino asidine karşılık gelen öznitelik hesaplanmıştır.

$$F('L') = \sum_{j\in\{2,6,7,15,20\}} \frac{j}{N} = \frac{2}{26} + \frac{6}{26} + \frac{7}{26} + \frac{15}{26} + \frac{20}{26} = 1.92 \quad (4.4)$$

Şekil 4.3. Verilen bir protein dizisi için öznitelik çıkarım örneği

– Adım 3. KPE tahmininde kullanılacak öz nitelik vektörünün oluşturulması

Bu aşamada amino asit dizisine sırasıyla adım 1 ve adım 2'de anlatılan alt dizilere ayırma ve öznitelik çıkarma işlemleri uygulanır. Dizinin sonlarında yer alan amino asitler, indislerinden dolayı daha büyük özniteliklere sahip olacaktır. Bu amino asitlerin, frekansı yüksek amino asitlerle karışmaması için her bir alt dizinin devrik haline de adım 2 uygulanır. Proteinin tümünü temsil edecek öz nitelik vektörü her bir alt diziden elde edilen özniteliğin birleştirilmesi ile elde edilir. Bağımsız bir dizilime ait öznitelik vektör boyutu 1×20 olmaktadır. Buna göre LTK algoritması, L=5 değeri için 1×200 (5 alt dizi ve devrik halleri için toplamda 10 farklı dizilim elde edilir) boyutunda öznitelik vektörü oluşturur. Anlatılan işlemler konak ve patojen proteinler için yapılıp sonuçlar bir araya getirildiğinde, PKE tahmininde kullanılacak nihai öznitelik vektörü 1×400 boyutunda olmaktadır.

Şekil 4.4. PKE tahmininde kullanılacak nihai öz nitelik vektörü çıkarma örneği

Şekil 4.4.'te LTK yönteminin örnek protein üzerine uygulanış aşamaları görülmektedir. LTK yöntemi amino asit dizi uzunluğundan bağımsız olarak sabit boyutta öznitelik vektörü oluşturur. Vektör uzunluğu alt dizi sayısını gösteren L parametresine göre değişir. PKE tahmininde optimum L değerinin kaç olacağı veri setine göre değişkenlik gösterebilir. Bu nedenle en iyi performansın elde edildiği alt dizi uzunluğu deneysel olarak belirlenebilir.

4.3. PROSES

Son yıllarda proteinlerin aminoasit dizi bilgileri çok hızlı bir şekilde artmış ve bu veriler çevrimiçi veri tabanları aracılığıyla ücretsiz erişime açılmıştır. Dizilim verilerinin hızla artmasıyla birlikte proteinler ile ilgili problemlerin çözümünde amino asit dizilim verisinin kullanıldığı makine öğrenmesi algoritmalarının kullanımı da yaygınlaşmıştır. Problemlerin bu algoritmalarla çözümünde, yeterli miktarda veri, verilerin ayırt edilmesinde kullanılacak güçlü kodlama yöntemi ve probleme uygun sınıflandırma algoritması ile başarılı sonuçlar alınmaktadır. Biyoinformatik alanında çalışan ve programlama bilgisi olmayan veya kodlama ile zaman harcamak istemeyen araştırmacılar için her bir aşamada kullanılacak açık kaynak kodlu veya ücretli yazılım araçları mevcuttur. Biyolojik verilerin indirildiği veri tabanlarında genellikle verinin analizi ve görselleştirmesi için gerekli modüller bulunmaktadır. Aynı şekilde veri setinden özniteliklerin çıkarılmasından sonra sınıflandırma yapmak için kullanılan Weka, Rapidminer benzeri birçok yazılım mevcuttur. Ancak çalışılan probleme bağlı olarak kullanılan veri setlerinin çok çeşitlilik göstermesinden dolayı

her veriye uygun kodlama aracı bulunmamaktadır. Tezde yapılan deneysel çalışmaların tümünde, kullanılan veri setinde yer alan proteinlere ait amino asit dizilimleri farklı yöntemler ile kodlandı. Amino asit dizilim kodlamasında PROFEAT [89], [90], iFrag [91], PseAAC[92], offline matlab protein encoding toolbox [93] gibi internet üzerinden veya çevrimdışı kullanılabilecek programlar mevcuttur Ancak bu programlarda literatürde geçen kodlama yöntemlerinin tamamını bulmak mümkün değildir. Ayrıca bu programların bazılarına ücretsiz erişim izni yoktur.

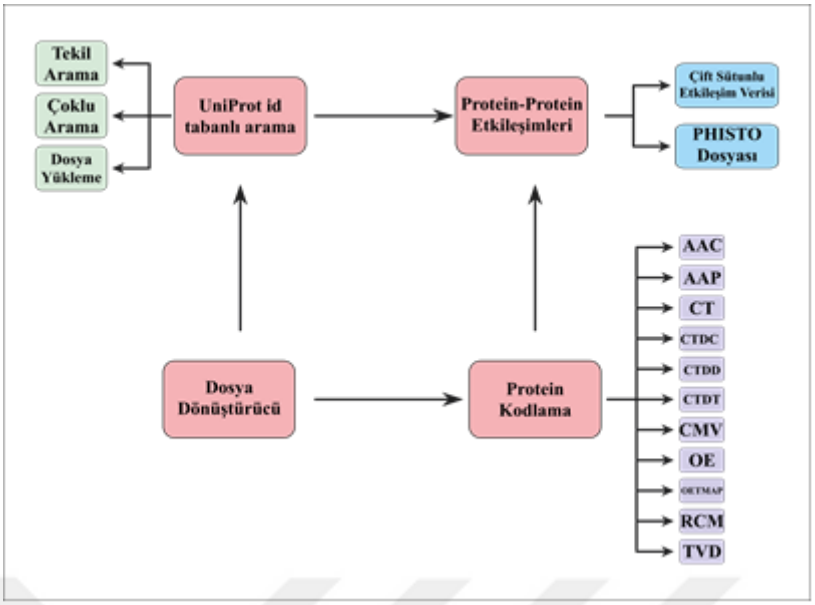
Tezde önerilen yöntemlerin doğrulanması amacı ile kullanılan veri setleri, veri tabanlarından indirilmesinden tahmin metotlarına verilmesine kadar birçok ön işlemden geçmiştir. Veriler üç farklı kaynaktan (STRING, PHISTO ve UniProt) sorgulanmıştır. PHISTO ve STRING veri tabanlarından proteinler arası etkileşim verileri çekilmiştir. UniProt veritabanından deneylerde kullanılan organizmalara ait amino asit dizilimlerinin tümü indirilmiştir. Etkileşim ağında yer alan proteinlere ait öznitelik vektörünün çıkarılması için ayrıca bu proteinlerin amino asit dizilimlerinin bilinmesi gerekmektedir. Ağdaki proteinlere ait dizilimler UniProt veri tabanından indirilen veri içerisinden UniProt id anahtarı kullanılarak çekildi ve etkileşimde yer alan proteinler ile eşlendi. Yapılan eşlemeden sonra her bir amino asit dizilimi farklı yöntemler ile kodlandı. Şu ana kadar yapılan işlemler ile veri setindeki pozitif etiketli sınıf oluşturuldu. Negatif sınıftaki örnek sayısı yapılan deneysel çalışmaya göre belirlenmektedir. İstenilen sayıda negatif örnek, her bir organizmanın proteinleri arasından rastgele seçilerek eşlenir. Negatif örneklere ait dizilimler de benzer yöntemler ile kodlanır ve pozitif sınıf ile birleştirilerek veri seti oluşturulur.

Yukarıda anlatılan işlem adımlarına ait kodlar PHP web programlama dili ile kodlanarak PROSES yazılımı tasarlandı. Tezde yapılanlara ek olarak literatürde geçen birçok protein kodlama yöntemi de yazılıma eklendi. Yazılım, protein kodlama, protein etkileşimi, arama ve dosya dönüştürme olmak üzere dört ana modülden oluşmaktadır. Bu modüller bir sonraki başlıkta detaylı olarak açıklanmıştır. PROSES şu anda Yalova Üniversitesi resmi web sunucusu altında yer alan http://proses.yalova.edu.tr adresinde kullanıma açılmıştır

4.3.1. PROSES modülleri

PROSES, kullanıcı dostu ve esnek bir ara yüz ile tasarlanmıştır. Yazılım protein kodlama, arama, protein-protein etkileşimi ve dosya dönüştürücü olmak üzere dört ana modülden oluşmaktadır. Modüllerin her biri farklı bir sayfada yer almakla birlikte herhangi bir modül içinde yer alan fonksiyon diğer modüller içinde de kullanılmaktadır. Örneğin arama modülünde proteinlerin UniProt kodlarına göre yapılan amino asit dizi sorgulama fonksiyonu hem protein kodlama modülünde hem de protein-protein etkileşim modülünde kullanılmaktadır. Modüllere ait diyagram ve modüller arası ilişki Şekil 4.5.'te görüldüğü gibidir.

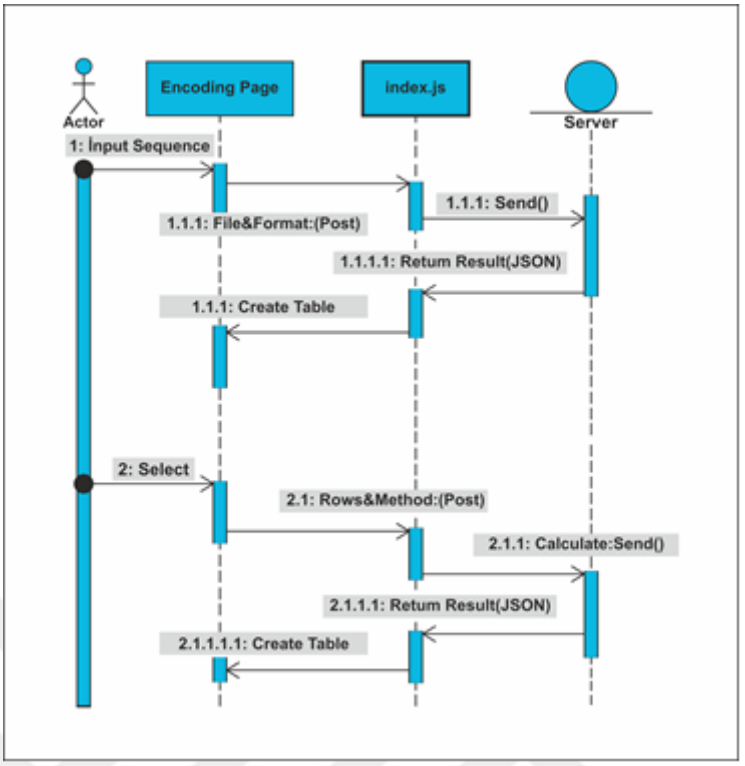
PROSES yazılımında, tümü dizilim tabanlı olmak üzere 12 farklı protein kodlaması yer almaktadır. Bu yöntemlerden amino asit kompozisyon (AAC), amino asit çifti (AAP), bitişik üçlü (CT), kompozisyon (CTDC), geçiş (CTDT), dağılım (CTDD) ve dipeptit kompozisyon (DC) yöntemleri amino asit dizi uzunluğundan bağımsız olarak eşit uzunlukta öznitelik vektörleri üretmektedir. Kompozisyon moment vektörü (CMV) yönteminde, M kaçıncı dereceden moment alınacağını belirten parametre olmak üzere, öznitelik vektör boyutu $20 \times M$ olmaktadır. Amino asit eşleme modeli (RCM) yönteminde, R öznitelik vektörünün rank parametresi olmak üzere, öznitelik vektör boyutu $R \times 400$ olmaktadır. Geri kalan Ortonormal kodlama (OE), OETMAP, Taylor Venn Diyagramı (TVD) yöntemlerinde ise öznitelik vektör boyutu protein dizi uzunluğu ile orantılı olmaktadır (Yöntemlerin detaylı açıklaması için bakınız bölüm 3.3).



Şekil 4.5. PROSES modülleri ve modüller arası ilişki diyagramı

Kullanılan kodlama yöntemlerinin her biri ayrı fonksiyon olarak yazılmıştır. Girilen aminoasit dizilimleri kullanıcının seçimine göre ilgili yönteme ait fonksiyonun çağrılması ile kodlanmaktadır. Kodlama modülüne ait çalışma prensibi Şekil 4.6.'da verilen akış diyagramında görülmektedir.

Protein etkileşim modülü, etkileşim ağında yer alan protein çiftlerini kodlama modülünde yer alan yöntemlerden herhangi biri ile kodlar ve her bir protein çiftine ait bir öznitelik vektörü üretir. Veriler sisteme, etkileşim halinde olan protein çiftlerine ait UniProt kodların her biri bir sütunda olan dosya halinde veya doğrudan PHISTO veri tabanından indirilen dosya olarak verilir. Etkileşim tablosunda yer alan her bir proteine ait aminoasit dizilimi arama modülü kullanılarak sorgulanır. Kullanıcının seçtiği kodlama yöntemine göre diziler kodlanır. Kodlanan proteinler etkileşim tablosunda verilen protein çiftlerine bakılarak öznitelik vektörleri birleştirilir. Sistem, aynı zamanda etkileşim ağı içerisinde yer alan proteinlerin tümüne ait amino asit dizilimlerinin tek dosyada indirilmesine olanak tanır.

