T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YAPAY ZEKÂ TEKNİKLERİNİN KULLANIMIYLA PROTEİN ETKİLEŞİMLERİNİN SEKANS BİLGİSİNE DAYALI TAHMİNİ

Yunus Emre GÖKTEPE

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Temmuz-2019
KONYA

TEZ KABUL VE ONAYI

Yunus Emre GÖKTEPE tarafından hazırlanan "Yapay Zekâ Tekniklerinin Kullanımıyla Protein Etkileşimlerinin Sekans Bilgisine Dayalı Tahmini" adlı tez çalışması 12/07/2019 tarihinde aşağıdaki jüri tarafından oy birliği / [redacted] ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda DOKTORA TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof. Dr. Hakan IŞIK	.....
Danışman Doç. Dr. Halife KODAZ	.....
Üye Prof. Dr. Harun UĞUZ	.....
Üye Dr. Öğr. Üyesi Ömer Kaan BAYKAN	.....
Üye Dr. Öğr. Üyesi Şaban GÜLCÜ	.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. Hakan KARABÖRK
Enstitü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Yunus Emre GÖKTEPE

12.07.2019

ÖZET

DOKTORA TEZİ

YAPAY ZEKÂ TEKNİKLERİNİN KULLANIMIYLA PROTEİN ETKİLEŞİMLERİNİN SEKANS BİLGİSİNE DAYALI TAHMİNİ

Yunus Emre GÖKTEPE

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Halife KODAZ

2019, 104 Sayfa

Jüri
Doç. Dr. Halife KODAZ
Prof. Dr. Hakan IŞIK
Prof. Dr. Harun UĞUZ
Dr. Öğr. Üyesi Ömer Kaan BAYKAN
Dr. Öğr. Üyesi Şaban GÜLCÜ

Proteinler hücre içerisindeki tüm olaylarda önemli roller oynarlar ve biyolojik yapıların temelini oluştururlar. Görevlerini yerine getirirken proteinler diğer protein ve moleküllerle farklı yapı ve şekillerde iletişim kurarlar. Protein-protein etkileşimleri denilen bu yapılar tüm biyolojik süreçlerin yürütülmesinde görev alırlar. Bu sebeple etkileşimlerin belirlenmesi araştırmacıların yoğunlaştığı önemli bir çalışma konusudur.

Proteinler arasında kurulan etkileşimlerin tespit edilmesi için farklı yöntemler kullanılmaktadır. Bu yöntemler temel olarak in-vivo, in-vitro ve in-silico olarak gruplandırılmaktadır. Etkileşimlerin deneysel olarak belirlenmesi laboratuvar ortamında yapılan yüksek hacimli çalışmalar olarak değerlendirilmektedir. Bu çalışmalar çok fazla zaman ve çaba gerektirmektedir. Ayrıca araştırmacılar bu tür çalışmaların yanlış pozitif ve yanlış negatif oranlarının da oldukça yüksek olduğunu vurgulamaktadırlar. Laboratuvar ortamında yapılan bu çalışmaları desteklemek ve onlara ön bilgi sunmak gibi amaçlarla hesapsal etkileşim tahmini yöntemleri yoğun olarak çalışılmaktadır.

Bu çalışmada proteinler arasında meydana gelen etkileşimlerin tahmini için hesapsal metotlar önerilmiştir. Bu metotlarda eldeki veri tabanlarından alınan protein sekans bilgilerini kullanan özellik çıkarım adımları üretilmiştir. Bu şekilde önerilen metotlar tüm protein veri kümelerine uygulanabilmekte olup proteinler hakkında farklı özelliklerin bilinmesine ihtiyaç duyan hesapsal metotlardan farklılık göstermektedir. Elde edilen özellik matrisleri kullanılarak destek vektör makinesi tabanlı bir sınıflandırma sistemi ile etkileşim tahmini yapılmaktadır.

Önerilen sistemlerin performansları sık kullanılan değerlendirme ölçütleri ile test edilmiştir. Önceki çalışmalarla karşılaştırılarak önerilen sistemlerin farklı veri kümeleri üzerinde elde ettiği sonuçların başarılı ve kabul edilebilir oldukları görülmüştür.

Anahtar Kelimeler: destek vektör makineleri, makine öğrenmesi, özellik çıkarımı, protein-protein etkileşimleri, protein yapıları

ABSTRACT

Ph.D THESIS

PREDICTION OF PROTEIN INTERACTIONS BY USING ARTIFICIAL INTELLIGENCE TECHNIQUES BASED ON PROTEIN SEQUENCE DATA

Yunus Emre GÖKTEPE

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Assoc. Prof. Dr. Halife KODAZ

2019, 104 Pages

Jury
Assoc. Prof.Dr. Halife KODAZ
Prof.Dr. Hakan IŞIK
Prof.Dr. Harun UĞUZ
Assist. Prof.Dr. Ömer Kaan BAYKAN
Assist. Prof.Dr. Şaban GÜLCÜ

Proteins play important roles in all events in the cell and form the basis of biological structures. While carrying out their tasks, proteins communicate with other proteins and molecules in different structures and forms. These structures, called protein-protein interactions, are involved in carrying out all biological processes. Therefore, determining the interactions is an important subject of research.

Different methods are used to detect interactions between proteins. These methods are mainly divided into 3 groups as in-vivo, in-vitro and in-silico. Experimental determination of the interactions is evaluated as high throughput studies made in laboratory. These studies require a lot of time and effort. In addition, researchers emphasize that the false positive and false negative rates of such studies are quite high. In order to support these studies conducted in the laboratory environment and to provide them with preliminary information, computational interaction prediction methods are studied extensively.

In this study, computational methods have been proposed for the estimation of interactions between proteins. In these methods, feature extraction steps using the protein sequence information from the present databases were generated. The models proposed in this way can be applied to all protein datasets and differ from computational methods that need to know different and complex properties about proteins. By using the obtained feature matrices, interaction prediction was made with a support vector machine based classification system.

The performances of the proposed systems were tested with commonly used evaluation criteria. Compared to previous studies, it was seen that the results of the proposed systems on different datasets were successful and acceptable.

Keywords: feature extraction, machine learning, protein structures, protein-protein interactions, support vektor machines

ÖNSÖZ

Bu tezin yürütülmesinde bilgi ve tecrübesiyle beni yönlendiren, her konuda anlayış gösteren danışman hocam Doç. Dr. Halife KODAZ'a teşekkür ederek başlamak istiyorum. Onun rehberliği ve yardımları bu çalışmanın tamamlanmasını sağladı.

Değerli öneri ve katkılarıyla çalışmama destek veren tez izleme komitesinde yer alan Prof. Dr. Hakan IŞIK ve Dr. Öğr. Üyesi Ömer Kaan BAYKAN'a,

Değerli yorum ve önerileri için Prof. Dr. Şirzat KAHRAMANLI, Prof. Dr. Harun UĞUZ ve Dr. Öğr. Üyesi Şaban GÜLCÜ hocalarıma,

Tezimi okumak ve yorum yapmak için kıymetli zamanlarını bana ayıran değerli mesai arkadaşlarım Öğr. Gör. Dr. Ahmet ÇAL ve Öğr. Gör. Ergun ELİTOK'a,

Son olarak, sağladıkları her türlü destek, teşvik ve anlayış için sevgili aileme teşekkür ederim.

Yunus Emre GÖKTEPE
KONYA-2019

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

ÇİZELGELER LİSTESİ

SİMGELER VE KISALTMALAR

Kısaltmalar

AAC	Standart Aminoasit Kompozisyonu
AUC	Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan
DIP	Etkileşen Proteinler Veri Tabanı
DNA	Deoksiribo Nükleik Asit
DVM	Destek Vektör Makinesi
EM	Elektron Mikroskopisi
GPCR	G Protein Kenetli Reseptör
HPRD	İnsan Protein Referans Veri Tabanı
MINT	Moleküler Etkileşim Veri Tabanı
NCBI	Ulusal Biyoteknoloji Bilgi Merkezi
NHGRI	Ulusal İnsan Genomu Araştırma Enstitüsü
NLM	Birleşik Devletler Ulusal Tıp Kütüphanesi
NMR	Nükleer Manyetik Rezonans
PPE	Protein Protein Etkileşimi
PSSM	Konuma özel puanlama matrisi
RCSB PDB	Yapısal Biyoinformatik için Araştırma Ortaklığı Protein Veri Bankası
RNA	Ribo Nükleik Asit
ROC	Alıcı İşletim Karakteristiği
TAP	İkili Benzerlik Saflaştırması
UniProt	Uluslararası Protein Kaynağı Veri Tabanı
Y2H	Maya İki-Hibrit
WWPDB	Dünya Çapında Protein Veri Bankası Organizasyonu

1. GİRİŞ

Canlılardaki tüm hücresel süreçlerin temelinde proteinler bulunmaktadır. Proteinler, biyolojik yapıların temel unsurlarıdır ve biyolojik süreçler içerisinde çeşitli fonksiyonların yerine getirilebilmesi için gereklidirler. Hücrelerin yaşam döngüsünde meydana gelen yapısal, işlevsel ve düzenleme amaçlı görevlerin tamamı aminoasit yapıtaşlarından oluşan bu karmaşık moleküllerin varlığı ile gerçekleşmektedir (Yao ve ark., 2019).

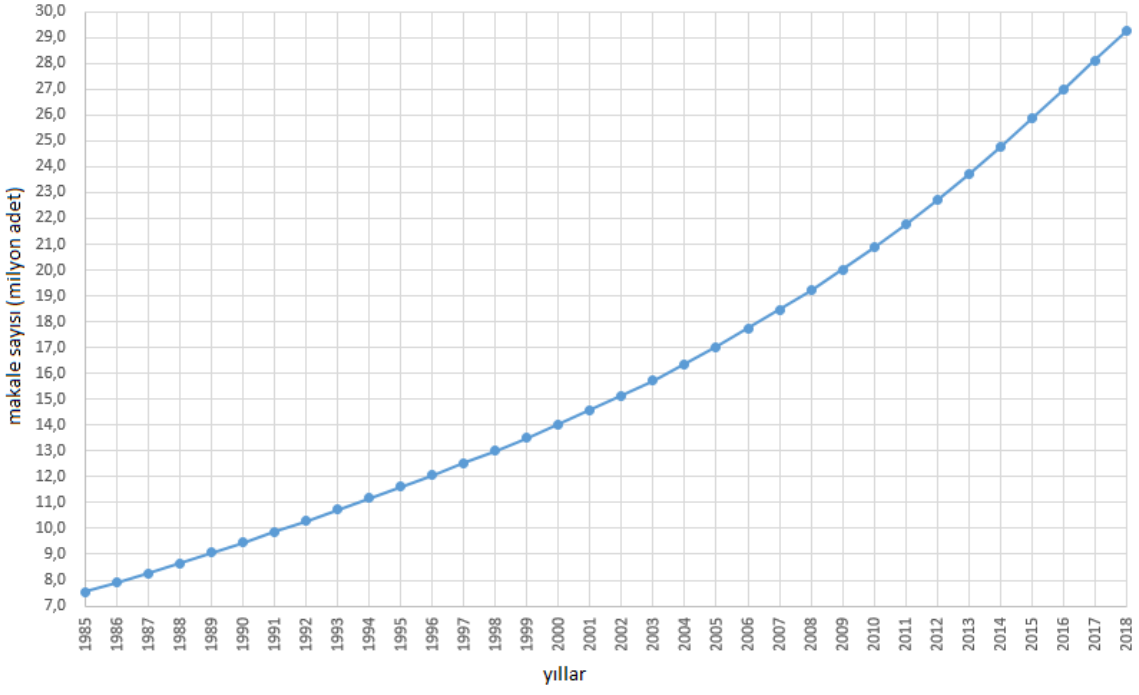
Bu görevlerin gerçekleşmesinde proteinler nadiren tek başlarına çalışırlar ve genellikle diğer moleküllerle birlikte etkileşim kurarlar. Proteinler arasında belirli bir amaç için kurulan bu etkileşimler Protein-Protein Etkileşimleri (PPE) olarak adlandırılır (Rao ve ark., 2014).

PPE'ler, hücresel döngü kontrolü, metabolik yolların çalışması, sinyal iletimi, besin alımı, DNA replikasyonu, RNA transkripsiyonu ve protein biyosentezi gibi birçok biyolojik olayın gerçekleşmesinde önemli rol oynar. Bununla beraber, anormal protein etkileşimlerinin alzaymır, Creutzfeldt-Jacob ve kanser gibi birçok hastalıkla ilişkili olduğu düşünülmektedir (Schuster-Bockler ve Bateman, 2008). Bu sebeple, biyolojik süreçlerin anlaşılması ve daha iyi analiz edilebilmesi, proteinlerin fonksiyonlarının tahmin edilmesi, proteinlerin hastalıklarla ilişkilendirilmesi ve yeni ilaçların tasarlanması gibi amaçlarla PPE'lerin tespit edilmesi çalışmaları büyük önem kazanmaktadır.

PPE tespit yöntemleri temel olarak *in vivo*, *in vitro* ve *in silico* olmak üzere 3 kategoride sınıflandırılır. Canlı organizma üzerinde belirli prosedürler uygulanarak gerçekleştirilen etkileşim tespit yöntemleri *in vivo* olarak anılır. *Maya iki-hibrit (Y2H, Y3H)* (Fields ve Song, 1989) ve *protein-parça tamamlama deneyleri* (Pelletier ve ark., 1999) bu grupta sayılabilir. Canlı organizmanın dışında bir laboratuvar ortamında kontrollü olarak gerçekleştirilen *Protein Mikrodizileri* (Templin ve ark., 2002)*, İkili Benzerlik Saflaştırması (TAP)* (Puig ve ark., 2001)*, Nükleer Manyetik Rezonans Spektroskopisi (NMR)* (Wuthrich, 1989) gibi teknikler *in vitro* sınıfındaki yöntemlerdendir. PPE tespitinin bilgisayar simülasyonu yoluyla yapıldığı yöntemler *in silico* sınıfındaki çalışmalardandır. *Sekans tabanlı çalışmalar*, *protein yapısına dayalı yaklaşımlar*, *filogenetik ağaçlar* (Pazos ve Valencia, 2001), *kromozom yakınlığı ve gen komşuluğu* (Dandekar ve ark., 1998), *gen ifadesine dayalı yöntemler* bu gruptaki çalışmalardan bazılarıdır.

Etkileşim tahmini için geliştirilen deneysel yöntemler yüksek hacimli yöntemler olarak ifade edilmektedir. Bu yöntemlerin etkileşim tahmininde genellikle önemli oranlarda yanlış-pozitif ve yanlış-negatif sonuçlar ürettiği belirtilmektedir (An ve ark., 2016). Aynı zamanda bahsedilen yöntemlerle çalışmanın çok fazla zaman gerektirdiği ve çok yüksek maliyetli olduğu bilinmektedir. Bununla beraber bilinen etkileşim sayıları da henüz çok sınırlı sayıda olduğu için bu konuda hâlâ yeni çalışmalara ihtiyaç duyulmaktadır (Stumpf ve ark., 2008). İnsan PPE'lerinin henüz yaklaşık olarak sadece %10 oranındaki bir kısmının bilindiği sanılıyor. Bu açıdan bakıldığında tespit edilmeyi bekleyen büyük miktarda PPE bilgisi bulunmaktadır. Bir başka açıdan, bilinen PPE'leri oluşturan proteinler insan proteinlerinin üçte ikisine karşılık gelmektedir. İnsan proteinlerinin yaklaşık üçte biri hakkında elimizde bilinen bir PPE bilgisinin bulunmadığı belirtilmektedir (Kotlyar ve ark., 2015).