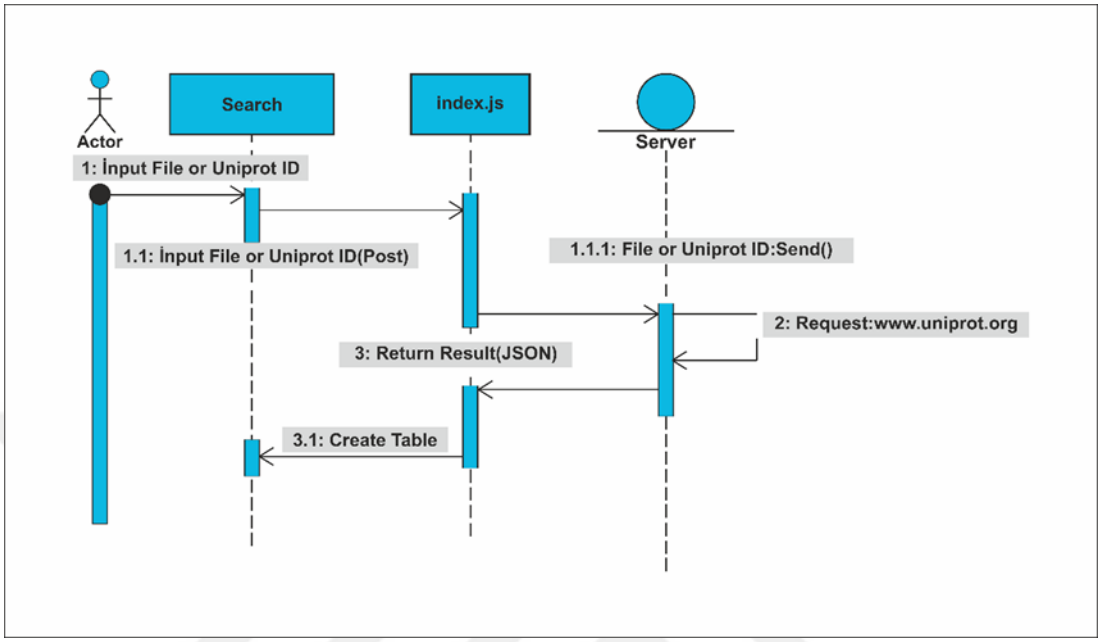


Şekil 4.6. Protein kodlama modülüne ait akış diyagramı

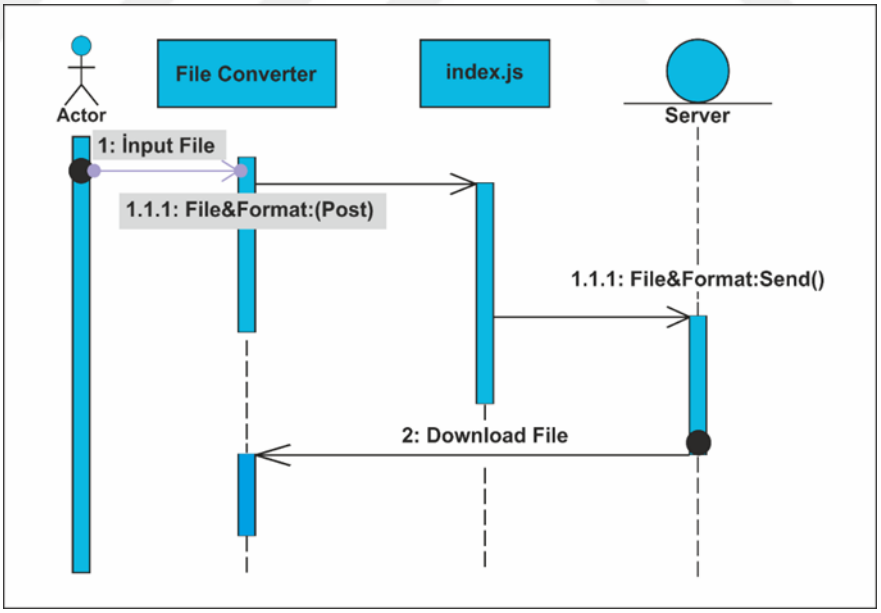
Arama modülü, kodlanmak istenen proteinlerin amino asit dizilimine PROSES üzerinden erişmeyi sağlar. Dizilimler UniProt koduna göre http://www.uniprot.org/ adresindeki veri tabanından sorgulanmaktadır. Sorgulamada UniProt kodu, yazılımda yer alan metin kutusu aracılığıyla manüel olarak veya kodların tek sütun halinde yer aldığı dosyanın sisteme yüklenmesi ile girilebilir. Veriler girildikten sonra her bir proteine ait kod, ilgili fonksiyona parametre olarak verilir. Fonksiyon, parametre olarak aldığı UniProt kimliğine ait dizilimi geri döndürür. Şekil 4.7.'de arama modülünün çalışmasına ait akış diyagramı verilmiştir.

PROSES yazılımında kullanıcıların farklı dosya formatları ile çalışmasına imkân vermek ve sistem içinde üretilen verileri farklı formatlara dönüştürmek amacıyla dosya dönüştürme modülü hazırlanmıştır. Dosya dönüştürme fonksiyonları aynı zamanda diğer modüllerin ürettiği verileri kullanıcıların farklı formatta indirmelerine olanak sağlamaları için de kullanılmaktadır. Proteinlere ait amino asit dizilimleri genellikle fasta formatında tutulur. PROSES yazılımında da protein kodlama modülü verileri fasta formatında almaktadır. Ancak verilerin kullanıcılar tarafından sıkça kullanılan "txt, xls, xlsx, arff" gibi formatlara da çevrilmesinin faydalı olacağı

düşünülerek bu modül hazırlanmıştır. Dosya dönüştürme modülünün işleyişine ait akış diyagramı Şekil 4.8.'de verilmiştir.



Şekil 4.7. Arama modülüne ait akış diyagramı



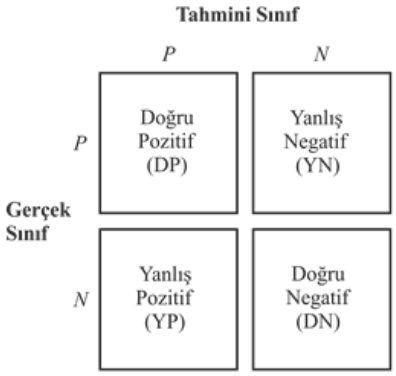
Şekil 4.8. Dosya dönüştürme modülüne ait akış diyagram

BÖLÜM 5. ARAŞTIRMA BULGULARI

Bu bölümde, önerilen iki yöntemin başarısı test edilmiş ve değerlendirme sonuçları verilmiştir. Yöntemlerin doğrulanmasında kullanılan değerlendirme metrikleri açıklanmıştır. İlk deneyde genişletilmiş ağ modelinin etkileşim tahmininde yaptığı iyileştirmeler ölçülmüştür. Bu amaçla konak canlının insan, patojen canlının bacillus anthracis ve ebola olduğu veri setleri kullanılmıştır. İkinci deneyde LTK yönteminin tahmin doğruluğunu nasıl etkilediği test edilmiştir. LTK yönteminin doğrulanmasında yersinia pestis ve bacillus anthracis patojenlerinin insan proteinleri ile yaptığı etkileşim verileri kullanılmıştır.

5.1. Değerlendirme Metrikleri

PKE probleminde tahmin metotları, ikili sınıflandırma (var/yok, doğru/yanlış v.b.) yapmaktadır. Deneysel çalışmalarda önerilen yöntemlerin başarı/başarısızlık değerlendirmesi doğruluk (accuracy), özgüllük (precision), duyarlılık(recall), F1, AUC ve MCC metrikleri kullanılarak test edildi. Bu metrikler karmaşıklık matrisi (confusion matrix) denilen 2×2 boyutunda bir matristen türetilmektedir. Karmaşıklık matrisinde, tahmin modelinin test aşamasında yaptığı tahminler (predicted class) ve tahmin edilen verilerin gerçek sınıfları (actual class) yer alır. Şekil 5.1.'dekarmaşıklık matrisi verilmiştir. Bu matriste doğru pozitif (DP) ve doğru negatif (DN) sırasıyla doğru tahmin edilen pozitif ve negatif örnek sayılarıdır. Yanlış pozitif (YP) ve yanlış negatif (YN) değerleri ise sırasıyla yanlış tahmin edilen pozitif ve negatif örnek sayılarıdır.



Şekil 5.1. Karmaşıklık Matrisi

Modelin, test verilerini sınıflandırmasının ardından pozitif ve negatif etiketli örneklerin sınıflandırma sonucuna bakılarak karmaşıklık matrisi oluşturulur.

Denklem 5.1'de verilen doğruluk metriği, tahmin modelinin negatif ve pozitif ayrımı yapmadan, doğru tahmin edilen örnek sayısının tüm örnek sayısına oranıdır. Doğruluk metriği sınıflandırmanın genel başarısı hakkında fikir vermektedir. Bu metriğe bakarak modelin hangi sınıfı daha iyi tahmin ettiği anlaşılmaz. Doğruluk, veri setinde sınıflara ait örnek sayısının çok farklı olduğu durumlarda, model başarımını ölçmede kullanılacak sağlıklı bir metrik değildir. Modelin öğrenme aşamasında negatif veya pozitif sınıfı daha iyi öğrenmiş olma ihtimali vardır. Bu durumda model test için verilen tüm örnekleri tek bir sınıfa atayabilir. Örneğin, tezin ilk deneysel çalışmasında negatif örneklerin pozitiflere oranı 1/10 olarak seçilmiştir. Bu veri seti için negatif lehine öğrenmenin gerçekleştiği model tüm pozitif örnekleri yanlış tahmin edip, negatifleri doğru tahmin etmesi durumunda bile doğruluk oranı %90 olmaktadır. Dolayısıyla bu tür veri setlerinin değerlendirilmesinde doğruluk tek başına, sınıflandırma başarısı hakkında fikir vermez.

$$Doğruluk = \frac{DP + DN}{DP + YP + DN + YN} \tag{5.1}$$

Tahmin modellerini sağlıklı bir şekilde analiz etmek için başka metriklere de ihtiyaç duyulur. Sınıflandırma probleminde doğru tahmin edilmesi istenen asıl örnekler hedef sınıf olarak verilir ve pozitif olarak etiketlenir. Bu çalışmada aralarında

etkileşim olan proteinler pozitif olarak etiketlenmiştir. Pozitif sınıfın ne kadar iyi tahmin edildiğini görmek için özgüllük ve duyarlılık metrikleri kullanılmaktadır.

Denklem 5.2'de verilen özgüllük metriği, doğru bulunan pozitif örnek sayısının, tahmin sonucu bulunan pozitif örnek sayısına oranıdır. Özgüllük metriğine göre modelin pozitif sınıfı ne kadar iyi tahmin ettiği anlaşılabilir. Düşük özgüllük değeri, negatif olarak etiketlenmesi gereken örneklerin doğru tahmin edilmediğini gösterir. Bu metrik negatif örneklerin tahmin başarısı hakkında fikir vermez.

$$\text{Özgüllük} = \frac{DP}{DP + YP} \tag{5.2}$$

Denklem 5.3'te duyarlılık metriğinin elde edilmesi için kullanılan denklem verilmiştir. Duyarlılık, değeri doğru bulunan pozitif örnek sayısının veri setindeki pozitif etiketli örnek sayısına oranıdır. Herhangi bir tahmin modelinde duyarlılık değerinin düşük olması modelin pozitif olarak sınıflanması gereken örnekleri negatif olarak etiketlediği anlamına gelir.

$$Duyarlılık = \frac{DP}{DP + YN} \tag{5.3}$$

Özgüllük ve duyarlılık metrikleri ayrı ayrı sınıflandırma başarısı hakkında fikir vermektedir. Bu metrikler arasındaki farkın fazla olması, sınıflandırmanın pozitif veya negatif sınıf lehine olduğunu göstermektedir. Bu iki metriğe ayrı ayrı bakmak yerine her ikisinin kullanımıyla bulunan F1 skor değeri her iki metrik hakkında fikir verir. Denklem 5.4'te verilen F1 skor değeri özgüllük ve duyarlılık metriklerinin harmonik ortalaması alınarak bulunur. Özgüllük ve duyarlılık değerlerinden birinin yüksek, diğerinin düşük olması durumunda ikisinin harmonik ortalamasını veren F1 skor değerine bakarak sınıflandırmayı daha sağlıklı yorumlamak mümkündür.

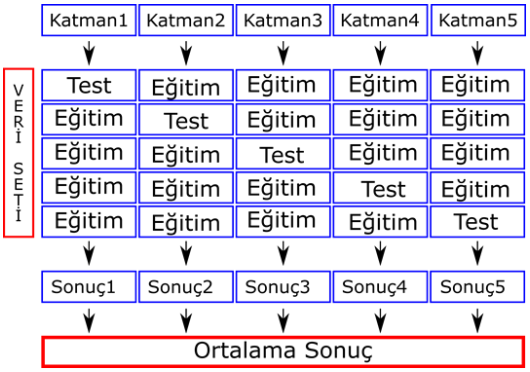
$$F1 = \frac{2 * \text{Özgüllük} * Duyarlılık}{\text{Özgüllük} + Duyarlılık} \tag{5.4}$$

Deneysel çalışmalarda kullanılan bir diğer değerlendirme metriği MCC (Mathews's correlation coefficient) değeridir. Denklem 5.5'te verilen karmaşıklık matrisine göre MCC'nin nasıl hesaplandığı görülmektedir.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad (5.5)$$

Yapılan deneylerde modellerin başarıları özgüllük-duyarlılık grafikleri kullanılarak da görsel ve sayısal olarak değerlendirilir. F1 skor değerine benzer şekilde sınıflandırma başarısı, çizdirilen özgüllük-duyarlılık eğrisine bakarak yorumlanır. Bu yorumun sayısal sonucu eğrinin altında kalan alanın hesaplanması ile çıkarılır.

Deneysel sonuçlar, kalite metriklerinin yanı sıra veri setinin test ve eğitim olarak nasıl ayrıldığına göre de değişir. Modelin eğitimi sırasında kullanılan verilerin, test aşamasında kullanılmaması gerekmektedir. Bu nedenle veri seti test ve eğitim verisi olarak ayrılır. Literatürde yapılan çalışmalarda veri setinin eğitim ve test olarak ayrılmasında farklı yollar izlenmektedir. Bazı çalışmalarda test verisi eğitim verisine eşit veya küçük olacak şekilde belirli bir yüzdeye göre ayrılır. Kullanılan bir diğer yöntem ise çapraz doğrulama (cross validation) yöntemidir. Çapraz doğrulama yönteminde veri seti, kullanıcının seçtiği 'k' sınıfa ayrılır. Bu sınıflardan 'k-1' tanesi eğitim sırasında, geri kalan sınıf ise test aşamasında kullanılır. Bu işlem her seferinde farklı bir test sınıfının kullanılması ile 'k' kez tekrar eder. Her tekrar sonrası gelen sonuçlar toplanıp ortalama alınır. Ortalama sonuç modelin başarımı olarak kabul edilir. Şekil 5.2.'de çapraz doğrulama yönteminin çalışma mantığına ait görsel verilmiştir.



Şekil 5.2. Çapraz doğrulama yöntemi ile veri setinin test ve eğitim verisi olarak ayrılıp sonucun değerlendirilmesi.

5.2. Deneysel Çalışma 1

Bu bölümde genişletilmiş ağ modeli yönteminin başarısı test edilip yorumlanmıştır. Önerilen yöntemin başarısı farklı veri setleri üzerinde, farklı tahmin ve kodlama yöntemlerinin çaprazlanması ile test edilmiştir. Konak ve patojen organizmalara ait proteinlerin amino asit dizilimleri AAC, AAP ve CTDC yöntemleri (Yöntemlerin detayı için bakınız bölüm 3.3) ile kodlanarak öznitelikler çıkarılmıştır. Kodlanmış patojen ve konak canlılara ait proteinler arası etkileşimler BN, NB, RF, C4.5, kNN, k* ve MF yöntemleri (detaylar için bakınız bölüm 3.2) ile sınıflandırılmıştır.

5.2.1. Veriseti

Birinci deneysel çalışmada genişletilmiş ağ modeli yöntemi, bacillus anthracis ve ebola patojenlerinin insan proteinleri ile yaptığı etkileşim verileri üzerinde test edilmiştir.

Genişletilmiş ağ modeli patojen ve konak proteinler arası bilinen etkileşimlerin yanı sıra ağ içindeki proteinlere ait tür içi etkileşim ağlarını da kullanmaktadır. Bu sebeple türler arası etkileşim verileri, mevcut yöntemlerin uygulanması için yeterli olurken, genişletilmiş ağ modeli türler arası etkileşimlerin yanı sıra tür içi etkileşimlere de ihtiyaç duyar.

Deneysel çalışmada kullanılan türler arası etkileşim verileri PHISTO veri tabanından indirildi. Etkileşim ağında her bir türe ait proteinler ayrılıp, bu proteinlerin tür içi etkileşimleri STRING veri tabanından sorgulandı. Yapılan sorgulamada çok sayıda etkileşim verisi dönmektedir. Yöntemin doğrulanması için kullanılan veri setinde tür içi ve türler arası yöntemlere ait örnek sayısının bir birine yakın olması gerekmektedir. Bu sebeple STRING veri tabanından yapılan sorguda etkileşimler güvenirliklerine göre sıralanarak PKE ağında yer alan örnek sayısına eşit olacak şekilde ayrıldı.

Tablo 5.1.'de çalışmada kullanılan ebola ve bacillus verilerine ait bilgiler verilmiştir. Tabloda, organizmaların bilinen tüm protein ve PKE ağında geçen farklı protein sayıları yer almaktadır. Her iki veri setinde de konak olarak insan proteinleri seçilmiştir. Tabloya bakıldığında türler arası PK, patojen organizma içi PP ve konak organizma içi HH etkileşim sayıları görülmektedir.

Aralarında etkileşimin olmadığı kabul edilen negatif etiketli veriler patojen ve konak organizmalara ait proteinlerin rastgele eşlenmesi ile oluşturulmuştur. Literatür çalışmalarında veri setinde kullanılan pozitif örnek sayısı genellikle negatif örnek sayısından az olmaktadır. Bu deneyde pozitif örneklerin negatif örneklere oranı 1/10 olarak seçildi.

Tablo 5.1. Veri setinde olan proteinler ve ağ içindeki etkileşim sayıları.

	B. anthracis	Ebola
Protein sayısı	5493	90
Ağ içindeki protein sayısı	907	3
Bilinen PK etkileşim sayısı	3050	147
Negatif etiketli etkileşim sayısı	30500	1470
Ağ içindeki konak protein sayısı	1568	399
Ağ içindeki HH etkileşim sayısı	1550	147
Ağ içindeki PP etkileşim sayısı	1500	-

Tabloda da görüldüğü gibi ebola virüsünün etkileşim ağındaki protein sayısı bacillus'a göre daha azdır. Ayrıca bu virüsün proteinleri arasındaki etkileşimlere erişilemediğinden bizim yöntem için Şekil 4.1.'de verilen etkileşim matrisinde PP alanı boş kalmaktadır. Önerdiğimiz yöntem, ilişki matrisinde PP ve HH alanlarından birinin eksik olması veya olmaması durumunda da mevcut verilere göre sınıflandırma yapabilmektedir.

5.2.2. Veri setlerinin ayrık değerlendirilmesi

Bacillus anthracis ve Ebola veri setlerine ait sonuçlar sırasıyla Tablo 5.2. ve Tablo 5.3.'te verilmiştir. Yöntemlerin test sonuçları, veri seti, protein kodlama ve tahmin yöntemi bazında yorumlanmıştır. Yorumlar kodlama yöntemlerinin tahmin metotlarına göre başarısı, tahmin metotlarının kodlama yöntemlerine göre kıyaslanması ve bu sonuçların ebola ve bacillus veri setlerinde nasıl değiştiği şeklinde yapılmıştır.

AAC kodlama ile yapılan tahmin sonuçlarına bakıldığında GAM yönteminin tüm deneylerde F1 sonucunu iyileştirdiği görülmektedir. Deneylerde F1 sonucunun hem özgüllük hem duyarlılık metrikleri ile birlikte yükselmesi önemlidir. Bu durum dikkate alınarak tablodaki sonuçlara bakıldığında, RF ve MF metotlarının duyarlılık değerleri dışında, tüm deneyler için her iki metriğin arttığı görülmektedir. Tüm tahmin metotları kıyaslandığında en iyi F1 sonucu klasik tahmin yönteminde BN, GAM yönteminde ise RF yöntemi ile alınmıştır. RF yöntemi ile klasik yöntemde çok düşük duyarlılık sonucu alınmıştır. Düşük duyarlılık, pozitif örneklerin iyi tahmin edilmediğini göstermektedir. GAM yönteminde bu değer özgüllük sonucu ile birlikte yükselmiş bu da F1 sonucuna yansımıştır. Örnek tabanlı kNN ve K* sonuçlarına bakıldığında GAM yöntemi ile sonuçların iyileşmesi ile beraber bu iki yöntemin birbirine yakın sonuçlar verdiği görülmektedir. MF tahmin metodu, hem klasik hem GAM veri setinde yüksek duyarlılık, düşük özgüllük değerleri üretmiştir. Bu da yöntemin, örnekleri pozitif olarak sınıflandırma eğiliminde olduğunu göstermektedir. MF yönteminde, sayıca fazla olan negatif örneklerin yanlış sınıflandırılmasından dolayı doğruluk değerleri düşük çıkmıştır. Benzer sebepten dolayı, NB yöntemi ile de düşük doğruluk değerleri elde edilmiştir.

AAP kodlama ile yapılan deneylerin tümünde F1 sonucunun arttığı görülmektedir. En iyi F1 sonucu klasik veri setinde BN yöntemi ile alınırken, GAM veri setinde kNN ve RF yöntemleri ile alınmıştır. kNN ile yapılan tahminlerde özgüllük, duyarlılık metrikleri arasındaki fark daha az olduğundan dolayı daha sağlıklı sonuçlar verdiği söylenebilir. kNN ve K* örnek tabanlı sınıflandırıcılarda GAM yöntemi ile

başarımın önemli ölçüde arttığı görülmektedir. Bu yöntemlerde özgüllük ve duyarlılık sonuçları yakın olduğundan dolayı pozitif ve negatif örnek sınıflamasında hata oranları neredeyse aynıdır. MF yöntemi ile yapılan sınıflandırmada duyarlılık oranı AAC yöntemine göre daha düşüktür. Ancak burada pozitif sınıflandırma eğilimi AAC kodlamasına göre daha azdır. NB yöntemine bakıldığında ise tüm metriklerde AAP kodlamanın AAC'ye göre daha iyi sonuçlar verdiği görülmektedir.

Tablo 5.2. Bacillus Anthracis veri seti için deneysel sonuçlar

		PKE				GAM			
Öznitelik	Metot	F1	Özg.	Dyr.	Dğr.	F1	Özg.	Dyr.	Dğr.
	BN	0.4090	0.3540	0.4840	87.31	0.6230	0.5430	0.7320	85.50
	NB	0.3450	0.2280	0.7080	75.58	0.5420	0.4140	0.7850	78.23
	RF	0.1650	0.9620	0.0900	91.70	0.6290	0.9410	0.4720	90.86
AAC	C4.5	0.2700	0.3010	0.2450	87.97	0.5670	0.5740	0.5610	85.98
	kNN	0.3090	0.2770	0.3500	85.87	0.5550	0.4770	0.6620	82.57
	K*	0.3330	0.2350	0.5740	79.12	0.5380	0.4060	0.7960	77.54
	MF	0.2142	0.1259	0.8342	40.87	0.3706	0.2614	0.7342	56.07
	BN	0.3960	0.2790	0.7110	80.33	0.5710	0.4750	0.7160	82.15
	NB	0.3520	0.3140	0.4010	86.62	0.4590	0.5240	0.4090	84.01
	RF	0.2640	0.7860	0.1590	91.96	0.6460	0.9110	0.5010	90.89
AAP	C4.5	0.3480	0.3670	0.3320	88.72	0.6060	0.6180	0.5950	87.16
	kNN	0.3370	0.7180	0.2200	92.13	0.6460	0.8660	0.5150	90.62
	K*	0.4700	0.6400	0.3720	92.10	0.5870	0.7460	0.4840	87.34
	MF	0.2348	0.2094	0.4715	67.45	0.4262	0.8356	0.3533	76.65
	BN	0.3010	0.2100	0.5340	77.48	0.5000	0.3840	0.7140	76.22
	NB	0.2730	0.1720	0.6610	68.01	0.4570	0.3220	0.7880	68.86
	RF	0.1520	0.8850	0.0830	91.57	0.6480	0.9580	0.4900	91.16
CTDC	C4.5	0.1820	0.3030	0.1300	89.38	0.5730	0.6200	0.5320	86.81
	kNN	0.2460	0.2840	0.2170	87.93	0.6090	0.5790	0.6420	86.29
	K*	0.2800	0.2220	0.3790	82.30	0.5680	0.4690	0.7180	81.81
	MF	0.2068	0.1234	0.7389	47.10	0.3313	0.2110	0.8309	42.81

CTDC kodlamada, klasik veri seti ile en iyi F1 sonucu BN sınıflandırıcı ile alınmıştır. Diğer kodlama yöntemlerine benzer şekilde GAM veri setinde, en iyi sonuç RF sınıflandırıcı ile alınmıştır. RF sınıflandırıcıda pozitif örnek sayısının artması her üç kodlama yönteminde de tahmin doğruluğunu arttırmış, ayrıca her üç yöntemde de düşük duyarlılık, yüksek özgüllük değerleri gözlendi. Karar ağacı yöntemleri kendi aralarında kıyaslandığında klasik veri setinde C4.5 metodunun RF'e göre daha iyi sonuç verdiği görülmüştür. C4.5 metodunun, GAM veri setinde RF'e göre daha düşük F1, daha yüksek duyarlılık sonucu ürettiği görülmektedir. CTDC kodlamada dikkat çeken bir diğer sonuç MF tahmin yönteminin çok düşük doğruluk değerine sahip olmasıdır. MF yönteminde her iki veri seti ile düşük özgüllük ve duyarlılık değerinden dolayı F1 ve doğruluk sonuçları da diğer yöntemlere göre daha düşük kalmıştır. GAM veri seti örnek tabanlı sınıflandırıcılar

olan Knn ve K* yöntemlerinde, F1 sonucunu %100 oranından fazla iyileştirmiştir. Karar ağacı yöntemleri olan RF ve C4.5 yöntemlerinde de benzer şekilde GAM veri seti ile tahmin sonuçlarında %100'den fazla bir iyileşme görülmektedir.

Bacillus veri setindeF1 skoru göz önüne alınıp, her üç protein kodlama yöntemi karşılaştırıldığında klasik veri setinde en iyi tahmin, AAC kodlama ve BN sınıflandırıcı ile alınmıştır. GAM veri setinde en iyi F1 sonucu CTDC protein kodlaması ve RF tahmin metodu ile alınmıştır. Doğruluk bizim yöntem için her üç kodlama yönteminde de RF tahmin metodunda en yüksek değere ulaşmıştır. RF sınıflandırıcı, diğer metrikler de göz önüne alındığında GAM yönteminde etkileşim tahmini için yüksek doğrulukta olduğu görülmüştür. Ancak PKE veri seti ile yapılan deneylerde pozitif sınıf tahmininde yetersiz kalmıştır.

Ebola veri setinde AAC kodlama ile en yüksek F1 0.756 ile RF sınıflandırıcıda görülmüştür. Bu değer bizim yöntemle 0.1390 arttırılarak 0.8950 olurken, bizim yöntemin, tüm metrikler için tahmin doğruluğunu arttırdığı görülmüştür. Ebola veri setinde AAC kodlama ve RF ile bulunan duyarlılık değerine bakıldığında bacillus verisinden farklı olarak negatif sınıf tahmininde de başarılı olduğu görülmektedir. Sınıflandırıcılar, ebola veri seti ve AAC kodlama yöntemi ile daha yüksek tahminler yapmıştır. Doğruluk değerlerine bakıldığında bizim yöntemin, en iyi artışı BN yönteminde %10 oranında arttırdığı görülmüştür. Bizim yöntem BN dışında NB, RF, K* ve MF sınıflandırıcıları içinde doğruluk değerini arttırmıştır.

Tablo 5.3. Ebola veri seti için deneysel sonuçlar

		PKE				GAM			
Öznitelik	Metot	F1	Özg.	Dyr.	Dğr.	F1	Özg.	Dyr.	Dğr.
	BN	0.4790	0.3170	0.9770	80.68	0.7690	0.6300	0.9890	90.16
	NB	0.5760	0.4170	0.9300	87.56	0.8010	0.8600	0.7470	93.82
	RF	0.7560	0.8490	0.6820	96.01	0.8950	0.9340	0.8590	96.65
AAC	C4.5	0.7310	0.7210	0.7420	95.05	0.8240	0.8280	0.8210	94.19
	kNN	0.7140	0.6420	0.8030	94.15	0.8230	0.7770	0.8700	93.75
	K*	0.6480	0.5470	0.7950	92.16	0.7960	0.7200	0.8900	92.43
	MF	0.2994	0.2017	0.7502	64.39	0.4406	0.4938	0.6077	70.51
	BN	0.4790	0.3170	0.9770	80.68	0.6480	0.4820	0.9850	82.22
	NB	0.4600	0.3010	0.9770	79.17	0.4460	0.2930	0.9320	61.60
	RF	0.7020	0.8490	0.5980	95.39	0.8620	0.9220	0.8100	95.71
AAP	C4.5	0.6720	0.6940	0.6520	94.22	0.8100	0.8100	0.8100	93.69
	kNN	0.7060	0.7410	0.6740	94.91	0.8720	0.8770	0.8670	95.77
	K*	0.5470	0.6200	0.4900	90.03	0.6510	0.7030	0.6070	83.29
	MF	0.3547	0.3857	0.4746	84.47	0.4045	0.4431	0.5104	75.73

Tablo 5.3. (Devamı)

	BN	0.6420	0.4780	0.9770	90.10	0.7850	0.6520	0.9850	91.04
	NB	0.5450	0.3880	0.9170	86.11	0.6360	0.6130	0.6620	87.45
	RF	0.7380	0.7750	0.7050	95.46	0.8620	0.8940	0.8330	95.58
CTDC	C4.5	0.1820	0.3030	0.1300	89.38	0.5730	0.6200	0.5320	86.81
	kNN	0.6390	0.6340	0.6440	93.40	0.7810	0.7680	0.7950	92.62
	K*	0.7020	0.6290	0.7950	93.88	0.8320	0.7860	0.8820	94.07
	MF	0.1811	0.1013	0.9036	25.97	0.4260	0.7044	0.4221	78.71

Ebola AAP kodlamasında, RF ve kNN sınıflandırıcıları ile birbirine yakın sonuçlar vermiştir ve diğer sınıflandırıcılardan daha iyi performans göstermiştir. Özgüllük-duyarlılık arasındaki fark RF için daha büyükken kNN için bu sonuçlar daha dengelidir. kNN bizim yöntemde tüm metrikler için daha yüksek sonuç vermekle beraber, 0.872 F1 ve 95.77 doğruluk değeri ile en yüksek sonuçlar alınmıştır. F1 sonuçlarına bakıldığında bizim yöntemin NB haricindeki tüm sınıflandırıcılarda tahmin doğruluğunu arttırdığı görülmüştür. kNN, RF ve BayesNet sınıflandırıcıların doğruluk değerleri bizim yöntemde daha yüksek çıkmıştır.