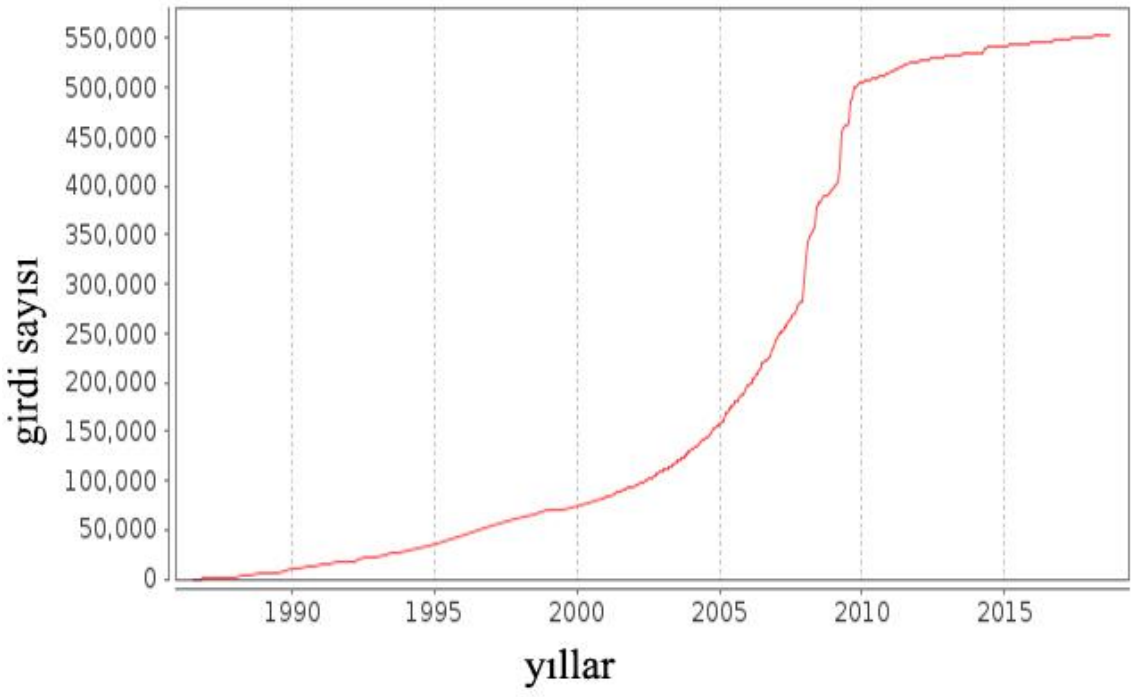
Teknolojik ilerlemelerin de motivasyonuyla biyoinformatik ve biyomedikal konularıyla ilgilenen araştırmacıların sunduğu çalışmalar her geçen gün daha da artmaktadır. Bu konudaki çalışmaların tutulduğu en önemli veri tabanlarından birisi Medline'dır (Medline, 2019). ABD Ulusal Tıp Kütüphanesi (U.S. National Library of Medicine - NLM) tarafından kurulan Medline veri tabanına 2018 yılı içerisinde 1,333,307 adet yeni makale eklenmiştir. Veri tabanının içerdiği makale sayısının yıllara göre değişimi Şekil 1.1'de görülmektedir.



Şekil 1.1. Yıllara göre Medline veri tabanındaki toplam makale sayısının ulaştığı rakamlar.

Yapılan yeni araştırmalar ile elde edilen protein sekans, fonksiyon ve etkileşimlerine ait veriler sürekli olarak artmaktadır. Protein sekanslarına ve fonksiyonlarına ilişkin kapsamlı bir veri tabanı olan UniProt'un 13 Şubat 2019 tarihinde yayınladığı UniProtKB/Swiss-Prot 2019_02 sürümü 559,228 adet sekans bilgisini içermektedir.

Şekil 1.2, UniProtKB/Swiss-Prot veri tabanının ihtiva ettiği sekans girdileri sayısının 1985-2019 yılları arasında göstermiş olduğu artışı ortaya koymaktadır (UniProt, 2019).



Şekil 1.2. Zaman içerisinde UniProtKB veri tabanındaki girdi sayısı (UniProt, 2019).

Etkileşim tahmini için kullanılan deneysel yöntemlerin çalışması için çok zaman gerekmesi, bu yöntemlerin pahalı olması ve sıklıkla yüksek yanlış pozitif ve yanlış negatif değerlerini üretmesi gibi dezavantajlara sahip olması sebebiyle, bu yöntemleri desteklemek için hesapsal yöntemlerin geliştirilmesine ihtiyaç duyulmaktadır (An ve ark., 2016). Bu amaçlar için araştırmacılar tarafından farklı tahmin metotları geliştirilmiş olmasına rağmen bu metotların çalışabilmesi için proteinler hakkında her durumda ulaşılması mümkün olmayan bir takım özelliklere ihtiyaç duyulduğu görülmektedir. Bununla beraber üretilen metotların tahmin doğrulukları henüz yeterince yüksek değerlere ulaşamamıştır.

Bu çalışmada proteinler arasında oluşan etkileşimlerin tahmin edilmesi için sekans tabanlı yöntemler üzerinde çalışılmıştır. Bu amaçla, protein birincil yapısından elde edilebilecek verileri kullanarak yüksek doğruluğa sahip etkileşim tahmin sistemleri önerilmektedir. Proteinlerin birincil yapısından elde edilen bilgiler kullanılarak daha etkin özellik çıkarım aşamaları geliştirilmeye çalışılmıştır. Bu özellik çıkarım yöntemleriyle eldeki etkileşim verileri daha anlamlı bir şekilde ifade edilebilmiştir. Bu veriler Destek Vektör Makineleri (DVM) tabanlı sınıflandırma sistemlerine tabi tutulmuştur. Elde edilen sonuçlar önerilen sistemlerin etkileşim tahmini doğruluk oranlarının daha iyi olduğunu göstermiştir.

1.1. Tez Organizasyonu

Tez çalışması altı ana bölümden oluşmaktadır;

İlk bölüm "*giriş*" bölümüdür. Burada tezle ilgili özet bilgiler, protein, protein-protein etkileşimleri (PPE) hakkında kısa tanımlamalar yapılmıştır.

İkinci bölümde PPE'ler ve tahmin yöntemleri ile ilgili literatür çalışmaları hakkında bilgi verilmiştir. Bu çalışmaların PPE tahminine olan katkıları açıklanmıştır. Mevcut sistemlerin sahip olduğu eksiklikler belirtilerek yetersiz kaldığı noktalar vurgulanmıştır. Bu çalışmada önerilen PPE tahmin yöntemlerinin amaçları anlatılmış ve üstünlüklerinden bahsedilmiştir.

Üçüncü bölümde tez çalışmasında kullanılan materyal ve yöntemlerden bahsedilmiştir. Bu bölümde proteinin yapısı, PPE, özellik çıkarım yöntemleri, DVM sınıflandırma yöntemi, performans değerlendirme ölçütleri ve tez çalışmasında kullanılan veri kümeleri hakkında detaylı bilgiler verilmiştir.

Dördüncü bölümde etkileşim tahmini amacıyla iki yeni metot önerilmiştir. Önerilen "Protein-Protein Etkileşimlerinin Ağırlıklandırılmış Bir Psödo Aminoasit Kompozisyonu Tabanlı Yöntemle Tahmini" ve "Sekans Tabanlı Bir Birleşik Metot Kullanarak Protein-Protein Etkileşim Tahmini" yöntemleri açıklanmış ve yöntemlerin aşamaları anlatılmıştır.

Beşinci bölümde önerilen sistemlerden elde edilen sonuçlar listelenmiştir. Bu sonuçlar literatürdeki diğer çalışmaların sonuçları ile farklı ölçütler üzerinden karşılaştırılarak incelenmiştir.

Son bölümde ise tez çalışması değerlendirilmiş ve elde edilen sonuçlar genel olarak değerlendirilmiştir. Tez çalışmasının genel katkıları anlatılmıştır. Sonraki çalışmalar konusunda önerilerde bulunulmuştur.

2. KAYNAK ARAŞTIRMASI

Bu bölümde, PPE'lerin önemi ve tahmin edilmesi konularını ele alan çalışmalara ilişkin geniş bir inceleme sunulmaktadır. Bu amaçla, ilk olarak protein-protein etkileşimlerinin tahmin edilmesinin önemi hakkındaki çalışmalardan bahsedilmiştir. Daha sonra genel olarak hesapsal etkileşim tahmin yöntemlerine ilişkin çalışmalardan bahsedilmiştir. Son olarak ise etkileşim tahminini sekans tabanlı yöntemlerle yapan çalışmalar incelenmiştir.

2.1. Protein-Protein Etkileşimleri (PPE) ile İlgili Çalışmalar

Proteinler bulundukları ortamda tek başlarına hareket etmeyip, bir ya da daha fazla sayıdaki diğer proteinlerle kalıcı ya da geçici olarak birleşerek bir işlevi yerine getirirler. Bu birleşime protein-protein etkileşimi denir ve bu etkileşimler hücre içi sinyal iletimi, metabolik yolların çalıştırılması, sentezleme ve taşıma gibi hücresel olayların hemen hepsinin temelini oluşturur. Bu sebeple araştırmacılar proteinler arasındaki mevcut etkileşimlerin tespit edilmesi, tespit edilen etkileşimlere ilişkin bilgilerin yeni çalışmalarda kullanılabilmesi için veri tabanlarında kaydedilmesi ve muhtemel yeni etkileşimlerin belirlenmesi amaçlarıyla çalışmalar yapmaktadır. Literatürde bu konudaki çalışmaların sayısı ve proteinler konusunda elde edilen veri miktarı zaman içinde hızlı bir artış sergilemektedir (Zahiri ve ark., 2013b).

Rao ve ark. (Rao ve ark., 2014), proteinlerin %80'inden fazlasının tek başlarına çalışmadıklarını ve genellikle bir ya da daha fazla yapıyla birleşerek fonksiyonlarını yerine getirdiklerini göstermektedir. Hücrenin biyokimyasını anlamak, sistem biyolojisini incelemek ve ilaç hedeflerini belirlemek gibi amaçlar doğrultusunda bir proteom içindeki proteinlerin etkileşimlerini bulmanın önemli olduğunu ortaya koymuştur. PPE verilerinin çok yüksek hacimlere sahip olmasından dolayı etkileşimlerin belirlenmesi için sadece deneysel yöntemleri kullanmanın yeterli olmadığını ve etkileşim tahminlerinin hesapsal yöntemlerle yapılmasının giderek daha zorunlu hale geldiğini göstermiştir.

Stumpf ve ark. (Stumpf ve ark., 2008) çalışmasında, organizmaların mevcut protein etkileşimlerinden yola çıkarak tam protein etkileşim ağının büyüklüğü hakkında güvenilir bir istatistiksel tahmin yapmanın mümkün olduğunu göstermişlerdir. Önerilen istatistiksel yöntemle, insanlarda protein etkileşim ağının tamamlandığında yaklaşık

650.000 adet etkileşimden oluşacağını hesaplamışlardır. Farklı organizmalar için tam protein etkileşim ağının muhtemel boyutları hakkında tahminler hesaplayan bu çalışma, bilinen etkileşim sayısının henüz çok yetersiz olduğunu göstermektedir.

Bilinen PPE ağının henüz tamamlanmış olmaktan çok uzak olduğunu belirten Kotlyar ve ark. (Kotlyar ve ark., 2015) çalışmalarında, güvenli bir şekilde PPE tahmini yapabilmek için *FpClass* isimli veri madenciliği tabanlı bir hesapsal yöntem önermektedir. İnsan PPE'lerinin henüz yaklaşık olarak sadece %10 civarındaki bir kısmının bilindiği ve bu bilinen PPE'lerin insan proteinlerinin üçte ikisi arasında meydana geldiği belirtilmekte ve yine insan proteinlerinin üçte birine ait bilinen bir etkileşim verisinin bulunmadığı gösterilmektedir. Çalışmada önerilen yöntemde, bir etkileşimin gerçekleşmesini sağlayabilecek tek protein özelliği yerine bir protein özellik grubunun varlığı aranır. Uyumlu özellik grubunun arandığı bu yöntemde aynı zamanda etkileşim olasılığını azaltabilecek olan uyumsuz özellik çiftleri de aranır. *FpClass* yöntemi ile düşük Yanlış Keşif Oranı elde edilmiş ve yöntemin önemli bir katkısının da, sunduğu yüksek doğrulama oranı ile interaktom haritalama için beraber çalışacak bir hesapsal-deneysel yaklaşım fikrini önermesidir.

Zahiri ve ark. (Zahiri ve ark., 2013a) çalışmasında, PPE'lerin metabolik döngüler, DNA kodlaması, RNA transkripsiyonu ve sinyal iletimi gibi hemen tüm hücresel işlemleri düzenlediği için çok önemli olduğu belirtilmiştir. Proteinler bu hücresel fonksiyonları diğer proteinlerle beraber kurduğu etkileşimlerle yerine getirmektedir. Bu etkileşimleri öğrenmek büyük önem arz ettiği için farklı deneysel metotlar geliştirilmiştir. Elde edilen genom sekans bilgisinin miktarı hızla arttığı için bilinen PPE sayısı çok geride kalmıştır. Bu sebeple PPE tahmini için deneysel metotları destekleyecek ve onlara muhtemel yeni etkileşimlerin tespit edilmesi konusunda ön bilgi sağlayacak hesapsal metotların üretilmesi gerekmektedir. Bu doğrultuda, PPE tahmini amacıyla geliştirilen birçok metot vardır ve bu metotlar kullanılan protein bilgisi açısından farklılaşmaktadır.

Genetik hastalıklara sebebiyet veren genlerin henüz tanımlanamadığını belirten çalışmada (Oti ve ark., 2006), fiziksel olarak etkileşim kuran proteinlerin aynı hücresel işlemde yer alma eğiliminde oldukları ve bu proteinlerin içinde olabilecek mutasyonların da benzer hastalıklara sebebiyet verebileceği öne sürülmüştür. Bu tür genetik hastalıklara sebebiyet veren genlerin PPE'ler üzerinden tahmin edilebileceği vurgulanmıştır. Çalışmada, 10894 adet insan proteini arasındaki 72,940 adet PPE bilgisi kullanılarak 432 hastalık geni bölgesi aranmıştır. Bir kısmının teyit edildiği 300'e yakın

muhtemel hastalık geni tahmininin yapıldığı belirtilmektedir. PPE'leri ortaya çıkarmanın hastalığa sebep olabilecek muhtemel konumlardaki genleri tespit etme olasılığını artırdığı gösterilmiştir. PPE bilgisi kullanımının bu alandaki çalışmalarda yaklaşık 10 kat daha fazla gelişme sağlayacağı ifade edilmiştir. Bilinen etkileşim sayısı artırılırsa bu tür çalışmalarda kullanılacak olan veri miktarı ve kalitesinin de artacağı ve bu konuda daha fazla ilerleme kaydedilebileceği vurgulanmıştır.

Protein etkileşimlerinin genellikle protein çiftlerinin ilişkiyi kuran sadece belirli parçaları arasında oluştuğunu belirten Liu ve ark. (Liu ve ark., 2009) bu parçaların bulunmasının önemli olduğunu vurgulamıştır. Bölge ya da domain olarak isimlendirilen bu alanların etkileşime aracılık ettiğini söylemişlerdir. Etkileşimleri domain seviyesinde anlamanın PPE ağlarını anlamayı, etkileşim bölgelerinin kesin olarak belirlenebilmesini, etkileşim bölgelerindeki olabilecek zararlı mutasyonların sebeplerinin anlaşılmasını ve patolojik protein etkileşimlerinin önlenmesi için ilaç geliştirilmesini sağlamak amaçlarıyla çok büyük öneme sahip olduğu belirtilmiştir. Bununla beraber bilinen domain-domain etkileşimlerinin (DDE) bilinmeyen PPE'lerin bulunmasında kullanılabileceği söylenmiştir. Çalışmada K-GIDDI (DDE'lerin bilgi güdümlü çıkarımı) isimli bir metot önerilmiştir. Bu metotla PPE'lerin daha küçük bölgelere indirgenmesi temel alınmıştır. K-GIDDI türler arası PPE ağlarından başlangıç bir DDE ağı çıkarmakta ve Gen Ontolojisi yardımıyla çalışan bir böl ve fethet kümeleme algoritmasıyla bu DDE ağını genişletmektedir. Elde edilen sonuçlar önerilen yöntemin performansının önceki DDE tahmin yöntemlerinden ya daha iyi ya da onlarla karşılaştırılabilir olduğu belirtilmiştir. Yöntemin yaptığı bazı tahminleri destekleyen biyolojik kanıtların bulunduğu ve PPE verilerini kullanan metotlarla belirlenemeyen DDE'lerin bu yöntemle bulunabildiği vurgulanmıştır.

Pierce ve ark. (Pierce ve ark., 2014) çalışmasında protein komplekslerine ilişkin deneysel olarak belirlenmiş yapı bilgisi bulunmadığından dolayı moleküler tabanını anlayabilmek için bu etkileşimlerin modellenmesine ihtiyaç duyulduğunu belirtmiştir. Protein-protein komplekslerinin yapılarının tahmini için katı cisim yerleştirme uygulaması olan ZDOCK ve M-ZDOCK tabanlı bir web sunucu sistemin kullanımını önermiştir. Bu sunucu yazılımı ile araştırmacıların protein-protein kompleksleri ve simetrik polimerlere ilişkin yapısal modelleri kolaylıkla elde edebilecekleri öne sürülmektedir.

Liu ve ark. (Liu ve ark., 2015) çalışmasında, yüksek hacimli biyolojik deneylerden elde edilen eksik etkileşimleri düzeltmek için saptanmış protein

kompleksleri uyarlanabilir *k-çekirdek* metodu ile budanmakta ve çözülmektedir. Önerilen yöntemin bir PPE ağındaki farklı yapı, sayı ve boyutlardaki düğümleri içeren protein komplekslerine uygun olduğu belirtilmiştir. Tahmin edilen PPE'lerin güvenilir olduğu ve protein etkileşim ağlarındaki eksik bilgileri tamamlayabildiği vurgulanmıştır. Geliştirilen etkileşim ağlarının protein komplekslerini tespit etme ve kompleksler arası ilişkileri anlama konusunda katkıda bulunacağı öne sürülmüştür.

2.2. Hesapsal Etkileşim Tahmini Yöntemleri ile İlgili Çalışmalar

Zaman içinde proteinlere ilişkin elde edilen genomik dizi bilgisi hızla artmasına rağmen proteinlere ilişkin bilinen fonksiyon bilgisinin miktarının ve proteinlerin yaptığı etkileşimlere ilişkin bilinen verilerin oranının henüz çok yetersiz olduğu araştırmacılar tarafından sıkça vurgulanmaktadır. Bu alanda deneysel yöntemlerle yeni veriler elde etmek yoğun emek ve zaman gerektirmektedir. Yeni verilerin elde edilmesini kolaylaştırmak ve deneysel çalışmalara ön bilgi sağlamak gibi amaçlarla etkileşim tahmininde hesapsal yöntemlerin kullanılması gerekmektedir (Zahiri ve ark., 2013a; Kotlyar ve ark., 2015).

Bock ve Gough (Bock ve Gough, 2003), genetik olarak benzer organizmalardaki protein etkileşimlerine dair bir veri tabanı geliştirmek amacıyla, daha önce kendileri tarafından önerilmiş olan *filogenetik önyükleme* adındaki algoritmayı proteom çapında etkileşim madenciliği için geliştirmişlerdir. Önerilen etkileşim madenciliği insan mide bakterisi olan 1039 adet doğrulanmış *Helicobacter pylori* veri kümesi üzerinde kurulan bir öğrenme sistemi ile gerçekleştirilmiştir. 10 kat çapraz geçerleme ile kesinlik ve hassasiyet ölçütlerinde sırasıyla %80 ve %69 değerleri elde edilmiştir.

Ben-Hur ve Noble (Ben-Hur ve Noble, 2005) çalışmasında, organizmaların etkileşim ağlarının tamamlanmış olmaktan çok uzak olduğunu ve hesapsal etkileşim tahmin yöntemlerine olan ihtiyacın giderek arttığını ifade etmiştir. PPE tahmini için veri kaynaklarının birleşimini kullanan bir metot üzerinde çalışılmıştır. Çalışmada, PPE tahmininde kullanılan çekirdek modelleri gösterilmiş ve performansı artırabilmek için bu çekirdeklerin birlikte kullanımı önerilmiştir.

An ve ark. (An ve ark., 2016) çalışmasında, PPE'lerin tahmin edilmesinin, biyolojik sistemlerde çalışan mekanizmaları anlayabilmek için gerekli olan protein etkileşim ağlarının oluşturulabilmesi için, çok önemli bir görev olduğunu söylemiştir.

Bu görevi yerine getirmeye çalışan birçok deneysel yöntemin geliştirilmiş olmasına rağmen bu yöntemlerin yüksek maliyet gerektirmesi, çok uzun zaman alması ve sıklıkla yanlış pozitif sonuçlar üretmesi gibi dezavantajlarının olduğu anlatılmaktadır. Bu dezavantajlardan dolayı araştırmacıların hesapsal etkileşim tahmin yöntemleri üzerine yoğunlaştığı ve bu alanda hâlâ yeni çalışmalara ihtiyaç duyulduğu belirtilmektedir. Bu çalışmada, etkileşim tahmini için *Bi-Gram ihtimalleri* ile *Bağlantı Vektör Makineleri* yönteminin beraber kullanıldığı bir yöntem önerilmektedir. Özellik vektörü *Bi-Gram* yöntemiyle elde edilmektedir ve *Temel Bileşenler Analizi* ile gürültü azaltışı yapılmaktadır. Elde edilen vektör *Bağlantı vektör makinesi* yöntemi ile sınıflandırmaya tabi tutulmaktadır. 5 kat *çapraz geçerleme* uygulanarak *Maya* ve *Helicobacter Pylori* veri kümelerinde sırasıyla %94.57 ve %90.57 doğruluk oranlarının elde edildiği belirtilmektedir.

PPE'lerin hücresel olayların büyük çoğunluğundan sorumlu olduğunu belirten ve PPE tahmini için geliştirilen deneysel yöntemlerin eksikliklerini ortaya koyan You ve ark. (You ve ark., 2014a) *Aşırı Öğrenme Makinesi* tabanlı bir etkileşim tespit modeli önermiştir. *Aşırı öğrenme Makineleri*, tek gizli katmanlı ileri öğrenme sinir ağlarına benzer bir yapıya sahiptir. Bu tür yapılarda gizli katmandaki başlangıç ağırlık değerleri sistem tarafından rastgele atanan modellerdir. Atanan bu değerler eğitim süresi boyunca güncellenmektedir. Bu modelde öğrenme süreci, diğer modellerle kıyaslandığında, daha hızlı gerçekleşmektedir. Önerilen sistem protein dizilerine ilişkin yerel ve genel özellik değerlerini çıkararak özellik vektörünü oluşturmaktadır. *İnsan* (İnsan Protein Referans Veri Tabanı) veri kümesi üzerinde çalıştırılan ve 5 kat çapraz geçerleme ile elde edilen sonuçlara göre önerilen sistemin %84.8 oranında bir doğruluğa sahip olduğu belirtilmiştir.

Hue ve ark. (Hue ve ark., 2010) çalışmasında, çözülmüş protein üç boyutlu yapılarının sayısının gittikçe arttığını, iki protein arasındaki olası etkileşimlerin tahmini için bu bilgileri kullanan hesapsal metotların daha çok önem kazanmaya başladığını belirtmiştir. Çalışmanın amacı, bir etkileşimin pahalı deneysel yöntemlerle gerçekleştirilmesinden önce protein çiftleri arasındaki olası etkileşimlere ilişkin geniş ölçekli bir görüntülemenin sunulmasıdır. Bu amaçla, iki proteininin etkileşip etkileşmediğini tahmin etmek için hesapsal bir metot önerilmiştir. Çalışmada, protein çiftleri arasındaki geniş ölçekli etkileşim tahmini için sekiz adet makine öğrenmesi metodu ile çalışılmıştır. Bu metotlarda sınıflayıcı olarak bir sinir ağı ya da bir DVM kullanılmıştır. Yazarlar önerilen metotta proteinler hakkında yapısal ve sekans

bilgilerinin kullanıldığını belirtmiştir. Her bir etkileşimin ayrı ayrı tahmin edilmesi de yöntemin dezavantajı olarak değerlendirilmiştir.

Organizmalar için yapısal bilgileri tam olarak bilinen protein sayısı bilinen toplam protein sayısı ile karşılaştırıldığında henüz oldukça az olmasına rağmen, PPE tahmininde proteinlerin yapısal bilgilerini kullanan çalışmalar da önerilmektedir. Zhang ve ark. (Zhang ve ark., 2012a) çalışmasında PPE tahmini için proteinlerin 3-boyutlu yapısı hakkındaki bilgilerin kullanılabileceğini söylemiştir. Yapısal olan ve yapısal olmayan etkileşim verilerini Bayes istatistiklerini kullanarak birleştiren bir PPE tahmin yaklaşımı önerilmiştir. Bu yaklaşımda kurulan *PrePPI* isimli algoritmanın elde ettiği sonuçların beklenmedik PPE'leri ortaya çıkarabildiği ve etkileşim tahmininde yapısal bilgi kullanımının oldukça etkili bir yöntem olduğu anlatılmıştır.

Chen ve Liu (Chen ve Liu, 2005) çalışmasında, proteinlerin belirli bölgelerden oluştuğunu ve PPE'lerin de bu bölgelerin etkileşim kurmasıyla meydana geldiğini ifade etmektedir. Bu fikirden yola çıkarak etkileşimlerin bulunabilmesi için PPE tahmini ikili bir sınıflandırma problemine dönüştürülmektedir ve bölge tabanlı bir Rastgele Orman metodu önerilmektedir. Çalışmada 3713 protein arasındaki 9,834 PPE verisi ikiye bölünerek test ve eğitim verisi olarak kullanılmaktadır. Etkileşim verileri içinde olmayan rastgele üretilmiş 8000 adet etkileşmeyen çift de negatif veri kümesini oluşturulmaktadır. Önerilen sistemin değerlendirilmesi ise Hassaslık ve Belirlilik karşılaştırma ölçütleri uygulanarak yapılmaktadır. Metot, Hassaslık ve Belirlilik ölçütlerinde sırasıyla %79.78 ve %64.38 değerlerini elde etmektedir.

Nanni (Nanni, 2005a) çalışmasında, farklı uzunluklardaki aminoasit sekanslarını, etkileşen ve etkileşmeyen protein çiftlerini ayırt edebilmek için gereken bilgiyi içerecek şekilde, vektör yapılarıyla göstermeyi amaçlamıştır. Bu amaç için K-yerel hiperdüzlem yöntemini kullanmıştır. Protein çiftlerini oluşturan her bir protein için 2-gram özellikleri çıkartılmaktadır ve her bir çift için bu özellik vektörlerinin toplamı hesaplanmaktadır. Veri kümesi olarak *Helicobacter Pylori* ve *Human* veri setlerini kullanan çalışma bir *K-En Yakın Komşuluk* metoduyla sınıflandırma yapılmaktadır. Önerilen yöntemin *Helicobacter Pylori* ve *Human* veri setleri için elde ettiği en yüksek değerler sırasıyla, doğruluk ölçütünde 0.84 ve 0.66, kesinlik ölçütünde 0.84 ve 0.67 ve hassaslık ölçütünde 0.86 ve 0.68 olmuştur.

PPE bölgelerinin tahmini konusunun işlemsel biyolojinin çok ilgi çeken bir alanı olduğunu belirten Deng ve ark. (Deng ve ark., 2009), konunun çözülmüş olmaktan çok uzak olduğunu ifade etmiştir. Bu çözümün önündeki engeller olarak; PPE bölgelerini

tam olarak tanımlayan biyolojik özelliklerin çıkarılamaması, PPE bölgelerinin tahmini için aminoasit sekanslarından çıkarılan bilgilerin benzer olduğu ve etkileşen protein bölgelerinin sayısının etkileşmeyen bölge sayısına göre çok az olmasından dolayı bir dengesizlik probleminin varlığı gösterilmektedir. Çalışmada dengesizlik problemini çözmek ve protein etkileşim bölgelerinin tahmin performansını arttırmak için bootstrap yeniden örneklendirme tekniği, DVM tabanlı birleşik sınıflayıcılar ve ağırlıklandırılmış oylama stratejisini beraber kullanan bir yöntem önerilmiştir. Yöntemde aminoasit sekanslarından elde edilen özellikler 4 gruba ayrılmıştır. Bu gruplar geleneksel dik kodlama ve Otokovaryans (AC) metotlarıyla dönüştürülerek 8 farklı özellik uzayı elde edilmiştir. Yazarlar, kullanılan birleşik sınıflayıcının 10 kat çapraz geçerleme ile eğri altındaki alan (AUC), hassaslık ve belirlilik değerlendirme ölçütlerinde sırasıyla 0.86, 0.76 ve 0.78 değerlerini elde ettiğini belirtmişlerdir. Çalışmada elde edilen sonuçlarla birleşik sınıflayıcıların iyi performans gösterdiği öne sürülmektedir.

Proteinler arasındaki etkileşimlere sebep olan bölgeleri temsil eden özelliklerin çıkarılmasının önemli bir sorun olduğunu belirten Cho ve ark. (Cho ve ark., 2009), bu konuda yeni özellik çıkarımları önermiştir. Önerilen özellikleri geleneksel özelliklerle birleştirerek etkileşim bölgelerinin tahmini için bir model geliştirmişlerdir. Özellik seçimini karar ağacı tabanlı bir yöntemle yapan çalışmada, seçilen özellikler 2 ayrı DVM tabanlı tahmin modeli ile işlenmektedir. Önerilen modelin performansı 10 kat çapraz geçerleme yöntemi ile değerlendiriliyor. Modelin elde ettiği sonuçların, seçilen özelliklerle, etkileşim bölgelerinin enerjik özelliklerini iyi bir şekilde yansıttığı ifade edilmektedir.