Ebola veri setinde CTDC kodlamasında en yüksek F1 0.73 ile RF sınıflandırıcısında gözlenmiştir. Bu sonuç bizim yöntemle 0.16 artarak 0.862 olmuştur. F1 ve doğruluk sonucuna göre en iyi ikinci sınıflandırıcı K* olmuştur. Bu kodlama yöntemine göre kNN ve BN sınıflandırıcılarda doğruluk sonuçları %80 üzerinde ve özgüllük-duyarlılık değerleri dengelidir. NB ve MF sınıflandırıcıları düşük özgüllük sonucu ile pozitif etkileşimlerin tahmininde yetersiz kalırken bizim yöntemde bu değerler yükselmiş olup daha dengeli sonuçlar gözlenmiştir.

Ebola veri seti bacillus verilerine kıyasla daha az etkileşim içermektedir. Ancak kodlama yöntemleri ve sınıflandırıcılar için genel sonuçlara bakıldığında bacillus verileri ile alınan sonuçlar arasında paralellik vardır.

Bacillus veri setinde, tüm protein kodlama yöntemleri ve sınıflandırıcılar için F1 sonuçları göz önüne alındığında en iyi sınıflandırıcının K*, en iyi kodlama yönteminin ise AAP olduğu görülmektedir. Bu sınıflandırıcı-kodlama eşlemesi için bizim yöntem F1 değerini 0.47'den 0.587'e yükseltmiştir. Bizim yöntem için etkileşim tahmininin en yüksek F1 sonucu CTDC kodlaması ile birlikte kullanılan RF sınıflandırıcısıdır. Sınıflandırıcıların farklı kodlama yöntemleri ile kullanımında

genel olarak önemli bir değişim görülmemiştir. Kodlama yönteminin sınıflandırıcı üzerindeki etkisi en fazla K* ve RF yöntemlerinde AAP ve CTDC kodlamaları arasında gözlenmiştir. Diğer yöntemlerde kodlamanın etkisi 0,1'den daha düşük olmuştur.

Ebola veri setinde, tüm protein kodlama yöntemleri ve sınıflandırıcılar için F1 sonuçları göz önüne alındığında 0.756 ile en iyi sınıflandırıcının RF, en iyi kodlama yönteminin ise AAC olduğu görülmektedir. Bu sınıflandırıcı-kodlama eşlemesi için GAM yöntemi F1 değerini 0.139 arttırarak 0.895'e yükseltmiştir. Bizim yöntemde de etkileşim tahmininde en yüksek F1, RF ve AAP yöntemleri ile elde edilmiştir. Bu veri setinde bizim yöntemde tahmin doğruluğunun düştüğü tek yöntem AAP kodlaması ile kullanılan NB sınıflandırıcısıdır. Sınıflandırıcılar arasındaki tahmin doğruluğu kıyaslandığında genel olarak MF yönteminin diğer sınıflandırıcılardan geride olduğu görülmüştür.

Bacillus ve ebola veri setleri için tablolarda verilen sonuçlara bakıldığında genişletilmiş ağ modelinde önerilen hipotezi desteklediği görülmektedir. Her iki veri seti ile yapılan 42 deneyin 41'inde F1 skor değerinin genişletilmiş ağ modeli ile arttığı görülmektedir. F1 skor değerinde kullanılan özgüllük ve duyarlılık metriklerine bakıldığında 42 deneyin 40'ında daha iyi özgüllük, 42 deneyin 36'sında daha iyi duyarlılık değerine ulaşıldığı görülmektedir. Bazı deneylerde çok yüksek özgüllük sonucu elde edilmesine rağmen duyarlılık sonucunun çok düşük olduğu (özellikle RF metodunda) görülmektedir. Bu durum protein etkileşim tahmininde modelin negatif sınıf lehine bir öğrenme gerçekleştirildiğini göstermektedir.

Patojen ve konak organizmaların proteinleri arası var olan etkileşimler olası tüm etkileşimlerin çok az bir kısmını oluşturmaktadır. Bu sebeple patojen ve konak proteinler arası oluşan etkileşim ağlarının seyrek olduğu söylenebilir. Örneğin Tablo 5.1.'de bacillus anthracis ve insan proteinleri arası etkileşim bilgilerine bakıldığında konak ve patojen organizmalar arası bilinen etkileşimlerin 3050 olduğu görülmektedir. Bu etkileşim ağında patojen canlıya ait 907 protein, konak canlıya ait 1568 protein olduğu görülmektedir. Protein sayıları göz önüne alındığında patojen ve

konak organizma arası olası tüm etkileşim sayısı yaklaşık 1,4 milyondur (907×1568). Dolayısıyla etkileşim ağında yer alan bilinen etkileşimler olası tüm etkileşimlerin %0,2'si kadardır. Deneysel çalışmalarda veri setleri oluşturulurken bu durum göz önüne alınarak öğrenme ve test aşamasında kullanılan veri setlerinde negatif etkileşim sayısının pozitif etkileşimlerden sayıca fazla olmasına dikkat edildi. Bu deneysel çalışmada pozitif etkileşimlerin negatif etkileşimlere oranı 1/10 olarak seçildi.

Negatif veri setinin pozitif verilere oranın fazla olması her iki veri seti içinde değerlendirme metriklerini etkilemiştir. Deneysel sonuçlarda hedef sınıfın pozitif kabul edildiği özgüllük, duyarlılık metrikleri ve bunların harmonik ortalaması olan F1 değerinin deneylerin çoğunda daha başarılı olduğu görülürken negatif sınıfında göz önüne alındığı doğruluk metriği için öyle olmadığı görülmektedir. Sonuç olarak yapılan değerlendirmelerde doğruluk metriğinin patojen-konak etkileşim tahmininin başarısını yorumlamak için sağlıklı bir yöntem olmadığı, bu metrik yerine hedef sınıfın pozitif olduğu özgüllük ve duyarlılık ile birlikte bu değerlerin harmonik ortalaması olan F1 metriği ile daha sağlıklı değerlendirilebileceği söylenebilir.

5.2.3. Çoklu veri seti ile yapılan tahmin değerlendirmesi

Literatürde PKE tahmini üzerine yapılan çalışmalara bakıldığında en büyük problemlerden birinin veri yetersizliği olduğuna önceki bölümlerde değinilmişti. Yapılan bazı çalışmalarda, farklı patojen-konak etkileşim ağının tahmin modeli başarısını nasıl etkilediği gözlemlenmek istenmiştir. Bu bölümde benzer amaçla daha önce ayrı ayrı değerlendirilen bacillus ve ebola veri setlerini birleştirerek yapılan testlerde, tahmin doğruluğunu nasıl etkilediği test edilmiştir. Bu amaçla birleştirilmiş veri seti önceki bölümde kullanılan üç kodlama yöntemi, yedi tahmin yöntemi ile çaprazlanarak klasik PKE ve GAM yöntemleri ile etkileşim tahminleri yapılmış ve bulunan sayısal sonuçlar Tablo 5.4.'te verilmiştir.

Önceki bölümde her bir türe ait etkileşimin ayrı değerlendirildiği sonuçlar ile birleştirilmiş veri setleri üzerine yapılan testler kıyaslanarak veri seti birleştirmenin

tahmin doğruluğunu nasıl etkilediği yorumlanabilir. Örneğin önceki bölümde bacillus veri setinde en iyi tahminin elde edildiği K* sınıflandırıcısı ve AAP kodlama sonuçlarını göz önüne aldığımızda birleştirilmiş verilerde F1 değeri 0.506 olurken bu değer bacillus için 0.47, ebola için 0.547 olarak bulunmuştu. Dolayısıyla veri setlerinin birleştirilmesi bacillus için yapılan tahmini arttırmışken, ebola için yapılan tahminde düşüşe sebep olmuştur.

Benzer şekilde ebola için en iyi sonucun alındığı RF yönteminde F1 değeri 0.415 olarak bulunmuştur. Bu değer bacillus verisinde 0.1650, ebola veri setinde 0.756 olarak bulunmuştu. Sonuç olarak birleştirilmiş veriler üzerinden yapılan tahmin bir veri setinde düşerken diğerinde yükselmiştir. Birleştirilmiş verilerde, ayrı yapılan değerlendirmelerle benzer şekilde yüksek özgüllük, düşük duyarlılık değerleri elde edilmiştir. F1 sonuçlarına bakıldığında en iyi sonucun bacillus veri setine benzer şekilde AAP protein kodlaması ve K* sınıflandırıcısı ile alındığı görülmüştür. GAM veri seti ile yapılan tahminlerde en iyi F1 sonucu AAP kodlama ile kullanılan kNN sınıflandırıcısında gözlenmiştir.

Birleştirilmiş veri setleri üzerinde klasik PKE tahmin metotları ve GAM yöntemi ile bulunan sonuçlar kıyaslandığında tüm sınıflandırıcı ve kodlama yöntemlerinde F1 değerinin arttığı görülmektedir. Doğruluk sonuçlarına bakıldığında GAM ile bulunan sonuçların diğer yönteme yakın sonuçlar verdiği görülmektedir.

Tablo 5.4.'te verilen sonuçlar önceki testler ile kıyaslanıp genel olarak yorumlandığında bir türe ait etkileşimlerin ayrı ayrı değerlendirilmesi ve bir başka tür ile birleştirilmesi sonucu yapılan öğrenmenin etkileşim tahminine önemli bir katkı sağlamadığı görülmektedir. Tahmin doğruluğunun önemli bir artışın olmamasında farklı sebepler aranabilir. Bunlardan ilki farklı patojen proteinlerine ait aminoasit dizilerinin farklı dizilim karakteristiğine (dolayısıyla dizilimden çıkarılan özniteliklerin farklılaşması) sahip olması düşünülmektedir. Bir diğer sebep türlere ait bilinen etkileşim sayılarının farklı olmasından dolayı öğrenme modelinin etkileşim sayısı fazla olan tür lehine öğrenmenin gerçekleşmiş olma olasılığıdır. Yapılan

deneyde bacillus veri setinin eboladan fazla olması ve tahmin doğruluğunun da bacillus verisinde daha yüksek olması bu ihtimali güçlendirmektedir.

Bu bölümde birleştirilen etkileşim verilerinde organizmaların filogenetik yakınlığı göz ardı edilmiştir. Ayrıca yukarıda da değinildiği gibi veri setlerindeki örnek sayıları da bir birbirinden farklıdır. Veri yetersizliğinin olduğu organizmalar arası etkileşim tahminlerinde doğruluğu arttırmak için özellikle filogenetik yakınlığın olduğu organizmalara ait etkileşim verilerinin birleştirilmesinde fayda olacağı düşünülmektedir. Ayrıca birleştirilen veri setlerine ait öznitelik vektörleri oluşturulurken organizmalar arası durumları göz önüne alarak, proteinlere ait bir ağırlık katsayısı kullanmak da mümkündür.

Tablo 5.4. Birleştirilmiş veri seti için değerlendirme sonuçları

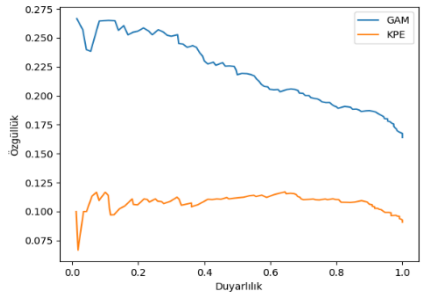
		PKE				Genişletilmiş Ağ Modeli			
Öznitelik	Metot	F1	Özg.	Dyr.	Dğr.	F1	Özg.	Dyr.	Dğr.
	BN	0.3970	0.3610	0.4430	87.81	0.6200	0.5270	0.7520	84.82
	NB	0.3670	0.2520	0.6770	78.78	0.5150	0.3900	0.7560	76.54
	RF	0.4150	0.9090	0.2699	93.12	0.6840	0.9410	0.5370	91.82
AAC	C4.5	0.4260	0.4340	0.4180	89.76	0.6230	0.6290	0.6170	87.70
	kNN	0.4360	0.3880	0.4990	88.30	0.6180	0.5570	0.6950	85.86
	K*	0.4130	0.3020	0.6530	83.15	0.5980	0.4700	0.8220	81.81
	MF	0.1978	0.1165	0.7275	45.80	0.3137	0.1955	0.8217	40.16
	BN	0.3740	0.2580	0.6800	79.33	0.4910	0.3980	0.6380	77.97
	NB	0.3150	0.2860	0.3500	86.17	0.3490	.0.3720	0.3280	79.64
	RF	0.4080	0.9000	0.2640	93.05	0.6990	0.9440	0.5550	92.06
AAP	C4.5	0.4310	0.4490	0.4160	90.05	0.6450	0.6500	0.6410	88.29
	kNN	0.4900	0.7000	0.3770	92.87	0.6950	0.8210	0.6030	91.22
	K*	0.5060	0.6270	0.4240	89.96	0.5940	0.7250	0.5040	84.83
	MF	0.2291	0.1966	0.4563	68.77	0.3584	0.3173	0.5021	68.29
	BN	0.3910	0.3490	0.4450	87.43	0.5900	0.4810	0.7610	82.41
	NB	0.2520	0.1550	0.6770	63.55	0.4090	0.2800	0.7600	63.56
	RF	0.3940	0.7520	0.2670	92.54	0.7210	0.9300	0.5880	92.42
CTDC	C4.5	0.6550	0.6290	0.6820	93.47	0.8020	0.8050	0.7980	88.15
	kNN	0.3400	0.3780	0.3080	89.12	0.6440	0.6280	0.6620	87.86
	K*	0.4090	0.3390	0.5160	87.86	0.6460	0.5630	0.7570	86.22
	MF	0.1850	0.1076	0.8548	29.25	0.3232	0.2148	0.8069	42.54

5.2.4. Özgüllük-duyarlılık grafikleri

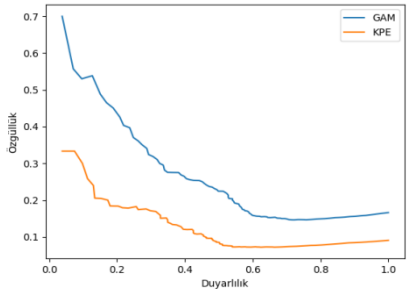
Bu deneysel çalışmanın son değerlendirmesi, matris faktorizasyon yöntemi ile klasik PKE ve GAM tahminlerine ait özgüllük-duyarlılık grafikleri yorumlanarak yapıldı. Matris faktorizasyon yönteminde, iki protein arasındaki etkileşime Denklem 3.1'den elde edilen sayısal sonuca bakarak karar verilir. Etkileşim kararının verilmesinde model, bulunan sayısal sonucun hangi sınıfa ait olduğu kararını vermek için bir eşik

değere ihtiyaç duyar. Sınıflandırma yapmadan önce belirlenen eşik değer modelin başarısını doğrudan etkilemektedir. Bulunan metriklerin başarısı verilen eşik değere göre değişmektedir. Yapılan tahminlerde eşik değer özgüllük-duyarlılık değerlerinin harmonik ortalaması olan F1 skorun maksimum olduğu yer olarak belirlendi. Böylece en iyi F1 performansının sağlandığı eşik değer seçilerek modelin başarısı maksimize edildi.