Tian ve ark. (Tian ve ark., 2019) çalışmasında, PPE tahmini için çoklu bilgi birleşimine dayanan bir yöntem önermiştir. PPIs-WDSVM adını verdikleri yöntem ile önce PseAAC (Psödo aminoasit kompozisyonu), AC (Oto kovaryans) ve EBGW (Grup Ağırlığına Bağlı Kodlama) metotları kullanılarak *Helicobacter pylori* ve *Saccharomyces Cerevisiae* veri kümelerinde protein sekanslarına ilişkin özellik çıkarımları yapılmaktadır ve çıkarılan bu özellikler birleştirilmiştir. Sonra birleştirilen özellik vektörü üzerinde gereksiz verilerin yok edilmesi için 2-boyutlu dalgacık gürültü azaltma metodu uygulanmıştır. Elde edilen gürültüden arındırılmış özellik vektörü DVM sınıflayıcısı tabanlı bir tahmin sisteminde girilerek değerlendirilmiştir. Elde edilen sonuçların önerilen sistemin PPE tahmin başarısını artırdığı belirtilmiştir.

2.3. Sekans Tabanlı Etkileşim Tahmini ile İlgili Çalışmalar

PPE tahmini için geliştirilmiş olan birçok hesapsal metot bulunmasına rağmen bu metotların çoğunun etkileşim tahmini yapabilmek için proteinlere ilişkin farklı yapısal bilgilere ihtiyaç duyduğu görülmektedir. Bu tür verilerin her durumda elde edilememesinin bir dezavantaj olarak ortaya çıkması ve hesapsal tahmin sistemlerinin daha hızlı ve etkili çalışabilmesi için protein hakkında sadece sekans dizilimi üzerinden elde edilebilecek verileri kullanan sekans-tabanlı etkileşim tahmin sistemlerine ihtiyaç duyulmaktadır.

Martin ve ark. (Martin ve ark., 2005) çalışmasında, PPE tahmini için proteinlerin sekans tabanlı verilerini deneysel yöntemlerden elde edilen verilerle birleştirmeyi önermektedir. Önerilen metot, etkileşen proteinlerin işaret tanımlayıcı değerlerini kullanan yeni bir tanımlama yöntemi geliştirmiştir. Bu yeni işaret bileşikleri DVM tabanlı bir sınıflandırma sistemine uygulanmıştır. Önerilen metot *Helicobacter pylori*, *insan* ve *fare* verilerine uygulanmıştır. Sonuç olarak, 10 kat çapraz geçerleme ile %70 ila %80 arasında doğruluk değerlerine ulaşılmıştır.

Nanni ve Lumini (Nanni ve Lumini, 2006) PPE tahmini için K-en yakın komşuluk metodunun değiştirilmiş bir formu olan HKNN K-yerel hiperdüzlem mesafesi en yakın komşuluğu sınıflayıcılarının birleşimini önermiştir. HKNN sınıflayıcılarının her biri aminoasitlerin farklı bir fizyokimyasal özelliğini kullanarak eğitilmiştir. Aminoasit dizileri, aminoasit indisleri ve aminoasit kompozisyonlarının 2-gram özelliklerini beraber kullanan bir yöntemle kodlanmıştır. Sistem, *Helicobacter pylori* veri tabanı üzerinde doğruluk, kesinlik ve duyarlılık ölçütlerinde sırasıyla 0.866, 0.858 ve 0.885 en iyi değerlerini elde etmiştir. İnsan veri tabanı üzerinde ise doğruluk, kesinlik ve duyarlılık ölçütlerinde sırasıyla 0.7, 0.7 ve 0.708 en iyi değerleri elde edilmiştir.

Derin öğrenme algoritmalarının kullanımı zamanla biyoinformatik de dâhil olmak üzere birçok araştırma konusunda gittikçe artmaktadır. Fakat bu yöntemin PPE tahmini alanında kullanımı henüz yeterli ilgiyi görmemiştir. Sekans tabanlı PPE tahmini problemi için derin öğrenme algoritması kullanan bir yöntem öneren Sun ve ark. (Sun ve ark., 2017), insan PPE tahmini konusunda önceki çalışmaları geride bırakan sonuçların elde edildiğini göstermişlerdir. Çalışmada önerilen derin öğrenme algoritması yığın oto-kodlayıcı yapısından oluşmaktadır. Bu yapı çok boyutlu veriyi önce gizli katmanda sıkıştıran ve sonra yeniden üreten danışmansız bir öğrenme

algoritması olan bir yapay sinir ağı örneğidir. Araştırmacılar özellik çıkarımı olarak *otokovaryans* (Guo ve ark., 2008) ve *birleşik üçlü* (Shen ve ark., 2007) yöntemlerini kullanmıştır. Sekans tabanlı PPE tahmini için uygulanan bu derin öğrenme algoritmasında elde edilen en iyi modelin, insan PPE verilerini içeren HPRD (İnsan Protein Referans Veri Tabanı) veri kümesi üzerinde yapılan çalışmalarda, 10 kat çapraz geçerleme ile %97.19 doğruluk değerine ulaştığı belirtilmiştir.

Proteinler arasındaki etkileşimlerin tahmini için tek bir sınıflandırma sisteminin yüksek bir doğruluk oranına ulaşamayacağını ifade eden Xia ve ark. (Xia ve ark., 2010a) çalışmasında, sekans tabanlı bir çoklu sınıflandırma sistemi önermiştir. Etkileşen protein çiftlerinin kodlanması için iki protein sekansı arasındaki korelasyonu fizyokimyasal özelliklerine göre belirleyen *moran otokorelasyon* tanımlayıcısı kullanılmıştır. PPE tahmini için *rotasyon ormanı* çoklu sınıflandırma sistemi tercih edilmiştir. *Rotasyon ormanının* dönüşüm metodu olarak *temel bileşenler analizi* kullanılmıştır. Çoklu sınıflandırma sistemindeki her bir sınıflayıcı için taban sınıflayıcı olarak WEKA kütüphanesindeki *J48* karar ağacı seçilmiştir. Önerilen metot *saccharomyces cerevisiae* ve *helicobacter pylori* veri kümelerinde test edilmiştir.

Deneysel yöntemlerle elde edilmiş olan PPE çiftlerinin muhtemel PPE ağının sadece bir parçası olduğunu işaret eden çalışmada (You ve ark., 2013) etkileşim tahmininin sadece protein sekans bilgisi kullanılarak yapıldığı bir yöntem önerilmektedir. Çalışmada, DIP veri tabanından alınan 11,188 adet protein çifti için *otokovaryans, birleşik üçlü, yerel tanımlayıcı* ve *moran otokorelasyon* yöntemleri kullanılarak her bir protein sekansından dört çeşit sekans tabanlı veri elde edilmektedir. Karmaşıklığı azaltmak ve tahmin doğruluğunu arttırmak için *Temel Bileşenler Analizi* ile özellik azaltışı yapıldıktan sonra önerilen *Birleşik Aşırı Öğrenme Makinesi* yöntemi ile sınıflandırma işlemi yapılmaktadır. Önerilen yöntem, *Saccharomyces cerevisiae* maya türüne ait PPE verileri üzerinde doğruluk, hassaslık ve kesinlik ölçütlerinde sırasıyla %87, %86.15 ve %87.59 sonuçlarını elde etmiştir. *Helicobacter pylori* veri kümesi üzerinde aynı ölçütlerde sırasıyla %87.5, %88.95 ve %86.15 sonuçlarını elde etmiş, *Escherichia coli* bakterisine ait PPE verisi üzerinde ise %87.5 kesinlik değerine ulaşmıştır.

PPE tahminini etkili ve doğru şekilde yapabilen yaklaşımlara olan ihtiyaca dikkat çeken Wang ve ark. (Wang ve ark., 2017), DNN-LCTD isimli sekans tabanlı bir etkileşim tahmin yöntemini önermiştir. Derin sinir ağları ve yeni bir yerel birleşik üçlü özellik gösteriminin bir birleşimi olan yöntem *Saccharomyces cerevisiae* veri kümesi

üzerinde test edilmiştir. Yerel birleşik üçlü özellik gösteriminin (LCDT) sırasal olarak uzak fakat uzaysal olarak yakın olan aminoasit kalıntıları arasındaki etkileşimleri yerel tanımlayıcılar (LD) ve birleşik üçlü (CT) metotlarından daha iyi açıkladığı belirtilmiştir. Yöntem önce LCDT ile aminoasit sekanslarından özellik bilgisini elde etmektedir. Bu bilgilerle üç gizli katmanlı bir sinir ağını eğitmektedir ve eğitilen sistem yeni PPE tahmini için kullanılmaktadır. Önerilen sistem doğruluk, kesinlik, hassaslık ve alıcı işletim karakteristik eğrisi (ROC) altındaki alan değerlendirme ölçütlerinde sırasıyla %93.12, %93.75, %93.83 ve %97.92 sonuçlarını elde etmiştir.

You ve ark. (You ve ark., 2014b) çalışmasında, etkileşim tahmini için çok ölçekli sürekli ve süreksiz özellik gösterimi ve DVM yöntemlerini kullanan sekans tabanlı bir yaklaşım önerilmiştir. Bu çalışmada kullanılan çok ölçekli sürekli ve süreksiz özellik gösteriminin sırasal olarak birbirine uzak ama uzaysal olarak birbirine yakın olan aminoasitlerin temsili açısından etkili bir yöntem olduğu vurgulanmıştır. Özellik çıkarımından elde edilen etkileşen ve etkileşmeyen çiftlere ilişkin veriler aynı boyutlu numerik vektörlere dönüştürüldükten sonra özellik fazlalığını ve karmaşıklığı azaltmak ve optimal bir özellik kümesini seçebilmek için minimum fazlalık maksimum ilişki ölçütü uygulanmıştır. Etkileşim olasılığı tahmini için DVM tabanlı bir tahmin modeli geliştirilmiştir. Önerilen sistemin *Helicobacter pylori* veri kümesi üzerinde çalıştırılmasıyla hassaslık, kesinlik, doğruluk ve Mathews korelasyon katsayısı ölçütlerinde sırasıyla %83.24, %86.12, %84.91 ve %74.40 değerleri elde edilmiştir. *Maya* veri kümesi üzerindeki çalışmalarda ise aynı ölçütlerde sırasıyla %90.67, %91.94, %91.36 ve %84.21 sonuçlarına ulaşılmıştır.

Deneysel yöntemlerin, zaman tüketimi, yüksek maliyet ve yanlış pozitif oranları gibi kaçınılmaz dezavantajlarının olduğunu ortaya koyan çalışmada Huang ve ark. (Huang ve ark., 2015), PPE tahmini için protein sekans bilgilerini kullanan hesapsal bir metot önermektedirler. Bu metotta, *Değişim Matris Gösterimi* üzerinde *Ayrık Kosinüs Dönüşümü* yöntemi uygulanarak kullanılarak protein sekanslarının yeni bir gösterimi elde edilmiştir. *Ağırlıklandırılmış Ayrık Gösterim tabanlı Sınıflayıcı* ile sınıflandırma işlemi daha iyi yapılmaya çalışılmıştır. *Helicobacter pylori* veri kümesi üzerinde yapılan çalışmalarda doğruluk, kesinlik, duyarlılık, Mathews korelasyon katsayısı ve eğri altındaki alan ölçütlerinde sırasıyla %86.74, %87.01, %86.43, %76.99 ve % 89.85 değerleri elde edilmiştir.

Wei ve ark. (Wei ve ark., 2017) çalışmasında, veri kümesi oluşturma, özellik çıkarma ve sınıflandırma aşamalarından oluşan yeni bir hesapsal yöntem önermişlerdir.

Bu çalışmada etkileşim veri tabanlarında bulunan PPE (protein-protein etkileşimi) ve NPPE (protein-protein etkileşimi olmaması) verileri arasındaki orantısızlık ve bunun sonucunda oluşan dengesiz veri problemi çözülmeye çalışılmaktadır. NPPE verileri iki protein arasında bir etkileşim olmadığını göstermektedir ve fiziksel olarak gerçekten bir genetik etkileşimin olmayacağını ispatlamak çok zordur. Bu sebeple eldeki NPPE verilerinin yanlış negatif sonuçlar üretmesi kaçınılmaz olmaktadır. Bu çalışma üç farklı yöntem kullanılarak yeni bir negatif etkileşim veri kümesi oluşturulmaktadır. Bu küme oluşturulurken kullanılan yöntemlerden birisi eldeki NPPE veri kümeleri üzerinde çalışan rastgele seçim yöntemi olurken diğer ikisi pozitif etkileşim verileri üzerinde çalışan *RandomPairs* ve *RecombinePairs* (Shen ve ark., 2007) metotlarıdır. Özellik çıkarım metotlarıyla bir özellik kümesi üretilmektedir ve ardından sınıflandırmaya tabi tutulmaktadır. Sonuç olarak, PPE tahmini için en iyi değerleri elde eden yöntemin *RandomPairs* veri kümesi, 188 boyutlu bir özellik vektörü ve LibD3C (Lin ve ark., 2014) sınıflandırma sisteminin beraber kullanımıyla oluşturulduğu belirtilmiştir.

PPE tahmin kalitesini arttırmak için bir toplu tahmin yöntemi öneren Xia ve ark. (Xia ve ark., 2010b) altı farklı tahmin yönteminin sonuçlarını birleştirerek PPE tahmini yapan DVM tabanlı bir sistem geliştirmişlerdir. İki aşamadan oluşan yöntem önce aminoasit sekans bilgilerini altı ayrı DVM tabanlı tahmin sistemine sokmaktadır ve elde ettiği tahmin sonuçlarını toplamaktadır. Sonra bu değerleri kullanarak bir girdi vektörü oluşturup DVM tabanlı toplu-tahmin yönteminde işlemektedir. Önerilen yöntemin *Saccharomyces cerevisiae* veri kümesi üzerinde diğer altı yöntemden daha yüksek tahmin doğruluğu elde ettiği belirtilmiştir. Yöntemin türler-arası veri kümelerinde de iyi sonuçlar ürettiği öne sürülmektedir.

Yu ve ark. (Yu ve ark., 2010) çalışmasında, etkileşen protein çiftlerinin birincil yapılar üzerinden tahmin edilebileceğini söylemiştir ve sadece sekans bilgisine dayanan bir PPE tahmin sistemi önerilmiştir. Çalışmada protein sekanslarını özellik vektörlerine dönüştürmek için olasılık tabanlı bir sistem geliştirilmiştir. PPE tahmini için bir çekirdek yoğunluk tahmini algoritması olan RVKDE (gevşemiş değişken çekirdek yoğunluk tahmincisi) makine öğrenmesi yönteminden yararlanılmıştır. RVKDE yönteminin performansının da parametrelerin doğru seçilmesine bağlı olduğu belirtilen çalışmada bir grid araması ile parametre optimizasyonunun yapıldığı belirtilmiştir. Önerilen yöntemin iyi bir tahmin performansı gösterdiği öne sürülmüştür.