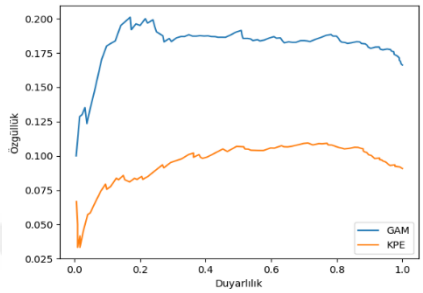
Özgüllük-duyarlılık grafikleri çizilirken, x_i ve y_i öznitelik vektörleri arası etkileşimlerin sayısallaştırıldığı $f(x_i,y_i)$ fonksiyonunda (denklem 3.1) tüm örneklerin pozitif ve negatif olduğu eşik değerler sırasıyla maksimum ve minimum olarak seçildi. Bulunan [max,min] değer aralığı 100 eşit parçaya bölündü. Bu aralıktaki her bir eşik değer için özgüllük ve duyarlılık değerleri hesaplanıp özgüllük-duyarlılık grafikleri çizildi. Şekil 5.3.'te her bir veri seti ve kodlama yöntemi için bulunan grafikler klasik PKE ve GAM yöntemleri için verilmiştir. Bu grafiklere bakıldığında GAM yönteminin dokuz deneyin sekizinde daha iyi sonuç verdiği görülmektedir. Grafiklere bakıldığında GAM yönteminin tüm duyarlılık değerleri için daha yüksek özgüllük değerine sahip olduğu görülmektedir. Her bir eşik değer için duyarlılık değerinin artması ile özgüllük değerinin azalma eğiliminde olduğu görülmektedir ancak önerilen yöntem büyük oranda klasik PKE tahmininden daha iyi sonuç vermektedir.



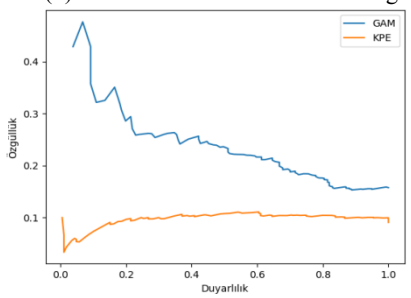
(a) Bacillus veri seti AAC kodlama grafiği



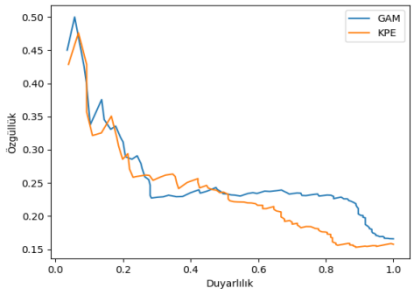
(b) Bacillus veri seti AAP kodlama grafiği



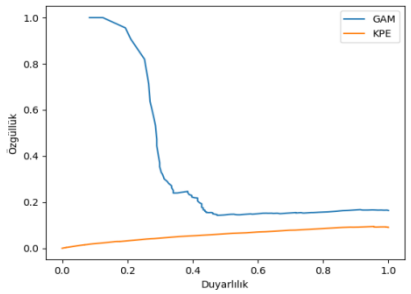
(c) Bacillus veri seti CTDC kodlama grafiği



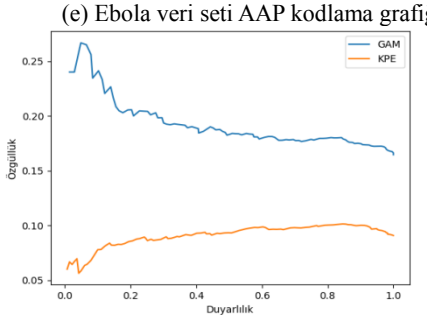
(d) Ebola veri seti AAC kodlama grafiği



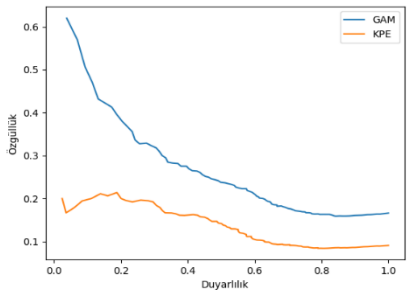
(e) Ebola veri seti AAP kodlama grafiği



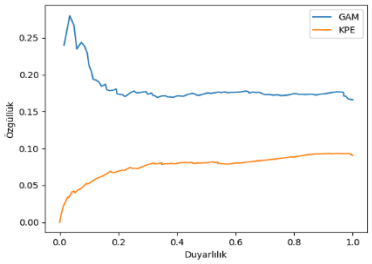
(f) Ebola veri seti CTDC kodlama grafiği



(g) Birleştirilmiş veri seti AAC kodlama grafiği



(h) Birleştirilmiş veri seti AAP kodlama grafiği



(i) Birleştirilmiş veri seti CTDC kodlama grafiği

Şekil 5.3.Matris faktorizasyon tahmin yöntemi için bulunan özgüllük-duyarlılık grafikleri

Yukarıda verilen grafiklere bakıldığında önceki bölümlerde, bacillus, ebola ve birleştirilmiş veri setleri için elde edilen sonuçlar ile benzerlik gösterdiği anlaşılmaktadır.

5.3. Deneysel Çalışma 2

Protein etkileşim tahmini için kullanılan makine öğrenmesi yöntemlerinde karşılaşılan en önemli problem veri yetersizliğidir. Son yıllarda proteinlere ait amino asit dizilim verilerinin hızla artması ile birlikte etkileşim tahminlerinde dizilim verilerine dayalı yöntemlere olan ilgi de artmıştır. Dizilim tabanlı tahmin modellerinin başarılı olmasındaki en önemli noktalardan biri verileri ayırt etmek için kullanılan öznitelik kodlama yöntemidir. Yeterli miktarda veri, verilerin ayırt edilmesinde kullanılacak güçlü kodlama yöntemi ve probleme uygun tahmin modeli etkileşim tahminindeki doğruluğu arttıracak etmenlerdir.

Bu bölümde konak-patojen protein etkileşim tahmininde doğruluğu arttırmak amacıyla tezde önerilen amino asit dizilimi tabanlı LTK yönteminin başarısı, literatürde sıkça kullanılan farklı öznitelik kodlama ve tahmin yöntemleri ile kıyaslanmıştır.

LTK, amino asitlerin protein dizilimindeki konumun, herhangi bir proteinin karakteristiğinin çıkarılmasında kullanılabileceği hipotezinden yola çıkarak önerilmiştir. Protein diziliminde geçen amino asitlerin her biri, zincir içindeki konumları dikkate alınarak, öznitelik vektöründe nümerik bir değerle kodlanmıştır. Konak ve patojen organizmaların proteinlerine ait dizilimlerden yola çıkarak bulunan öznitelik vektörleri birleştirilerek etkileşim tahmininde kullanılmıştır (detaylar için bakınız bölüm 4.2). Deneyler Bacillus Anthracis ve Yersinia Pestis veri setlerine uygulandı. Deneylerde: Naive Bayes, Bayesian Networks, Random Forest, C4.5 ve kNN tahmin metotları kullanıldı (detaylar için bakınız bölüm 3.2). Bu tahmin yöntemlerinden C4.5ve Random Forest karar ağaçları; Naive Bayes ve Bayesian Networks istatistiksel; kNN ise örnek tabanlı sınıflandırıcılardır. Tahmin yöntemlerinin tümü, bu çalışmada önerilen LTK yönteminin yanı sıra protein dizilim

tabanlı AAC, AAP ve CT protein kodlama yöntemleri (detaylar için bakınız bölüm 3.3) ile test edildi. Farklı kodlama ve tahmin yöntemi kombinasyonları ile yapılan deneylerde elde edilen sonuçların önemli bölümünde önerilen yöntemin tahmin doğruluğunu arttırdığı görüldü.

Sonuçlar değerlendirilirken daha önemli olduğu düşünülerek yorumlar genellikle F1 metriği üzerinden yapıldı. PKE tahmin probleminde asıl amaç etkileşimin olduğu protein çiftlerini tahmin etmek olduğundan dolayı özgüllük ve duyarlılık metriklerinin harmonik ortalaması olan F1 metriği tahmin sonuçlarının değerlendirilmesinde kullanılacak en önemli metriktir. Bu metrik sayesinde modelin, yanlış sınıflandırılmış pozitif ve negatiflerin pozitif örneklere oranı hakkında fikir edinilecektir. Her ne kadar PKE probleminde pozitif örneklerin doğru tahmini önemli olsa da aralarında etkileşim olmayan protein çiftlerinin bilinmesi de önemlidir. Bu sebeple verilen diğer metrikler ile kodlama ve tahmin metotlarının her iki sınıfın tahmini ile ilgili başarısı da tablolarda verilmiş ve dikkat çeken sonuçlar yorumlanmıştır.

5.3.1. Veri seti

Çalışmada deneysel yöntemler, konak canlının insan, patojenlerin bacillus anthracis ve yersinia pestis olduğu etkileşim verileri üzerinde test edilmiştir. Yapılan testlerin tümünde makine öğrenmesi temelli tahmin yöntemleri kullanılmıştır. Tahmin yöntemlerinin tümünde negatif ve pozitif (etkileşim olan ve olmayan) etiketli verilere göre öğrenme modeli çıkarılmıştır. Deneysel testlerde her iki tür için de kullanılan pozitif olarak etiketli etkileşim verileri PHISTO veritabanından indirilmiştir.

PKE ve PPE tahmin probleminde danışmalı öğrenme tabanlı tahmin metotlarında yapılan literatür çalışmalarının çoğunda, negatif veri seti konak ve patojen proteinler arasından rastgele seçilip eşlenmektedir. Konak ve patojen proteinler arası pozitif etkileşimler olası tüm eşlemelerin çok küçük oranını kapsadığından dolayı negatif sınıf içinde hata payının küçük olduğu düşünülmektedir. Eğitim verisi oluşturulurken bu oran dikkate alınarak genellikle negatif veri seti, pozitif veri setinden daha büyük

olacak şekilde seçilir. Bu deneyde benzer bir yol izlenerek pozitif verilerin negatif verilere oranı 1/5 olarak seçildi. Rastgele oluşturulan veri seti içinde pozitif verilerin olması ihtimaline karşı veri setleri karşılaştırılarak benzerliğin tespit edildiği protein çiftleri negatif veri setinden çıkarıldı.

Tablo 5.5. LTK yönteminin değerlendirilmesinde kullanılan veri setlerine ait bilgiler

	Bacillus Anthracis	**Yersinia Pestis**
Ağdaki pozitif etkileşim sayısı	3050	4097
Ağdaki negatif etkileşim sayısı	15250	20485
Ağdaki patojen protein sayısı	936	1227
Ağdaki konak protein sayısı	1686	2150
Toplam protein sayısı	5493	3909

Tablo 5.5.'te, kullanılan veri setlerine ait sayısal bilgiler verilmiştir. Tabloda kullanılan etkileşim ağı içindeki farklı patojen proteinlerin sayısı ve o türe ait bilinen tüm proteinlerin sayısı verilmiştir. Örneğin bacillus patojenine ait etkileşim ağında 1686 farklı protein olduğu görülmektedir. Bacillus patojenine ait bilinen tüm protein sayısının 5493 olduğu ve etkileşim ağında tüm proteinlerin yaklaşık %25'inin yer aldığı görülmektedir. Bu orana yersinia pestis patojeni için baktığımızda %55 civarında (2150/3909) olduğu görülmektedir. Çalışmada bilinen etkileşimler hariç tutulup olası tüm etkileşimler arasından bacillus için 15250, yersinia pestis için 20485 rastgele protein eşlemesi (pozitif verinin beş katı) negatif olarak etiketlenip veri seti oluşturulmuştur.

5.3.2. Bacillus anthracis veri setine ait sonuçlar

Tablo 5.6.'de bacillus anthracis için yapılan 20 farklı deneye ait değerlendirme sonuçları görülmektedir. Yöntemlerin kalite değerlendirmesinde özgüllük, duyarlılık, F1, doğruluk, MCC ve AUC metrikleri kullanılmıştır. Önerilen yöntemin dizilim tabanlı yöntemlerle kıyaslamanın yanında farklı tahmin metotları ile yapılan karşılaştırmanın nasıl değiştiği de değerlendirilmiştir.