Chen ve Jeong (Chen ve Jeong, 2009) çalışmasında, protein sekans bilgi sayısının giderek artmasından dolayı protein etkileşim bölgelerinin, sadece protein

sekans bilgisi kullanarak tahmininin de giderek daha önemli olduğunu öne sürmektedir. Bu sebeple hiçbir yapısal bilgiye ihtiyaç duymadan protein sekansından geniş ölçekte özellik çıkaran bir yöntem önerilmiştir. Önerilen yöntem protein veri bankasından (PDB) elde edilen polipeptit zincirlerine ait veriler üzerinde değerlendirilmiştir. Yöntem; fizyokimyasal özellikler, aminoasit uzaklıkları ve pozisyona özel skor matrisleri (PSSM) olmak üzere 3 grup özellik çıkarımı yapmaktadır. Bu özellikleri etkili bir şekilde kullanan rastgele orman tabanlı bir metot geliştirilerek bağlanma bölgesi tahmini yöntemlerinde sık karşılaşılan dengesiz veri sınıflandırma sorununu çözmeye çalışmışlardır. Her özellik grubu için her birisi 100 adet rastgele seçilen özellikle kurulan 100 adet ağaç üretilmektedir. Bir örnek için sınıflandırma kararı ağaçların sınıflandırma çoğunluğuna göre belirlenmektedir. Elde edilen sonuçların diğer sekans tabanlı yöntemlere göre daha başarılı olduğu ifade edilmektedir.

Zhang ve ark. (Zhang ve ark., 2014) çalışmasında, PPE tahmini için uygulanan deneysel yöntemlerin yanlış pozitif ve yanlış negatif oranları, çok zaman ve emek gerektirmesi ve yüksek maliyetli olması gibi dezavantajlarının bulunduğunu belirtmektedir. Bu dezavantajlarından dolayı etkileşim tahmini için hesapsal metotlara duyulan gereksinim vurgulanmaktadır. Literatürdeki hesapsal metotların; sekans tabanlı, genom tabanlı ve yapı tabanlı olarak üç gruba ayrılabileceği belirtilmektedir. Genom tabanlı ve yapı tabanlı metotların ise proteinler hakkında belirli ön bilgilere ulaşmadan uygulanamayacağı ifade edilmektedir. Sekans tabanlı metotların ise daha genel olduğu ve her durumda uygulanabileceğinden bahsedilmektedir. Çalışmada etkileşim tahmini için, bir özellik çıkarım yöntemi uygulanarak ikili çekirdek fonksiyonlu bir SVM modeli önerilmektedir.

Literatürdeki çalışmalarda vurgulandığı gibi, etkileşim tahmini için önerilen deneysel yöntemlerin çalışması için çok zamana ihtiyaç duyulmaktadır. Ayrıca bahsedilen yöntemlerin pahalı olması ve yüksek oranda yanlış pozitif ve yanlış negatif değerlerini üretmesi gibi dezavantajlarının bulunduğu belirtilmektedir. Bu eksikliklerden dolayı bu yöntemleri desteklemek için etkileşim tahminini hesapsal yöntemlerle yapan modellerin geliştirilmesine ihtiyaç duyulmaktadır. Bu amaçlar doğrultusunda araştırmacılar tarafından metotlar geliştirilmiştir. Fakat bu metotların çalışabilmesi için proteinler hakkında her durumda ulaşılamayacak bir takım özelliklere ihtiyaç duyulduğu görülmektedir. Bununla beraber geliştirilen etkileşim tahmin metotların ölçülen sınıflandırma doğruluklarının henüz yeterince yüksek değerlere ulaşamadığı belirtilmektedir.

Bu tez çalışmasında, proteinler arasında oluşan etkileşimlerin tahmin edilmesi amacıyla sekans tabanlı yöntemler üzerinde çalışılmıştır. Bu amaçla, proteinlerin birincil yapısından elde edilebilecek verileri kullanan, yüksek sınıflandırma doğruluğuna sahip etkileşim tahmin sistemleri önerilmektedir. Protein sekanslarının birincil yapısından elde edilen bilgiler kullanılarak daha etkin çalışan özellik çıkarım aşamaları geliştirilmeye çalışılmıştır. Bu özellik çıkarım yöntemlerinin kullanımıyla, eldeki etkileşim verileri daha anlamlı bir şekilde ifade edilebilmiştir. Bu veriler Destek Vektör Makineleri (DVM) tabanlı sınıflandırma sistemlerine tabi tutulmuştur.

Geliştirilen sistemlerle elde edilen etkileşim tahmin sonuçları yüksek doğruluk değerlerine ulaşmıştır. Önerilen sistemlerin sonuçları farklı ölçütlerle değerlendirilmiş ve önceki çalışmalardan daha iyi olduğu sonucuna ulaşılmıştır.

3. MATERYAL VE YÖNTEM

Aminoasitler, proteinlerin yapısı, sekans bilgisi, kimyasal bağlar, protein-protein etkileşimleri, etkileşim türleri, özellik çıkarım yöntemleri ve etkileşim tahmini için kullanılan sınıflandırma yöntemleri bu başlık altında verilmektedir. Protein konusunun geçtiği literatürdeki ilk çalışmalardan da bahsedilmiştir.

Laboratuvar ortamında yapılan deneysel çalışmalarla ve hesapsal metotların kullanımıyla ortaya çıkarılan protein bilgileri ve etkileşimleri bu amaç için kurulmuş çeşitli veri tabanlarında tutulmaktadır. Proteomik alanında çalışma yapan araştırmacıların kullanımına açık olan bu veri tabanları hakkında kapsam, veri formatı, dosya yapısı ve erişim şekli gibi bilgiler de bu bölümde incelenmektedir.

3.1. Protein Hakkında İlk Çalışmalar

Proteinler hakkında ilk temel analizler 1828 yıllarında yapılmıştır. Hollandalı bir kimyacı olan Gerardus Johannes Mulder'in yaptığı bu analizlerde incelediği moleküllere ilk olarak 1838 yılında İsveçli kimyacı Jöns Jakob Berzelius tarafından eski Yunancada birincil ve asli anlamlarına gelen πρωτεῖος (*proteios*) kelimesinden türetilen "protein" ismi verilmiştir (Braun ve Gingras, 2012).

Proteinler arasındaki etkileşimlere ilişkin yapılan ilk çalışma ise sindirim sisteminde bulunan bir enzim olan *tripsin'in* inhibisyonunu ve kinetik özelliklerini inceleyen Hedin (Hedin, 1906) tarafından raporlanmıştır. Protein etkileşimlerinin organizmalar için gereken enzim faaliyetlerinin oluşmasında hayati bir öneme sahip olduklarını kabul eden çalışmalar 1960'lı yıllarda yapılmaya başlanmıştır (Crick ve Orgel, 1964). Protein etkileşimlerinin önemi daha geç anlaşılmış ve çalışmalar maya iki-hibrit isimli protein etkileşim yönteminin geliştirilmesi ile önemli bir aşama kaydetmiştir (Fields ve Song, 1989). Bu yöntem, çok fazla sayıdaki girdi ve çıktı ile çalışıldığı için yüksek-hacimli (high-throughput) olarak isimlendirilen protein etkileşim tahmin yöntemlerinin ilki olarak geliştirilmiştir.

3.2. Proteinlerin Yapısı

Bir organizma içerisindeki temel bileşen olan proteinler, hücre içerisinde gerçekleşen tüm biyolojik sistemleri yönetirler. Proteinler hücre içerisindeki sinyal

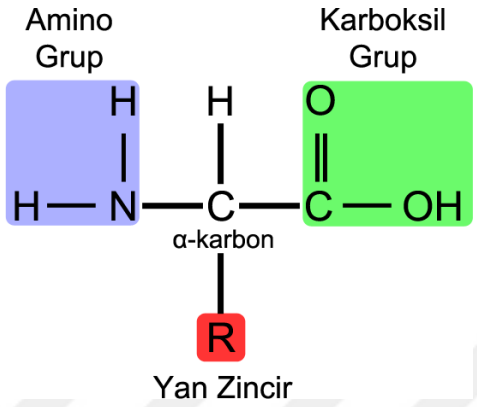
iletimi, metabolik yolların çalışması, besin alımı ve kan yoluyla iletimi, bakteri ve virüs yıkımı ile enfeksiyonlara karşı savaşma, kemik ve kıkırdak yapısının oluşumu ve hareket sisteminin çalışması, sindirim, protein biyosentezi, DNA replikasyonu ve RNA transkripsiyonu gibi tüm süreçlerde önemli rol alırlar.

Sahip oldukları biyolojik fonksiyonlara göre proteinler farklı gruplarda değerlendirilebilmektedir. *Enzimler;* hücre içerisindeki biyokimyasal tepkimeleri hızlandırmak amacıyla çalışan katalitik protein grubundadır. Örnek olarak peptit bağı hidrolizinde görev alan HIV-1 Proteaz (PDB kodu: 1DMP) ve fenilalanin yan zincirinin tirozine dönüşümünü katalize eden fenilalanin hidroksilaz (PDB kodu: 2PAH) birer enzimdir. *Hormonlar;* gelişim ve üreme gibi fizyolojik süreçlerin kontrolünü sağlayarak sistemlerin etkinliğini düzenlemekle görevlidir. Vücuttaki şeker metabolizmasının düzenlenmesini sağlayan *insülin* (PDB kodu: 3I40) düzenleyici bir diğer proteindir. Organizma içerisinde iyon, küçük molekül ya da makromoleküllerin hareket ettirilmesinde taşıyıcı proteinler görev almaktadır. Kalsiyum ve yağ asitleri gibi birçok maddenin taşınmasını sağlayan *serum albümin* (PDB kodu: 1AO6) ve oksijen taşıyan *hemoglobin* (PDB kodu: 1A3N) taşıyıcı proteinlerdendir. Diğer bir grup olan yapısal proteinler için saç ve tırnak yapısında bulunan *keratin* (PDB kodu: 3TNU) ve deri, kıkırdak ve bağ dokularında bulunan *kollajen* (PDB kodu: 1BKV) örnek verilebilir. Vücutta savunma mekanizmalarının çalışmasını sağlayan proteinler savunma proteinleri olarak adlandırılır. Antijenleri tanıyıp etkisizleştirerek organizmayı koruyan *immünglobülinler (antikorlar)* (PDB kodu: 1IGY) ve kanın pıhtılaşmasına yardımcı olan *fibrinojen* (PDB kodu: 3GHG) ve *trombin* (PDB kodu: 1AVG) bu grupta yer alan proteinlerdendir. Mikrotübülleri oluşturan *tübülin* (PDB kodu: 1TUB), kaslardaki kasılma ve gevşeme hareketlerini düzenleyen ve hücrelerin şekillerinin korunmasına yardımcı olan *miyozin* (PDB kodu: 1B7T) ve *aktin* (PDB kodu: 3HBT) kontraktil grup içinde en bilinen protein çeşitleridir.

3.2.1. Aminoasitler

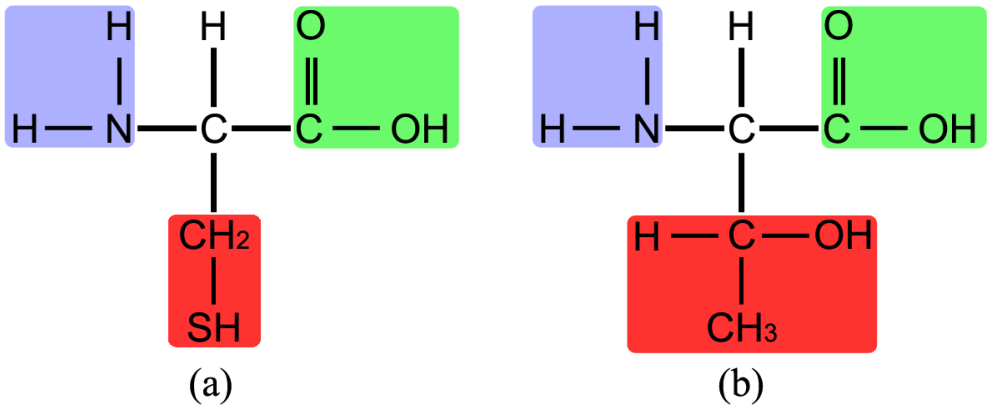
Proteinlerin temel yapı taşları aminoasitlerdir. Aminoasit çeşitlerinin belirli sayılarda ve özel bir dizilişle birbirine bağlanarak oluşturduğu zincire protein denir. Bir aminoasit ise: bir karboksil (–COOH) grubu, bir amino ($–NH_2$) grubu ve her aminoasitte farklı bir forma sahip olan organik yan zincir (R) grubunu içeren basit bir organik bileşiktir. Bir monomer olan bu molekülün merkezinde α–karbon olarak bilinen

bir karbon (C) atomu bulunur. Şekil 3.1'de görüldüğü gibi, bir amino grup ve bir karboksil grubunun bağlı olduğu α–karbon atomunun diğer iki tarafına bir hidrojen atomu (H) ve her aminoasidin birbirinden farklı olmasını sağlayan bir organik grup (R) bağlanır. Bu yan zincir grubu bir aminoasidin diğerlerinden farklı belirli kimyasal özelliklere sahip olmasını sağlar.



Şekil 3.1. Aminoasitlerin temel yapısı (merkezdeki α–karbon atomuna bağlanan kimyasal gruplar renklendirilerek gösterilmiştir).

Şekil 3.2'de görüldüğü gibi aminoasidin türünü R grubu belirlemektedir. R grubunun farklılaşmasıyla doğada 300'den fazla sayıda aminoasidin varlığından söz etmek mümkün olsa da, temel olarak protein oluşumunda görev alan 20 çeşit aminoasit bulunmaktadır.



Şekil 3.2. Aminoasitlerin kimyasal yapıları (a) Sistein, (b) Treonin.

Doğada en çok bulunan ve protein oluşumunda görev alan bu 20 aminoasidin isimleri, üç harf kısaltmaları, tek harf gösterimleri ve yan zincir bilgileri Çizelge 3.1'de gösterilmiştir.

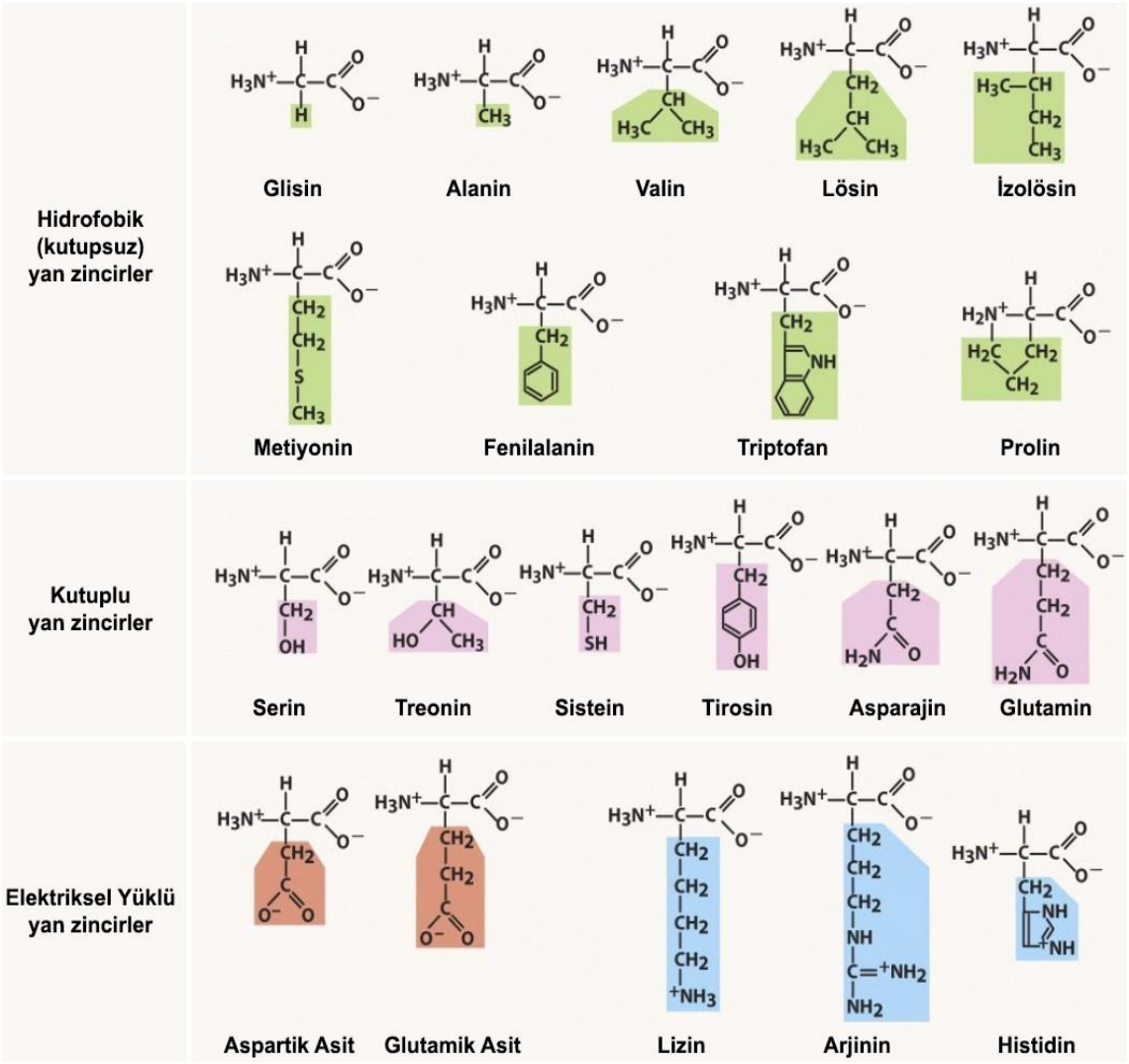
Çizelge 3.1. Protein oluşumunda görev alan 20 çeşit aminoasit listesi

Aminoasit	Üç Harf Kısaltma	Tek Harf Gösterim	Yan Zincir
Alanin	Ala	A	CH3
Arjinin	Arg	R	HN=C(NH2)-NH-(CH2)3
Asparajin	Asn	N	H2N-CO-CH2
Aspartik asit	Asp	D	HOOC-CH2
Fenilalanin	Phe	F	C6H5-CH2
Glutamin	Gln	Q	H2N-CO-(CH2)2
Glutamik asit	Glu	E	HOOC-(CH2)2
Glisin	Gly	G	H
Histidin	His	H	NH-CH=N-CH=C-CH2
İzolösin	Ile	I	CH3-CH2-CH(CH3)
Lösin	Leu	L	(CH3)2-CH-CH2
Lizin	Lys	K	H2N-(CH2)4
Metiyonin	Met	M	CH3-S-(CH2)2
Prolin	Pro	P	(CH2)3
Serin	Ser	S	HO-CH2
Sistein	Cys	C	HS-CH2
Treonin	Thr	T	CH3-CH(OH)
Triptofan	Trp	W	C6H4-NH-CH=C-CH2
Tirozin	Tyr	Y	HO-C6H4-CH2
Valin	Val	V	(CH3)2-CH

20 adet standart aminoasidin tamamının kendine has kimyasal özellikleri vardır. Bu aminoasitler yan zincirlerinin yapısı ile oluşan kimyasal özelliklere göre değerlendirildiğinde genel olarak; hidrofobik, kutuplu ve yüklü olmak üzere 3 temel gruba ayrılır (Bourne ve Weissig, 2003). Bu gruplara dâhil olan amino asitler Şekil 3.3'te gösterilmiştir (Freeman, 2005).

Hidrofobisite özelliği, aminoasidin su ile temas afinitesinin düşük olduğu anlamına gelmektedir. Kutuplu ya da yüklü aminoasitler ise suyla temas halinde olmaya eğilimli olmaktadır. Su ile temasa yatkın olma durumu hidrofilisite olarak ta bilinmektedir. Alanin, Valin, Lösin, İzolösin, Metiyonin, Prolin, Fenilalanin hidrofobik özellikte olan aminoasitlerdir. 20 genel aminoasitten birisi olan Glisin ise bir yan zincire sahip olmamasına rağmen protein içinde bulunma bölgeleri ve fonksiyonu dikkate alınarak hidrofobik aminoasit grubuna dâhil edilir.

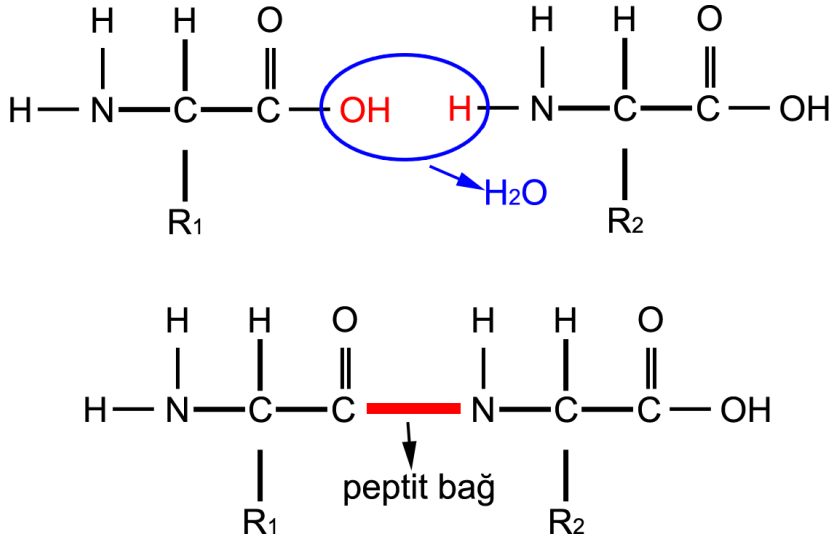
Kutuplu aminoasitler grubunda Glutamin, Asparajin, Serin, Sistein, Treonin ve Tirozin bulunurken Histidin, Lizin, Arjinin, Glutamik asit ve Aspartik asit yüklü aminoasitler grubuna dâhil edilir. Burada Histidin, Lizin ve Arjinin pozitif yüklü (bir elektron kaybetmiş) iken Glutamik asit ve Aspartik asit ise negatif yüklü (bir proton kaybetmiş) aminoasitlerdir (Freeman, 2005).



Şekil 3.3. Amino asitlerin kimyasal özelliklerine göre gruplandırılması (Freeman, 2005).

3.2.2. Amino asitlerin birleşmesi

Bir aminoasidin karboksil ucunun diğer bir aminoasidin amino ucuna bağlanmasıyla peptitler oluşur. İki aminoasit peptit bağıyla birleşirken birinin amino ucu bir hidrojen atomu (H) diğerinin karboksil ucu bir hidroksil molekülü (OH-) kaybederek bir su (H2O) molekülü açığa çıkar. Bu işlem bir kondansasyon tepkimesi ya da bir dehidrasyon sentezi şeklinde isimlendirilir. İki aminoasidin birleşerek bir peptidi oluşturma formu Şekil 3.4'te gösterildiği gibi olacaktır.



Şekil 3.4. Peptit bağı oluşumu.

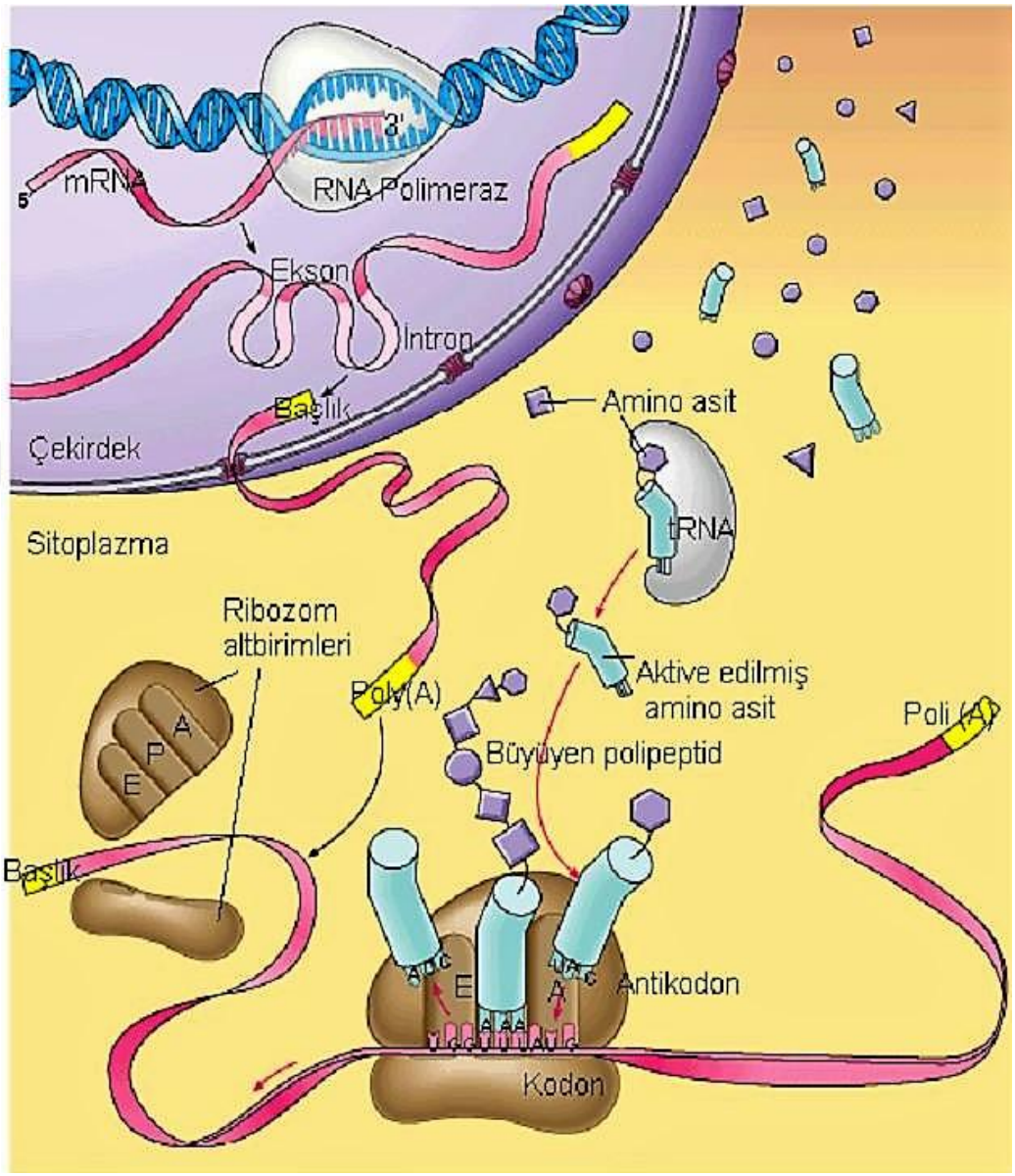
Aminoasitler bu şekilde peptit bağlarıyla birbirlerine bağlanarak polipeptit denen daha uzun zincirleri oluştururlar. Bu zincirler 20 genel aminoasidin farklı kombinasyon ve dizilişleri ile oluşturulmuş yapılardır. Aminoasitlerin polimerleri olan proteinler de bir ya da daha fazla polipeptitten oluşan yapılardır.

Polipeptitler kimyasal olarak iki farklı uca sahip olduğu için yönlü olma özelliğine sahiptir. Bir protein zincirinin amino grubuyla biten ucu *N-terminal uç* (amino terminal ucu) ve karboksil grubuyla biten ucu da *C-terminal uç* (karboksil terminal ucu) şeklinde isimlendirilir. Şekil 3.4'te oluşan en kısa zincirde sol uç N-terminal ve sağ uç ise C-terminal özelliğindedir.

3.2.3. Protein sentezi

Bir hücrede 10.000'in üzerinde farklı türde protein bulunabilmektedir (Wilhelm ve ark., 2014). Protein sentezi, hücrenin ihtiyaçlarına göre belirli proteinlerin üretilip hedef fonksiyonlara atanmalarını sağlayan biyokimyasal bir süreçtir. Bu sürecin ilk aşaması transkripsiyondur ve DNA'da bulunan genetik bilginin bir bölümünün bir mRNA (mesajcı RNA) vasıtasıyla okunması şeklinde çalışır. İkinci aşaması ise mRNA'daki bilginin bir polipeptit zincirine dönüştürülmesini sağlayan translasyon işlemidir. Translasyon işlemi ribozom moleküllerinde gerçekleşir ve tRNA'ların (transfer RNA), mRNA kodonundaki şifrelere uygun aminoasitleri taşıması ve mRNA kodonuna bağlaması şeklinde gerçekleşir.

İki aşamalı olan protein sentezi sürecinin genel akışı Şekil 3.5'te görüldüğü gibidir (Wikipedia, 2019).



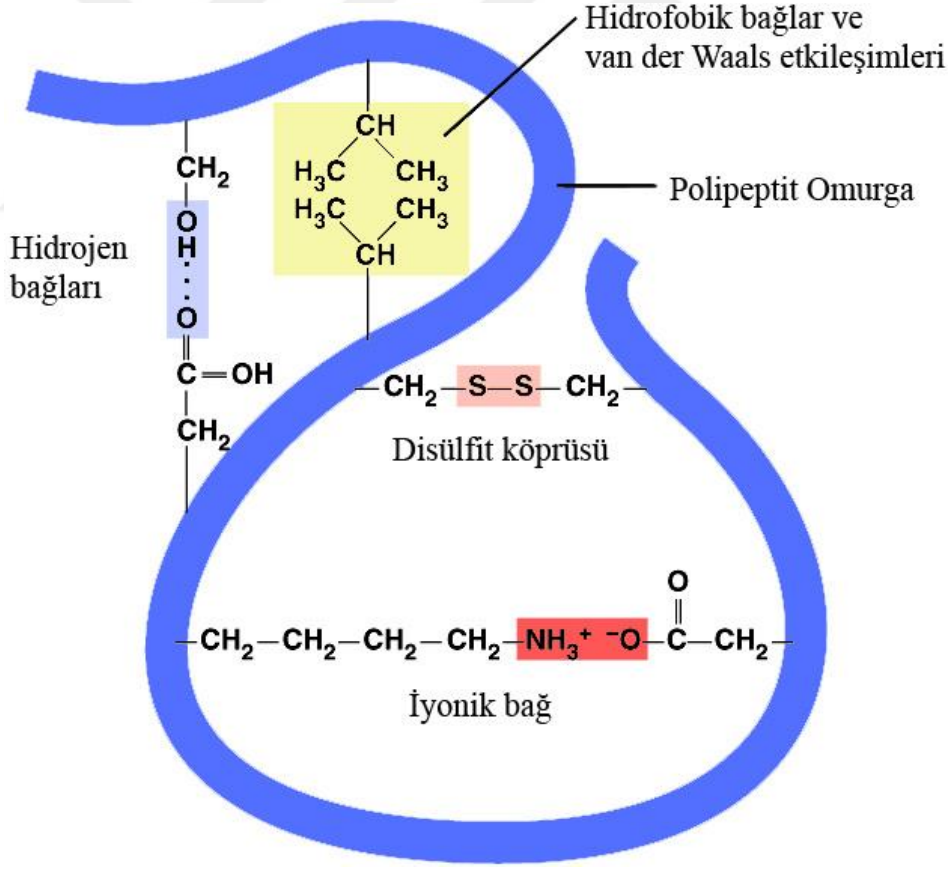
Şekil 3.5. Protein sentezlenmesi süreci (Wikipedia, 2019).