Tablo 5.6.'da tüm kodlama ve tahmin yöntemi kombinasyonları arasında her bir kalite metriği için en iyi sonuçlar tabloda kalın karakterler ile gösterilmiştir. Bu sonuçlara bakıldığında F1, MCC, Doğruluk ve AUC metriklerine göre en iyi sonuçlar

LTK kodlama yöntemi ve RF tahmin modeli ile elde edilmiştir. Ayrıca duyarlılık metriğinde en iyi sonucun LTK kodlama ve BN tahmin modeliyle elde edildiği görülmektedir. Sadece özgüllük metriği için CT kodlama yönteminin RF tahmin metoduyla daha iyi olduğu görülmektedir.

Deney sonuçlarının tümüne bakıldığında LTK yöntemine en yakın sonuçların AAC yöntemiyle alındığı görülmektedir. LTK algoritmasının doğası gereği frekansı yüksek olan amino asitlerin öznitelik değeri yüksek çıkmaktadır. Dolayısıyla tahmin metotlarına göre yapılan sıralamada AAC ile benzerlik görülmektedir. Örneğin F1, Doğruluk, MCC ve AUC metriklerine göre her iki kodlama yönteminde de C4.5 tahmin metodu NB metodundan daha iyi sonuçlar üretmiştir. Bu metrikler için RF, kNN ve BN tahmin metotlarında, sıralama ufak farklarla değişmiştir.

LTK yöntemi AAP metodu ile kıyaslandığında NB tahmin metodu dışında diğer yöntemlerin tümü ile daha iyi sonuçlar vermiştir. Bu iki kodlama arasındaki en önemli fark kNN tahmin metodunda görülmektedir. Örneğin doğruluk metriğinde %10'luk bir artış vardır. Yine AUC metriği için 0.102'lik bir artış olduğu görülmektedir. Bu sonuçlara göre amino asit diziliminde kimyasal özelliklere göre yapılan bir gruplamadan elde edilen frekans değerinin etkileşim tahmininde lokasyon bilgisinden daha az etkili olduğu söylenebilir. Diğer bir deyişle amino asitleri dizilim içinde ayrı ayrı değerlendirmenin, kimyasal gruplar halinde değerlendirmeden daha fazla ayırt edici özellik verdiği sonucunu çıkarabiliriz.

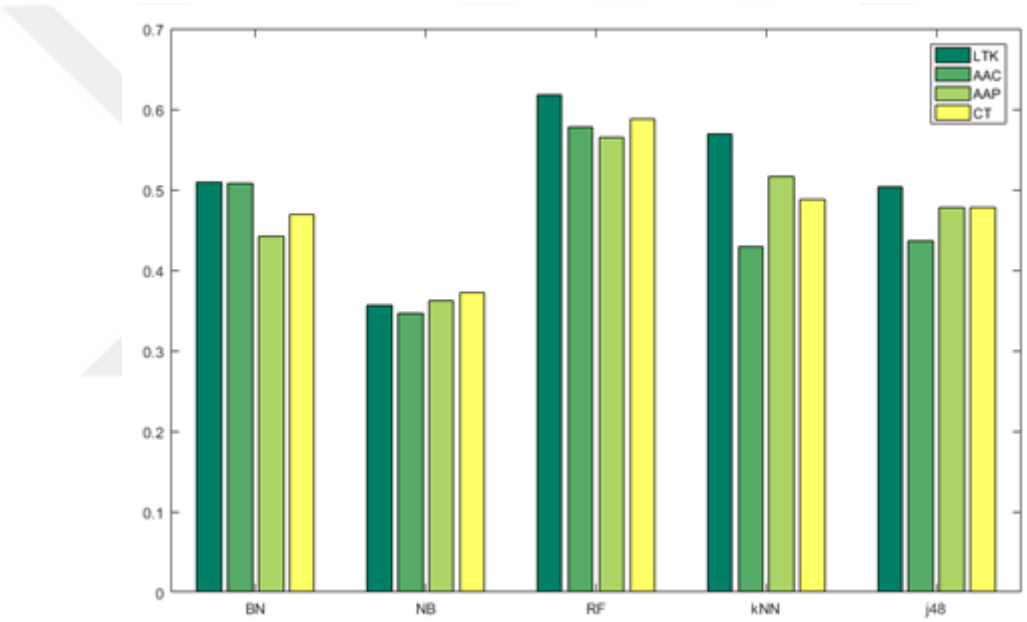
Tablo 5.6. Bacillus anthracis veri seti için bulunan değerlendirme sonuçları

		Özgüllük	Duyarlılık	F1	Doğruluk	MCC	AUC
LTK	BN	0,477	**0,807**	0,6	0,803	0,51	0,862
	NB	0,546	0,375	0,445	0,829	0,356	0,811
	RF	0,805	0,565	**0,664**	**0,895**	**0,618**	**0,931**
	kNN	0,688	0,602	0,642	0,877	0,57	0,857
	C4.5	0,586	0,607	0,596	0,85	0,504	0,725
AAC	BN	0,52	0,72	0,604	0,627	0,508	0,873
	NB	0,351	0,747	0,478	0,701	0,347	0,79
	RF	0,891	0,442	0,59	0,888	0,578	0,922
	kNN	0,432	0,728	0,542	0,775	0,43	0,755
	C4.5	0,548	0,527	0,537	0,834	0,436	0,724
AAP	BN	0,455	0,704	0,553	0,792	0,442	0,84
	NB	0,54	0,393	0,455	0,828	0,362	0,826
	RF	0,828	0,465	0,596	0,884	0,565	0,922
	kNN	0,667	0,529	0,59	0,866	0,516	0,737
	C4.5	0,569	0,579	0,574	0,843	0,478	0,708

Tablo 5.6. (Devamı)

	BN	0,45	0,778	0,571	0,786	0,47	0,851
	NB	0,354	0,802	0,491	0,696	0,372	0,795
CT	RF	**0,894**	0,453	0,601	0,89	0,588	0,927
	kNN	0,468	0,778	0,585	0,798	0,488	0,848
	C4.5	0,564	0,588	0,575	0,841	0,478	0,729

Son olarak LTK yöntemi CT yöntemi ile F1 metriği için kıyaslandığında 0.05 daha düşük sonuç ürettiği görülmektedir. F1 metriğinde LTK geri kalan tüm tahmin metotlarında daha iyi sonuçlar vermiştir. CT kodlaması ile yapılan tahminlerde dikkat çeken sonuçlardan biri BN yönteminin NB haricindeki tüm tahmin metotlarının gerisinde kalmasıdır.



Şekil 5.4. Bacillus veri seti için bulunan MCC sonuçlarının kodlama ve tahmin yöntemlerine göre kıyaslanması

Şekil 5.4.'te Bacillus veri seti için MCC metriğine göre kodlama ve tahmin metotlarına göre yapılan PKE sonuçları verilmiştir. Yukarıda tablodaki sayısal sonuçlara göre yapılan yorumların, tek bir metrik için görsel sonuçlarına bakıldığında LTK yönteminin NB tahmin metodu dışında tüm yöntemlerde en iyi kodlama yöntemi olduğu görülmektedir. Grafiğe göre tüm deneyler için sonuçlar göz önüne alındığında AAP-kNN eşlemesi dışında en iyi tahmin metodunun LTK-RF kombinasyonu ile elde edildiği görülmektedir.

Tablo 5.7.'de dört kalite analiz metriğine göre LTK yönteminin diğer kodlama yöntemleri ile yapılan kıyaslamada, daha iyi olduğu deney sayıları verilmiştir. Tablodaki sonuçlara bakıldığında LTK yönteminin, tahmin metotları ve sonuçlar arasındaki fark göz ardı edilerek sadece deney sayılarına göre yapılan kıyaslamada LTK yönteminin her üç kodlama yönteminden düzenli olarak daha iyi sonuçlar verdiği görülmüştür. Yine benzer yaklaşımla yapılan değerlendirmede AAC yönteminin AAP ve CT yöntemlerinin gerisinde kaldığı, AAP ve CT yöntemlerinin ise eşit sayıda deneyde başarılı olduğu görülmektedir.

Tablo 5.7. Bacillus veri seti için LTK ile diğer kodlama yöntemlerinin deney sayısı üstünlüğüne göre kıyaslanması

	AAC	AAP	CT	Toplam Başarı
Doğruluk	3	4	4	11/15
F1	5	5	5	15/15
MCC	5	4	4	13/15
AUC	5	4	4	13/15

5.3.3. Yersinia pestis veri setine ait sonuçlar

LTK yönteminin uygulandığı ikinci veri seti yersinia pestis organizmasıdır. Önerilen yöntem, farklı kodlama yöntemleri ile karşılaştırmanın yanında farklı organizmaya ait veri setine benzer deneyleri uygulayarak sonuçlar arasında kıyaslama yapıldı. Tablo 5.8.'de Yersinia pestis organizmasına ait değerlendirme sonuçları görülmektedir.

Aşağıdaki tabloda verilen 20 deney sonucuna bakıldığında en iyi F1, doğruluk, MCC ve AUC sonucunun LTK-RF kodlama- sınıflandırıcı yöntem eşlemesinde alındığı görülmektedir. Bacillus veri setinde de en iyi sonucun LTK-RF eşlemesi ile alındığı düşünüldüğünde bu çalışmada önerilen LTK yönteminin RF tahmin metoduyla PKE tahminin de en iyi sonuçları verdiği daha net görülmüş oldu. Diğer veri seti ile benzer olarak BN tahmin metodunda LTK ve AAC kodlama yöntemleri ile çok yakın sonuçlar alınmıştır. BN yöntemi ile AAP ve CT kodlamasında %2 seviyesinde yakın sonuçlar alınmıştır. Bu sonuçlar hem Yersinia hem bacillus veri setleri için geçerlidir.

AAC kodlaması, LTK kodlamasına göre BN tahmin metodunda F1 metriği için yakın sonuçlar alınmasına rağmen, doğruluk metriğinde %12 oranında bir düşüş olmuştur. Bu düşüş BN tahmin metodunun AAC kodlamasında negatif sınıf tahmininde daha kötü sonuçlar döndürdüğünü göstermektedir. Bu durum AAP ve CT kodlaması için geçerli olmayıp yöntemler doğruluk metriğinde de beklenen sonuçları vermiştir.

AAP ve CT kodlama yöntemleri LTK ile kıyaslandığında doğruluk metriğinde %1-2 oranında yakın sonuçların alındığı görülmektedir. Ancak F1 skor değerinde fark %3-5 seviyesine ulaşmaktadır. Bu durum LTK kodlaması ile pozitif etiketli örneklerin AAC ve CT kodlamalarına göre daha iyi tahmin edildiğini göstermektedir.

kNN tahmin metodu en iyi F1 sonucunu LTK kodlaması ile vermiştir. LTK'dan sonra en yakın sonuç AAP kodlamasıyla elde edilmiştir. Daha sonra gelen CT ve AAC kodlama doğruluk sonuçları neredeyse aynıdır.

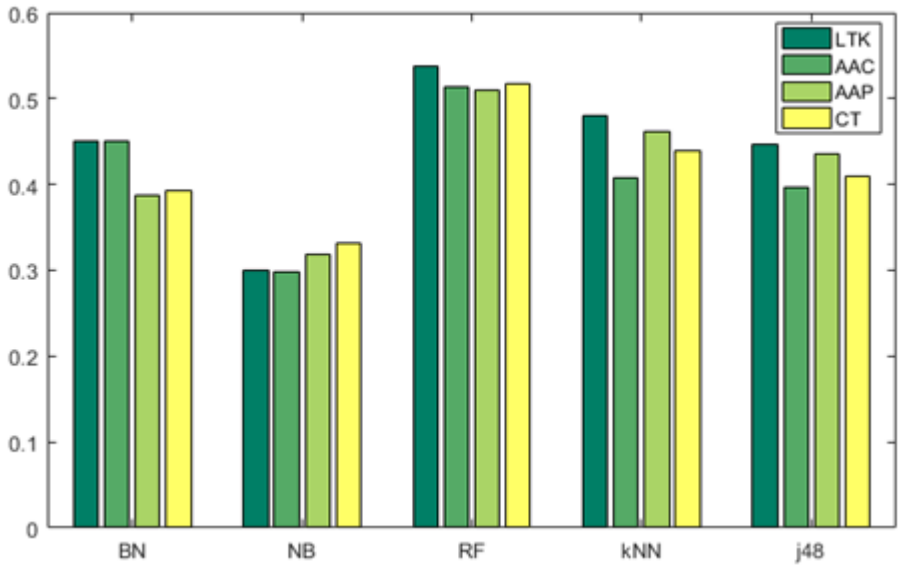
Yersinia pestis veri seti için elde edilen genel sonuçlara bakıldığında en kötü sonuçların NB tahmin metoduyla alındığı görülmüştür. Bunun yanında gene bayes tabanlı bir yöntem olan BN her üç kodlama yöntemi ile de yapılan tahminlerde diğer yöntemlere yakın sonuçlar döndürmüştür.

Şekil 5.6.'da yersinia veri seti için her bir kodlama ve tahmin yöntemi için bulunan MCC sonuçları grafikte görülmektedir. Bacillus veri seti ile benzer olarak en iyi sonuçlar RF-LTK yöntemleri ile elde edilmiştir. Gene diğer veri seti ile benzer şekilde LTK yöntemi NB tahmin metodunda tüm kodlama yöntemlerinde geride kalmakla birlikte AAC ile çok yakın sonuçlar vermiştir. LTK'nın diğer tüm tahmin yöntemlerinde en iyi sonucu verdiği grafikten anlaşılmaktadır.

LTK yöntemi AAC ile bayes tabanlı sınıflandırıcılarda (BN ve NB) yakın sonuçlar verirken, diğer sınıflandırıcılarda aradaki fark daha fazladır. Yersinia veri setine ait MCC grafiğine bakıldığında başarı sıralamasının bacillus veri seti ile çok yakın olduğu görülmektedir.