Protein sentezi çok karmaşık bir süreç olmasına rağmen araştırmacılar sentezleme hızının yaklaşık olarak prokaryotlarda saniyede 50 amino asit, ökaryotlarda ise saniyede 5 amino asit olduğunu belirtmektedir (Fedorov ve Baldwin, 1997).

3.2.4. Proteinlerin yapısında bulunan kimyasal bağlar

Proteinlerin yapılarında proteinleri bir arada tutan ya da diğer moleküllerle birleşmesini sağlayan farklı bağ türleri bulunmaktadır. Bunlar kovalent ya da kovalent olmayan güçlü ya da zayıf kimyasal bağlardır. Bu bağlar protein zincirlerinin birincil, ikincil, üçüncül ve dördüncül yapılarının kararlı hale gelmesinde rol oynamaktadırlar. Aminoasitler peptit bağlarıyla birleşerek lineer polipeptit sekansını oluştururlar. Polipeptitler belirli bir protein fonksiyonuna uygun hale gelecek şekilde katlanırken bu bağlardan yoğun olarak yardım alır.

Eğitim materyalleri ve değerlendirmesi konularında çalışan uluslararası bir firma olan Pearson (Pearson, 2018) tarafından geliştirilen gösterimin düzenlenmiş bir hali olan Şekil 3.6'da bu bağların protein sekansı üzerindeki oluşumu gösterilmektedir (Bobroff ve ark., 2016).



Şekil 3.6. Hidrojen bağı, iyonik bağ, disülfit köprüsü, Hidofobik bağ ve vander Waals etkileşimlerinin polipeptit omurgası üzerinde gösterimi (Pearson, 2018).

3.2.4.1. Peptit bağı

Protein sentezlenmesi sırasında iki aminoasit arasında gerçekleşen peptit bağı, yüksek ayrışma enerjisine sahip olan kuvvetli bir kovalent bağ türüdür. Oluşumu Şekil 3.4'te görülen ve bir aminoasidin α-karboksil karbonu ile diğer bir aminoasidin α-amino azotu arasında gerçekleşen peptit bağı, proteinin birincil yapısının kararlı hale gelmesinde önemli rol oynamaktadır.

3.2.4.2. Disülfit bağ

Protein zincirindeki iki Sistein (Cys) aminoasidi arasında gerçekleşen bağ türüdür. Sisteinlerin yan zincirindeki sülfhidril grupları (tiyol, –SH) birbirine yaklaşınca oluşan güçlü bir kovalent bağdır. Disülfit köprüsü ya da S-S bağı olarak ta bilinen bu etkileşim proteinlerin üçüncül yapılarının kararlı hale gelmesine yardımcı olarak protein molekülünün şeklinin oluşmasını sağlar.

3.2.4.3. İyonik bağ

Genel anlamda zıt kutuplu iyonlar arasında kurulan bağ türüdür. Aminoasitlerin iyonik asidik ve bazik yan zincirleri arasında oluşmaktadır. Proteinlerin üçüncül ve dördüncül yapılarının kararlı hale gelmesine yardımcı olan bağlardır. Asidik ya da bazik ortamlarda proteinlerin doğal yapısı bozulabileceği için bu bağlar ortamın pH seviyesindeki değişimlerle ya da sulu ortamlarda bozulabilecek şekilde zayıf türde bağlardır.

3.2.4.4. Hidrojen bağı

Bu bağ türü proteinlerin ikincil, üçüncül ve dördüncül yapılarının kararlı hale gelmesine yardımcı olan hidrojen bağları bir hidrojen atomu ile yakınındaki oksijen ya da azot atomları arasında gerçekleşir. Hidrojen bağları, hafif elektropozitif olan hidrojen atomunun elektronegatif bir oksijen atomu içeren grup (–C=O) ya da bir azot atomu içeren grup ($–NH_2$) tarafından çekilmesiyle hidrojen köprüsü şeklinde (C=O···H···N) oluşan zayıf bir bağ çeşididir.

3.2.4.5. Hidrofobik bağ

Hidrokarbon içerikli yan zincire sahip olan aminoasitler kutupsuz olarak değerlendirilir. Alanin, izolösin, lösin, metiyonin ve valin gibi hidrofobik aminoasitler kutupsuzdur ve suyu iterler. Hücre içerisindeki sulu bir ortamda bir protein sekansı düşünüldüğünde, zincir boyunca kutupsuz aminoasitlerin yoğun olarak bulunduğu bölgeler suyu itmek için yan yana gelecek şekilde katlanırlar. Bu şekilde kutupsuz bölgeler arasında oluşan zayıf etkileşimler meydana gelir. Bu etkileşimler hidrofobik bağlar olarak adlandırılır.

3.2.4.6. Van der waals kuvveti

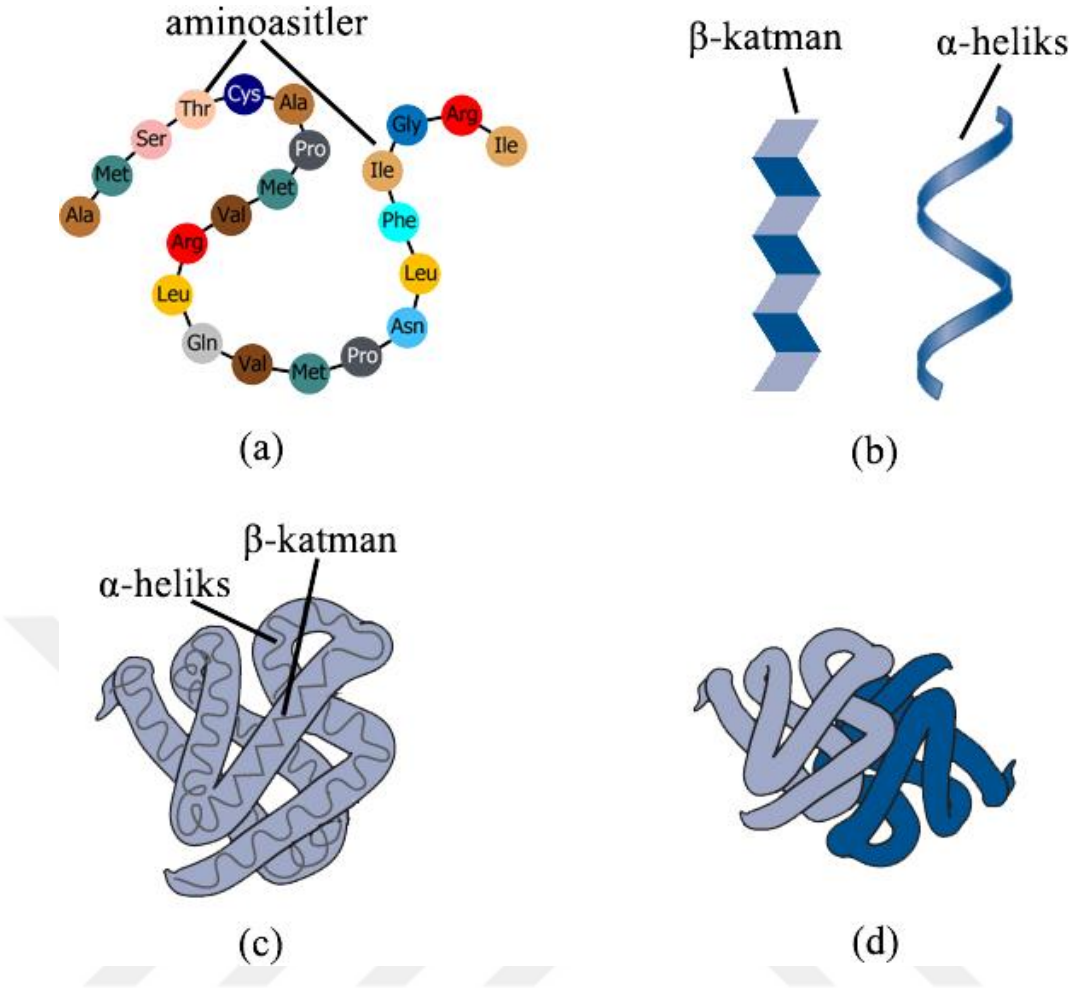
Birbirine yeterince yaklaşmış olan zıt kutuplu elektron tabakalı atomlar arasında meydana gelen çekim kuvvetidir. Bu çekim ancak belirli bir Van der Waals yarıçapı içerisinde olur. Çok zayıf olan bu kuvvet aynı anda çok sayıda gerçekleştiği için kararlı yapının korunmasında etkili olmaktadır.

3.2.5. Proteinlerin yapısal seviyeleri

Proteinlerin sahip oldukları şekilleri, yerine getirecekleri fonksiyonun gerçekleşmesi için çok önemlidir. Kendilerine has olan son şekillerine nasıl kavuştukları ise dört seviye üzerinden açıklanır. Proteinler genel olarak şu dört seviye ile gösterilirler:

- Birincil (primer),
- İkincil (sekonder),
- Üçüncül (tersiyer) ve
- Dördüncül (kuarterner) yapılar.

Protein yapısal seviyelerinin karmaşıklığı birincil yapıdan dördüncül yapıya doğru gidildikçe artmaktadır. Bu yapılar, NHGRI (Amerika Birleşik Devletleri'nde bulunan Ulusal İnsan Genomu Araştırma Enstitüsü) tarafından yapılan "protein organizasyon seviyeleri" adı çalışmanın düzenlenmiş bir formu olan Şekil 3.7'de görülmektedir (NHGRI, 2019).



Şekil 3.7. Proteinin dört yapısı (a)Birincil yapı: aminoasit zinciri, (b)İkincil yapı: α-heliks ve β-katman , (c)Üçüncül yapı, (d)Dördüncül yapı (NHGRI, 2019)

3.2.5.1. Birincil yapı

Aminoasitlerin peptit bağlarıyla birbirlerine bağlanmasıyla dizi şeklinde oluşan polipeptit zinciri proteinin birincil yapısını oluşturur. Aminoasitlerin bu doğrusal dizisi proteinin yapısını belirler. Proteinlerin ihtiva ettiği aminoasitler ve bu aminoasitlerin diziliş sırası hakkında bilgi veren en basit yapısıdır. Proteinlerin bu şekilde gösterimi protein sekansı olarak isimlendirilir.

3.2.5.2. İkincil yapı

Bu yapının oluşumunda polipeptit zincirinin omurgası en temel rolü oynar. Aminoasitlerin yan zincirinde bulunan atomlar dikkate alınmadan, ana zincirdeki atomlar arasındaki etkileşimlere bağlı olarak polipeptit içinde oluşan katlanmış yapıların

gösterimidir. Bu katlanmalar zincirdeki oksijen ve hidrojen atomları arasında oluşan hidrojen bağlarından kaynaklanır.

İkincil yapının bilinen en temel iki tipi *α-heliks* ve *β-katman* yapısıdır. *α-heliks* polipeptit zincirinin hidrojen bağlarıyla bir sarmal şeklinde sağa ya da sola kavislenmesiyle oluşur. *β-katman* yapısı ise tek bir zincir üzerinde değil bitişik zincirler arasında oluşan hidrojen bağları ile zincirin düzlemsel tabakalar halinde bir birine bağlanmasıyla oluşur. *α-heliks* ve *β-katman* yapıları proteinlerde görülen ikincil yapıların büyük çoğunluğunu oluşturur. Bunların dışında *ilmek* ve *bobin* olarak isimlendirilen düzensiz yapılar da nadiren görülmektedir.

3.2.5.3. Üçüncül yapı

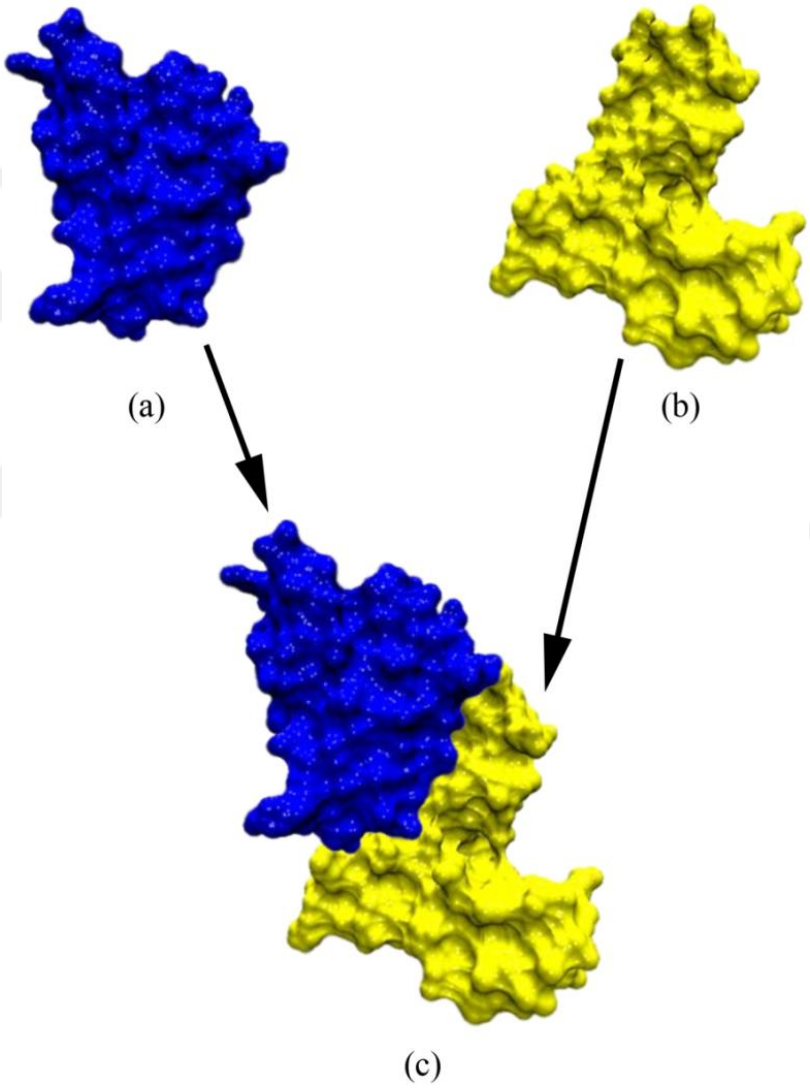
Bir protein için üçüncül yapı, polipeptit zincirinin genel üç boyutlu yapısı anlamındadır. Protein molekülü en düşük enerji ve en yüksek kararlılık durumuna geçebilmek için eğilir ve bükülür. Proteinin ikincil yapısında bulunan elemanlar doğru fonksiyona sahip olabilmesi için her zaman aynı üçüncül yapıyı oluşturacak şekilde aynı formda katlanır. Bu şekilde oluşan üç boyutlu şekil proteinin üçüncül yapısıdır. Bu yapının oluşumunda yan zincirler en önemli rolü oynar. Yan zincir atomlarının arasında oluşan etkileşimlerin etkisi vardır. Bu etkileşimler temel olarak hidrojen bağları, iyonik bağlar, hidrofobik bağlar ve disülfit bağları vasıtasıyla sağlanır.