Tablo 5.8. Yersinia pestis veri seti için bulunan değerlendirme sonuçları

		Özgüllük	Duyarlılık	F1	Doğruluk	MCC	AUC
LTK	BN	0,432	**0,778**	0,555	0,772	0,451	0,833
	NB	0,443	0,402	0,421	0,798	0,3	0,775
	RF	0,793	0,451	**0,575**	**0,878**	**0,538**	**0,914**
	kNN	0,604	0,539	0,569	0,851	0,48	0,825
	C4.5	0,542	0,558	0,55	0,833	0,447	0,688
AAC	BN	0,485	0,659	0,559	0,809	0,45	0,842
	NB	0,329	0,692	0,446	0,685	0,298	0,754
	RF	**0,929**	0,335	0,492	0,873	0,513	0,903
	kNN	0,447	0,642	0,527	0,789	0,407	0,797
	C4.5	0,523	0,484	0,503	0,825	0,397	0,711
AAP	BN	0,415	0,668	0,512	0,767	0,388	0,806
	NB	0,459	0,419	0,438	0,803	0,319	0,791
	RF	0,839	0,379	0,522	0,873	0,509	0,903
	kNN	0,625	0,478	0,542	0,852	0,461	0,827
	C4.5	0,539	0,54	0,54	0,831	0,436	0,698
Conjoint Triad	BN	0,394	0,735	0,513	0,744	0,392	0,809
	NB	0,331	0,775	0,464	0,672	0,331	0,761
	RF	0,915	0,349	0,505	0,875	0,518	0,906
	kNN	0,441	0,728	0,55	0,781	0,439	0,823
	C4.5	0,512	0,528	0,52	0,821	0,41	0,689



Şekil 5.5. Yersinia veri seti için bulunan MCC sonuçlarının kodlama ve tahmin yöntemlerine göre kıyaslanması

Tablo 5.9.'da yersinia veri seti ile yapılan deneylerde LTK yönteminin diğer kodlama yöntemlerine göre daha başarılı olduğu deney sayıları verilmiştir. Bu tablo yersinia veri seti için verilen sayısal sonuçlara bakılarak çıkarılmış ve okuyucuların daha rahat bir değerlendirme yapması için hazırlanmıştır. Tablo 5.9.'da başarılı olunan deneylerin hangileri olduğu göz ardı edilerek sadece kodlama yöntemlerinde başarılı olunan toplam deney sayıları verilerek tablo oluşturulmuştur. Burada amaç önerilen

LTK kodlamasının deneylerin büyük çoğunluğunda diğer kodlama yöntemlerinden daha iyi olduğunu göstermektir.

Tablo 5.9. Yersinia veri seti için LTK kodlama ile diğer kodlama yöntemleri arasında yapılan kıyaslamada başarılı olunan deney sayısı

	AAC	AAP	CT	Toplam Başarı
Doğruluk	4	3	5	12/15
F1	3	4	4	11/15
MCC	5	4	4	13/15
AUC	3	2	4	9/15

BÖLÜM 6. TARTIŞMA VE SONUÇ

Organizmaların proteinleri arasındaki olası etkileşim sayısının çok yüksek boyutta olması proteinler arası etkileşimleri tespit etmede kullanılan deneysel yöntemleri yetersiz kılmaktadır. PKE tahmininde hesaplamalı yöntemler kullanarak deneysel yöntemlerin yüksek maliyet ve uzun zaman dezavantajları ortadan kaldırılmak istenmektedir. Hesaplamalı yöntemler kullanılarak yapılan çalışmalarda karşılaşılan en önemli problem veri yetersizliğidir. Veri yetersizliğini iki açıdan değerlendirebiliriz. İlki patojen ve konak organizmalar arasında deneysel olarak doğrulanmış etkileşimlerin olmaması veya bu sayının yetersiz olmasıdır. İkincisi proteinler arası etkileşimin bilinmesine rağmen kullanılan hesaplamalı yöntemin ihtiyaç duyduğu protein bilgisine erişememektir. Örneğin proteinlerin ikincil yapısından yola çıkarak geliştirilen bir tahmin yönteminde az sayıda proteinin ikincil yapısının bilinmesi yöntemin uygulanabilirliği açısından önemli bir problemdir.

Tezde, PKE tahmini üzerine yapılan çalışmalarda genel olarak tahmin doğruluğunu arttırmak ve veri yetersizliğinden kaynaklanan eksikleri gidermek amaçlandı. Veri yetersizliği göz önüne alınarak, çalışmada PKE tahmini için en fazla verinin olduğu hesaplamalı yöntemden yola çıkıldı. Veri tabanlarında proteinlere ait en fazla veri, amino asit dizilim bilgisidir. Bu sebeple sadece amino asit dizilim verisi kullanılarak etkileşim tahminini arttırmaya yönelik çalışmalar yapıldı. Bu kapsamda ilk olarak konak ve patojen organizmalara ait yeterli sayıda doğrulanmış etkileşim verisinin olmadığı durumlar için GAM yöntemi önerildi. GAM yönteminde temel hipotez türler arası etkileşim ağında yer alan proteinlerin, tür içi etkileşim kurdukları proteinlerle benzer karakteristiğe sahip olduğudur. Dolayısıyla türler arası etkileşimde yer alan proteinlere ait tür içi etkileşimlerin PKE ağı ile birleştirilmesi sonucu tahmin doğruluğunun artacağı düşünüldü. İki farklı veri seti ile yapılan deneysel çalışmaların büyük çoğunluğu hipotezi destekledi. Bu bölümde ayrıca farklı

türlere ait veri setlerinin birleştirilmesi sonucu tahmin doğruluğunun nasıl değişeceği gözlenmek istendi. Birleştirilen veri setlerinin tahmin doğruluğuna önemli bir katkı sağlamadığı sonucuna ulaşıldı. Yapılan deneylerde doğruluğa etkisi olabilecek etkenlerden biri, veri setlerinde türler arası tüm proteinler arasından rastgele oluşturulan negatif veri sınıfıdır. GAM yöntemine ait deneylerde, pozitif sınıfın negatif sınıfa oranı 1/10 olacak şekilde rastgele negatif örnek seçildi. Bu oranın artması ile birlikte tahmin yöntemleri, negatif sınıfa ait örnekleri daha iyi öğrenip, daha az orana sahip pozitif sınıfı tahminde zayıf kalmaktadır. GAM modeline göre artan pozitif sınıf sayesinde sınıflandırma algoritmalarının tahmin doğruluğu artmaktadır. Tahmin doğruluğunun artması tür içi ve türler arası ağda yer alan proteinlerin benzer karakteristiğe sahip olduğu tezinin doğruluğunu göstermektedir. Tahmin metotları arasında rastsal orman gibi veri setini, alt veri setlerine ayıran yöntemlerde de sadece PKE verileri ile yapılan tahminlerde pozitif sınıf tahmini çok kötü sonuçlar vermiştir. Bunun sebebi 1/10 pozitif negatif sınıf oranına sahip veri setinde alt veri setlerinin çoğunun sadece negatif örneklerden oluşmasıdır. Rastsal orman yöntemi alt veri setlerinin her biri için bir karar ağacı oluşturup daha sonra test örnekleri bu karar ağaçlarından gelen sonuçların oylanması ile sınıflandırılmaktadır. Büyük çoğunluğu negatif verilerden oluşan karar ağaçları, pozitif örnek tahmininde daha az oy alacak, dolayısıyla yanlış sınıflandırma yapacaktır. GAM yöntemi ile pozitif örnek sayısı arttırılarak bu problemin de önüne geçilmiştir. Rassal orman metodu ile yapılan deneylerde özgüllük, duyarlılık ve F1 sonuçları önemli ölçüde artış göstermiştir. Benzer durum naif bayes ve bayes ağlarında da geçerlidir. Örnek tabanlı sınıflandırıcıların bu duruma daha toleranslı olduğu söylenebilir. Deneylerde GAM yöntemine ait sonuçlar daha az artış göstermesine rağmen her iki deney için özgüllük ve duyarlılık sonuçlarına bakıldığında sonuçların daha dengeli değiştiği görülmektedir.

Makine öğrenmesi algoritmalarında amino asit dizilimi tabanlı patojen konak etkileşim tahmininde başarımı etkileyen faktörlerden biri de kullanılan özniteliklerdir. Tezde tahmin doğruluğunu arttırmak amacıyla etkileşimleri daha iyi ayırt edecek protein kodlaması üzerine çalışıldı. Tahmin metotlarına öznitelik vektörü olarak verilecek kodlanmış amino asit dizilerinin sabit uzunlukta olması

gerekmektedir. Literatürde geçen ve sabit uzunlukta öznitelik vektörü üreten kodlama yöntemleri genellikle amino asitlerin kimyasal özelliklerini veya dizi içerisindeki frekanslarını göz önüne almaktadır. Çalışmada amino asitlerin frekans ve kimyasal özellikleri dışında dizi içerisindeki lokasyonlarının da proteinleri ayırt etmede kullanılabileceği düşünüldü. Dizide yer alan her bir amino asidin indis değeri dizi uzunluğuna bölünüp benzer amino asitlerin değerleri ile toplandı. Bu işlem sonrası yirmi amino asidin her birine karşılık bir nümerik bir değer elde edildi. Proteinler arası farklılığı arttırmak için amino asit dizi uzunluğu önceden belirlenen bir parametreye göre alt diziler ayrıldı. Bu alt dizilerden sonuncusu amino asit dizisinin tümünden oluşmaktadır. Alt diziler oluşturulduktan sonra her bir alt dizi için nümerik değerler hesaplandı.

Dizinin sonlarında yer alan amino asitler daha büyük indise sahip olduklarından dolayı benzer amino asidin başlarda yer aldığı ve daha yüksek frekanslara sahip olduğu proteinler ile karışma ihtimali vardır. Bu durumun önüne geçmek için aynı işlem amino asit dizisinin tersine de uygulandı. Alt dizi ve dizilerin tersinden oluşan nümerik vektörler birleştirilerek proteine ait öznitelik vektörü oluşturuldu. Alt dizilere ayırma ve dizinin tersine çevrilmesi ile farklı proteinlerden benzer nümerik vektör oluşma ihtimali önemli ölçüde düşürülmüştür.

LTK algoritmasının alt dizi sayısını belirleyen parametre, doğrudan öznitelik vektör uzunluğunu belirlemektedir. Amino asit dizi uzunluğu kısa olan proteinler göz önüne alınarak alt dizi parametresinin çok büyük seçilmemesi gerekmektedir. Ayrıca parametrenin büyük olması öznitelik vektörünün de büyük olmasına, dolayısıyla hesaplama maliyetinin artmasına sebep olacaktır. LTK yönteminde seçilmesi gereken kesin bir alt dizi sayısı yoktur. Optimum alt dizi sayısı kullanılan veri seti üzerinde farklı deneylere göre belirlenmelidir. LTK başarımının ölçüldüğü deneysel çalışmada alt dizi sayısı beş olarak belirlendi.

Önerilen öznitelik kodlama yöntemi ve genişletilmiş ağ modeli yapılan deneylerin büyük çoğunluğunda daha iyi sonuçlar verdi.

KAYNAKLAR

[1] S. Mei, "Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins," *PLoS One*, vol. 8, no. 11, p. e79606, 2013.

[2] S. De Bodt, S. Proost, K. Vandepoele, P. Rouzé, and Y. de Peer, "Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression," *BMC Genomics*, vol. 10, no. 1, p. 288, 2009.

[3] J. Shen, J. Zhang, X. Luo, and W. Zhu, "Predicting protein–protein interactions based only on sequences information," *Proc. ...*, vol. 104, no. 11, pp. 4337–4341, 2007.

[4] A. W. Crosby, *The Columbian Exchange: Biological and Cultural Consequences of 1492*. 1972.

[5] R. Acuna-Soto, D. W. Stahle, M. K. Cleaveland, and M. D. Therrell, "Megadrought and megadeath in 16th century Mexico," *Emerging Infectious Diseases*, vol. 8, no. 4. pp. 360–362, 2002.

[6] J. Diamond, "Guns, germs and steal : the fates of human societies," *Perspect. Biol. Med.*, p. 512, 1999.

[7] M. Naghavi *et al.*, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013," *Lancet*, vol. 385, no. 9963, pp. 117–171, 2015.

[8] M. Kshirsagar, J. Carbonell, and J. Klein-Seetharaman, "Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks," *NIPS Work. Mach. Learn. Comput. Biol.*, vol. 1, no. 1, pp. 3–6, 2013.

[9] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, "Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins," in *Bioinformatics*, 2010, vol. 27, no. 13, pp. i645–i652.

[10] H. Kitano, "Systems biology: a brief overview.," *Science*, vol. 295, no. 5560, pp. 1662–4, 2002.

[11] S. Durmus, T. Çakir, A. Özgür, and R. Guthke, “A review on computational systems biology of pathogen-host interactions,” *Frontiers in Microbiology*, vol. 6, no. APR. 2015.

[12] E. Nourani, F. Khunjush, S. Durmu\cs, and S. Durmus, “Computational prediction of virus-human protein-protein interactions using embedding kernelized heterogeneous data,” *Mol. Biosyst.*, vol. 12, no. 6, pp. 1976–1986, 2016.

[13] Z.-H. You, K. C. C. Chan, and P. Hu, “Predicting Protein-Protein Interactions from Primary Protein Sequences Using a Novel Multi-Scale Local Feature Representation Scheme and the Random Forest,” *PLoS One*, vol. 10, no. 5, p. e0125811, May 2015.

[14] E. Nourani, F. Khunjush, S. Durmu\cs, and S. Durmus, “Computational approaches for prediction of pathogen-host protein-protein interactions,” *Front. Microbiol.*, vol. 6, no. FEB, p. 94, 2015.

[15] B. A. Shoemaker and A. R. Panchenko, “Deciphering protein-protein interactions. Part I. Experimental techniques and databases,” *PLoS Computational Biology*, vol. 3, no. 3. pp. 0337–0344, 2007.

[16] M. Kshirsagar, J. Carbonell, and J. Klein-Seetharaman, “Multitask learning for host-pathogen protein interactions,” in *Bioinformatics*, 2013, vol. 29, no. 13, pp. i217--i226.

[17] L. Cai, Z. Pei, S. Qin, and X. Zhao, “Prediction of protein-protein interactions in saccharomyces cerevisiae based on protein secondary structure,” in *Biomedical Engineering and Biotechnology (iCBEB), 2012 International Conference on*, 2012, pp. 413–416.

[18] T. Guirimand, S. Delmotte, and V. Navratil, “VirHostNet 2.0: surfing on the web of virus/host molecular interactions data,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D583--D587, 2015.

[19] R. Winnenburg, T. K. Baldwin, M. Urban, C. Rawlings, J. Köhler, and K. E. Hammond-Kosack, “PHI-base: a new database for pathogen host interactions,” *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D459--D464, 2006.

[20] Z. Xiang, Y. Tian, and Y. He, “PHIDIAS: a pathogen-host interaction data integration and analysis system,” *Genome Biol.*, vol. 8, no. 7, p. R150, 2007.

[21] R. Kumar and B. Nanduri, “HPIDB-a unified resource for host-pathogen interactions,” in *BMC bioinformatics*, 2010, vol. 11, no. S6, p. S16.

[22] D. Szklarczyk *et al.*, “The STRING database in 2017: quality-controlled protein--protein association networks, made broadly accessible,” *Nucleic Acids Res.*, p. gkw937, 2016.

[23] H. Zhou, J. Jin, and L. Wong, “Progress in computational studies of host-pathogen interactions.,” *J. Bioinform. Comput. Biol.*, vol. 11, no. 2, p. 1230001, 2013.

[24] P. Uetz *et al.*, “A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.,” *Nature*, vol. 403, no. 6770, pp. 623–7, 2000.

[25] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, “A comprehensive two-hybrid analysis to explore the yeast protein interactome,” *Proc. Natl. Acad. Sci.*, vol. 98, no. 8, pp. 4569–4574, 2001.

[26] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin, “A generic protein purification method for protein complex characterization and proteome exploration,” *Nat. Biotechnol.*, vol. 17, no. 10, pp. 1030–1032, 1999.

[27] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proc Natl Acad Sci USA*, vol. 95, no. 25, pp. 14863–14868, 1998.

[28] A. . b K. J. A. . M. G. . Jones R.B.a Gordus, “A quantitative protein interaction network for the ErbB receptors using protein microarrays,” *Nature*, vol. 439, no. 7073, pp. 168–174, 2006.

[29] G. MacBeath and S. L. Schreiber, “Printing proteins as microarrays for high-throughput function determination,” *Science (80-.).*, vol. 289, no. September, pp. 1760–1763, 2000.

[30] H. Berman *et al.*, “The Protein Data Bank and the challenge of structural genomics,” *Nat. Struct. Biol.*, vol. 7 Suppl, pp. 957–959, 2000.

[31] P. Ye, B. D. Peyser, X. Pan, J. D. Boeke, F. A. Spencer, and J. S. Bader, “Gene function prediction from congruent synthetic lethal interactions in yeast.,” *Mol. Syst. Biol.*, vol. 1, no. 1, p. 2005.0026, Jan. 2005.

[32] G. P. Smith, “Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface.,” *Science*, vol. 228, no. 4705, pp. 1315–7, Jun. 1985.

[33] A. H. Y. Tong *et al.*, “Systematic genetic analysis with ordered arrays of yeast deletion mutants,” *Science (80-.).*, vol. 294, no. 5550, pp. 2364–2368, 2001.

[34] Y. Yan and G. Marriott, “Analysis of protein interactions using fluorescence technologies,” *Current Opinion in Chemical Biology*, vol. 7, no. 5. pp. 635–640, 2003.

[35] M. A. Cooper, “Label-free screening of bio-molecular interactions,” *Analytical and Bioanalytical Chemistry*, vol. 377, no. 5. pp. 834–842, 2003.

[36] Y. Yang, “Quantitative characterization of biomolecular assemblies and interactions using atomic force microscopy,” *Methods*, vol. 29, no. 2, pp. 175–187, 2003.

[37] W. Baumeister, R. Grimm, and J. Walz, “Electron tomography of molecules and cells,” *Trends in Cell Biology*, vol. 9, no. 2. pp. 81–85, 1999.

[38] J.-T. T. Yu and M.-Z. Z. Guo, “Prediction of Protein-Protein Interactions from Secondary Structures in Binding Motifs Using the Statistic Method,” in *2008 Fourth International Conference on Natural Computation*, 2008, vol. 5, pp. 100–103.

[39] X. Zhao, J. Li, Y. Huang, Z. Ma, and M. Yin, “Prediction of bioluminescent proteins using auto covariance transformation of evolutional profiles,” *Int. J. Mol. Sci.*, vol. 13, no. 3, pp. 3650–3660, 2012.

[40] N. Liu and T. Wang, “Protein-based phylogenetic analysis by using hydropathy profile of amino acids,” *FEBS Lett.*, vol. 580, no. 22, pp. 5321–5327, 2006.

[41] M. D. Dyer, T. M. Murali, and B. W. Sobral, “Computational prediction of host-pathogen protein-protein interactions,” in *Bioinformatics*, 2007, vol. 23, no. 13.

[42] M. D. Dyer, T. M. M. Murali, and B. W. Sobral, “Supervised learning and prediction of physical interactions between human and HIV proteins,” *Infect. Genet. Evol.*, vol. 11, no. 5, pp. 917–923, 2011.

[43] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman, “Prediction of interactions between HIV-1 and human proteins by information integration.,” *Pac. Symp. Biocomput.*, pp. 516–27, 2009.

[44] I. Nouretdinov, A. Gammerman, Y. Qi, and J. Klein-Seetharaman, “Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method,” *Pac. Symp. Biocomput.*, pp. 311–322, 2012.

[45] P. Baldi and S. S. Brunak, *Bioinformatics: the machine learning approach*. MIT press, 2001.

[46] J. R. Bock and D. A. Gough, “Predicting protein--protein interactions from primary structure.,” *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[47] K. C. Mondal, N. Pasquier, A. Mukhopadhyay, C. da Costa Pereira, U. Maulik, and A. Tettamanzi, “Prediction of Protein Interactions on HIV-1-Human PPI Data using a Novel Closure-based Integrated Approach.,” in *BIOINFORMATICS*, 2012, pp. 164–173.

[48] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and R. Eils, “Mining association rules from HIV-human protein interactions,” in *Proc. Int. Conf. Systems in Medicine and Biology*, 2010, pp. 344–348.

[49] A. Mukhopadhyay and U. Maulik, “Network-based study reveals potential infection pathways of hepatitis-c leading to various diseases,” *PLoS One*, vol. 9, no. 4, 2014.

[50] S. Ray, A. Mukhopadhyay, and U. Maulik, “Predicting annotated HIV-1-Human PPIs using a biclustering approach to association rule mining,” in *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*, 2012, pp. 28–31.

[51] S. Mei, “Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins,” *PLoS One*, vol. 8, no. 11, 2013.

[52] M. Kshirsagar, J. G. Carbonell, J. Klein-Seetharaman, and K. Murugesan, “Multitask Matrix Completion for Learning Protein Interactions Across Diseases,” in *Research in Computational Molecular Biology: 20th Annual Conference, RECOMB 2016, Santa Monica, CA, USA, April 17-21, 2016, Proceedings*, M. Singh, Ed. Cham: Springer International Publishing, 2016, pp. 53–64.

[53] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman, “Prediction of interactions between HIV-1 and human proteins by information integration,” in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2009, p. 516.

[54] M. D. Dyer, T. M. Murali, and B. W. Sobral, “Supervised learning and prediction of physical interactions between human and HIV proteins,” *Infect. Genet. Evol.*, vol. 11, no. 5, pp. 917–923, 2011.

[55] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, “Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins,” in *Bioinformatics*, 2010, vol. 27, no. 13, pp. i645–i652.

[56] Q. Xu, E. W. Xiang, and Q. Yang, “Protein-protein interaction prediction via collective matrix factorization,” in *Proceedings - 2010 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2010*, 2010, pp. 62–67.

[57] S. Orchard *et al.*, “Protein interaction data curation: The International Molecular Exchange (IMEx) consortium,” *Nature Methods*, vol. 9, no. 4. pp. 345–350, 2012.

[58] G. D. Bader, D. Betel, and C. W. V Hogue, “BIND: The Biomolecular Interaction Network Database,” *Nucleic Acids Res.*, vol. 31, no. 1, pp. 248–250, 2003.

[59] I. Xenarios, “DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 303–305, 2002.

[60] L. Licata *et al.*, “MINT, the molecular interaction database: 2012 Update,” *Nucleic Acids Res.*, vol. 40, no. D1, 2012.

[61] A. Chatr-Aryamontri *et al.*, “The BioGRID interaction database: 2017 update,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D369–D379, 2017.

[62] T. S. Keshava Prasad *et al.*, “Human Protein Reference Database--2009 update.,” *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D767-72, 2009.

[63] S. Kerrien *et al.*, “The IntAct molecular interaction database in 2012,” *Nucleic Acids Res.*, vol. 40, no. D1, 2012.

[64] R. A. Laskowski, V. V. Chistyakov, and J. M. Thornton, “PDBsum more: New summaries and analyses of the known 3D structures of proteins and nucleic acids,” *Nucleic Acids Res.*, vol. 33, no. DATABASE ISS., 2005.

[65] M. Rashid, S. Ramasamy, and G. P. S. Raghava, “A simple approach for predicting protein-protein interactions.,” *Current protein & peptide science*, vol. 11, no. 7. pp. 589–600, 2010.

[66] H. W. Mewes *et al.*, “MIPS: A database for genomes and protein sequences,” *Nucleic Acids Research*, vol. 27, no. 1. pp. 44–48, 1999.

[67] T. Beuming, L. Skrabanek, M. Y. Niv, P. Mukherjee, and H. Weinstein, “PDZBase: A protein-protein interaction database for PDZ-domains,” *Bioinformatics*, vol. 21, no. 6. pp. 827–828, 2005.

[68] S. Razick, G. Magklaras, and I. M. Donaldson, “iRefIndex: A consolidated protein interaction database with provenance,” *BMC Bioinformatics*, vol. 9, 2008.

[69] M. Kanehisa and S. Goto, “Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, pp. 27–30, 2000.

[70] S. Durmu\cs Tekir *et al.*, “PHISTO: pathogen--host interaction search tool,” *Bioinformatics*, vol. 29, no. 10, pp. 1357–1358, 2013.

[71] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert, “A new approach to collaborative filtering: Operator estimation with spectral regularization,” *J. Mach. Learn. Res.*, vol. 10, no. Mar, pp. 803–826, 2009.

[72] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Mach. Learn.*, vol. 29, no. 2–3, pp. 131–163, 1997.

[73] V. Muralidharan and V. Sugumaran, “A comparative study of Na{\"\i}ve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis,” *Appl. Soft Comput.*, vol. 12, no. 8, pp. 2023–2029, 2012.

[74] J. R. Quinlan, *C4.5: Programs for Machine Learning*, vol. 1, no. 3. 1992.

[75] J. R. Quinlan, “Improved use of continuous attributes in C4.5,” *J. Artif. Intell. Res.*, vol. 4, pp. 77–90, 1996.

[76] J. G. Cleary, L. E. Trigg, and others, “K*: An instance-based learner using an entropic distance measure,” in *Proceedings of the 12th International Conference on Machine learning*, 1995, vol. 5, pp. 108–114.

[77] D. Y. Mahmood and M. A. Hussein, “Intrusion detection system based on K-Star classifier and feature set reduction,” *IOSR J. Comput. Eng.*, vol. 15, no. 5, pp. 107–112, 2013.

[78] M. Bhasin and G. P. S. Raghava, “Classification of nuclear receptors based on amino acid composition and dipeptide composition,” *J. Biol. Chem.*, vol. 279, no. 22, pp. 23262–23266, 2004.

[79] J. Ruan, K. Wang, J. Yang, L. A. Kurgan, and K. Cios, “Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences,” *Artif. Intell. Med.*, vol. 35, no. 1–2, pp. 19–35, 2005.

[80] J. Chen, H. Liu, J. Yang, and K. C. Chou, “Prediction of linear B-cell epitopes using amino acid pair antigenicity scale,” *Amino Acids*, vol. 33, no. 3, pp. 423–428, 2007.

[81] S. MAETSCHKE, M. TOWSEY, and M. BODÉN, “Blomap: an Encoding of Amino Acids Which Improves Signal Peptide Cleavage Site Prediction,” in *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference*, 2005, pp. 141–150.

[82] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, “Prediction of protein folding class using global description of amino acid sequence.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 19, pp. 8700–4, 1995.

[83] M. Gök and A. T. Özcerit, “A new feature encoding scheme for HIV-1 protease cleavage site prediction,” *Neural Comput. Appl.*, vol. 22, no. 7–8, pp. 1757–1761, 2013.

[84] J. Guo, Y. Lin, Z. Sun, and A, “Novel Method for Protein Subcellular Localization: Combining Residue-Couple Model and,” in *SVM, in: Proceedings of Third Asia-Pacific Bioinformatics Conference, 17-21 January 2005, Singapore*, 2000, vol. pp, pp. 117–129.

[85] W. R. Taylor, “The classification of amino acid conservation,” *J. Theor. Biol.*, vol. 119, no. 2, pp. 205–218, Mar. 1986.

[86] S. L. Lo, C. Z. Cai, Y. Z. Chen, and M. C. M. Chung, “Effect of training datasets on support vector machine prediction of protein-protein interactions,” *Proteomics*, vol. 5, no. 4, pp. 876–884, 2005.

[87] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.-H. H. Kim, “Recognition of a protein fold in the context of the SCOP classification,” *Proteins Struct. Funct. Genet.*, vol. 35, no. 4, pp. 401–407, 1999.

[88] S. Maetschke, M. Towsey, and M. Boden, “BLOMAP: an encoding of amino acids which improves signal peptide cleavage site prediction,” in *Proceedings of the 3rd Asia-Pacific bioinformatics conference*, 2005, pp. 141–150.

[89] Z.-R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen, “PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence,” *Nucleic Acids Res.*, vol. 34, no. suppl_2, pp. W32--W37, 2006.

[90] H. B. Rao, F. Zhu, G. B. Yang, Z. R. Li, and Y. Z. Chen, “Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence,” *Nucleic Acids Res.*, vol. 39, no. suppl_2, pp. W385--W390, 2011.

[91] J. Garcia-Garcia *et al.*, “iFrag: A Protein–Protein Interface Prediction Server Based on Sequence Fragments,” *J. Mol. Biol.*, vol. 429, no. 3, pp. 382–389, 2017.

[92] H. B. Shen and K. C. Chou, “PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition,” *Anal. Biochem.*, vol. 373, no. 2, pp. 386–388, 2008.

[93] M. K. Wen Zhang, “Protein encoding: A Matlab toolbox of representing or encoding protein sequences as numerical vectors for bioinformatics,” *J. Chem. Pharm. Res.*, 2014.

ÖZGEÇMİŞ

İrfan Kösesoy, 2005 yılında başladığı Trakya Üniversitesi Bilgisayar Mühendisliği Bölümü'nü 2009 yılında bitirdi. 2009 yılındas Trakya Üniversitesi Bilgisayar Mühendisliği Bölümün'de yüksek lisans çalışmalarına başladı. 2010 yılında Yalova Üniversitesi bilgisayar mühendisliğinde Araştırma Görevlisi olarak çalışmaya başladı. 2011 yılında yüksek lisans çalışmasını tamamladı. 2012 yılında Sakarya Üniversitesi bilgisayar mühendisliği bölümünde doktora çalışmasına başladı. Halen Yalova Üniversitesi Bilgisayar Mühendisliği Bölümü'nde Araştırma Görevlisi olarak görev yapmaktadır.