3.2.5.4. Dördüncül yapı

Üçüncül yapı tek bir polipeptit zinciri için yapısal formu göstermesine rağmen proteinler genellikle tek bir zincir olarak işlev görmezler. Birden fazla sayıdaki üçüncül yapısına katlanmış polipeptidin birleşimi şeklinde bulunurlar. Polipeptitlerin birbirleri ile olan etkileşimleri vasıtasıyla kendi yapılarını düzenlemeleri ve bir araya gelmeleriyle dördüncül yapı oluşmaktadır. Örnek olarak bir hemoglobin molekülü dört adet globin molekülünün birleşmesiyle oluşmaktadır. Bu yapının oluşumunda da üçüncül yapıda bahsedilen bağlar rol oynamaktadır.

3.2.6. Protein-protein etkileşimi (PPE)

Bir protein molekülünün belirli bir görevi yerine getirmek amacıyla biyokimyasal olayların ya da elektrostatik güçlerin yardımıyla diğer proteinlerle yaptığı fiziksel temaslar protein-protein etkileşimi (PPE) olarak isimlendirilir. Örnek olarak, interlökin-2 ile onun alfa reseptörü arasındaki etkileşim Şekil 3.8'de gösterildiği gibidir (Chaurasia, 2014).



Şekil 3.8. İnterlökin-2 ile onun alfa reseptörü arasında oluşan etkileşim (a) İnterlökin-2 molekülü, (b) İnterlökin-2 için Alfa reseptörü ve (c) İnterlökin-2 ve alfa reseptörü ile oluşan kompleks yapı (Chaurasia, 2014).

Proteinler, hücre içerisindeki hemen hemen tüm olaylarda çok önemli rol oynarlar. Vücudun doku ve organlarının yapısı, çalışma süreçlerinin sürdürülmesi ve düzenlenmesi için gereklidirler. Görevlerini yerine getirirken tüm fonksiyonlarda

genellikle tek başlarına değil, biyolojik aktiviteye uygun şekilde diğer proteinlerle etkileşerek biyolojik süreçleri yürütürler. Bu şekilde etkileşim kurarak bir grup halinde çalışan proteinlerin oranının %80'den fazla olduğu bilinmektedir (Berggard ve ark., 2007).

Proteinler tarafından kurulan bu birleşimler biyolojik süreçler içerisinde ya da süreçler arasında bilgi aktarımı yapmak için iki ya da daha fazla protein arasında gerçekleşmektedir. Ayrıca, proteinin fonksiyonunun bilinmediği durumlarda, görev ve işlevleri bilinen diğer proteinlerle kurduğu etkileşimlere bakılarak fonksiyon belirlemesi yapılabilmektedir.

PPE'lerin görülebilecek etkilerinin çok farklı alanlarda ortaya çıkabildiğini belirten araştırmacılar PPE'lerin görülen etkilerini aşağıdaki temel gruplarda toplamıştır (Phizicky ve Fields, 1995);

- Proteinlerin kinetik özelliklerini değiştirir,
- Enzimlerle reaksiyona giren maddelere (substrat) kanal açarak alt birimler arasında hareket etmelerini sağlar,
- Genellikle küçük efektör (biyolojik aktiviteyi düzenlemek için proteine bağlanan) moleküller için yeni bir bağlanma noktası oluşturur,
- Proteinleri baskılar ya da etkisiz hale getirir,
- Proteinlerin özelliklerini, farklı bağlanma ortaklarıyla etkileşime sokarak, substratlarına göre değiştirir.

3.2.6.1. Protein domainleri

Domainler, bir protein sekansının üçüncül yapısında bulunan ve belirli bir üç boyutlu katlanmış yapıya sahip olan bölgelerdir. Protein sekansının diğer kısımlarından bağımsız olarak var olan ve işlev görebilen korunmuş bölgelerdir. Domainler genellikle belirli bir fonksiyonu yerine getirir ya da belirli bir etkileşimin oluşabilmesini sağlarlar. Etkileşimler protein zincirlerindeki bu özel domain bölgeleri arasında gerçekleşir. Proteinler diğer belirli proteinlerle bağ kurabilmek için bu tür yapısal domain bölgelerine sahiptir.

3.2.6.2. İnteraktom

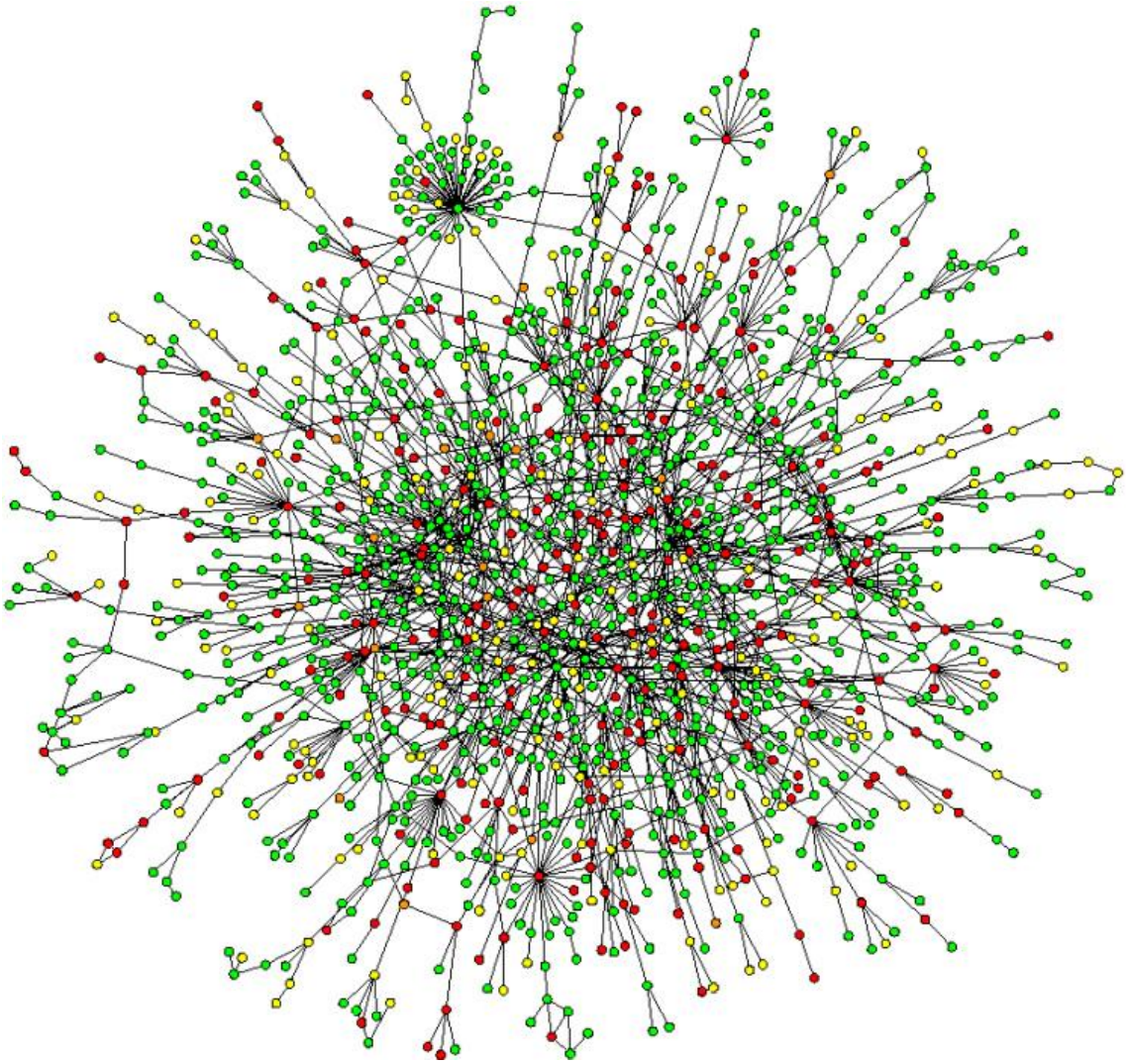
Moleküler biyoloji kapsamında belli bir biyolojik organizma ya da özel bir hücre için meydana gelen etkileşimlerin tamamına "interaktom" denilmektedir. İnteraktom, proteinler arasındaki etkileşimler gibi moleküller arasında oluşan fiziksel ilişkileri ifade eder. PPE'ler belirli bir hücre ya da organizma için tanımlanan protein kümesi ile bağlantılıdır. Belirli bir yapı için tanımlanan, ya da genomu tarafından ifade edilen bu protein kümesine "proteom" denilmektedir. Proteinlerin yapılarını inceleyip ortaya çıkarmayı amaçlayan ya da proteinlerin fonksiyonlarını keşfetmeye çalışan araştırmalara da proteom analizi ya da "proteomik" denilmektedir. Protein üretimi ve indirgenmesinin araştırılması, protein modifikasyonlarının analizi, protein hareketlerinin ve etkileşimlerinin incelenmesi bu alandaki çalışmalardandır.

3.2.6.3. PPE Ağları

Bir PPE ağı, belirli bir organizma içinde bulunan proteinlerin tamamına ilişkin etkileşimlerin, proteinlerin birer düğüm ile temsil edildiği, graf yapısı ile gösterilmesidir. Bu ağda, 2 protein arasında tespit edilen bir etkileşim düğümler arasına çizilen kenarlarla ifade edilir. Maya (Saccharomyces cerevisiae) için Jeong ve ark. (Jeong ve ark., 2001) çalışmasında verilen PPE ağının değiştirilmiş bir hali Şekil 3.9'da gösterilmiştir (Fossum, 2008).

3.2.6.4. PPE'lerin üç boyutlu yapıları

Araştırmacılar tarafından elde edilen PPE verilerinin sayısı sürekli artmaktadır. Deneysel olarak onaylanan etkileşimlere ait bilgi ve belgelerin depolanıp araştırmacıların hizmetine sunulan önemli veri tabanları bulunmaktadır. Bu veri tabanlarında varlığı kanıtlanmış etkileşimlere ait yapısal bilgiler, dokümantasyon ve etkileşimlere ait üç boyutlu gösterimler bulunmaktadır.

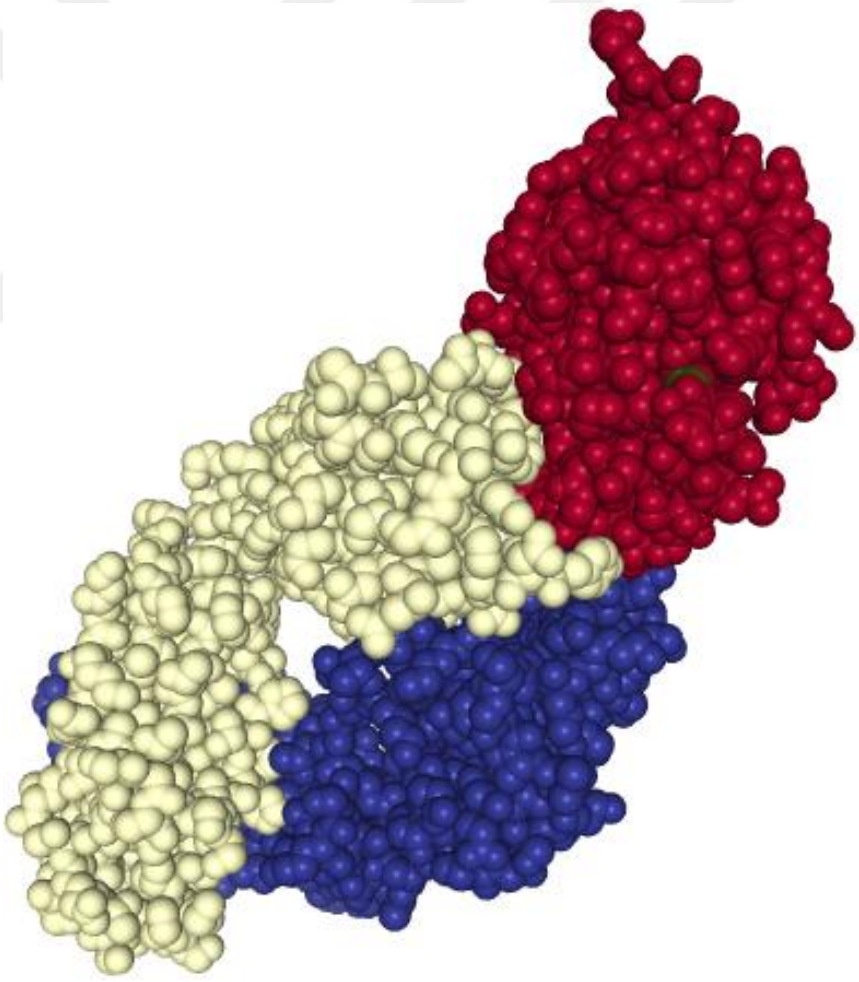


Şekil 3.9. Saccharomyces cerevisiae için PPE ağı haritası. Daireler protein düğümlerini, düğümler arası hatlar da bir PPE'yi temsil etmektedir. Renkler düğümün maya için önemini göstermektedir. Kırmızı düğümler maya için hayati önem taşırken yeşillerin önemi daha azdır. Turuncular büyüme hızını etkiler ve sarılar ise tespit edilmemiştir (Fossum, 2008).

WWPDB (Worldwide Protein Data Bank) (Berman ve ark., 2003) organizasyonunun bir üyesi olan Protein Veri Bankası (Research Collaboratory for Structural Bioinformatics Protein Data Bank - RCSB PDB) bu anlamda protein, nükleik asitler ve kompleks yapıların üç boyutlu şekilleri hakkında deneysel olarak belirlenmiş bilgileri içermektedir. RCSB PDB; moleküler, yapısal ve hesaplamalı biyoloji alanında verileri seçip organize eden ve açıklamalar getirerek araştırmacıların kullanımına sunan bir yapı üzerine kurulmuştur. 1971 yılında kurulduğunda sadece birkaç yapı hakkında bilgi içeren PDB, Nisan 2019 itibarıyla içerdiği yapısal bilgi sayısını 151,000'in üzerine çıkarmıştır. Bu verilerin büyük çoğunluğu protein yapılarına aitken, bir kısmı da DNA ve RNA nükleik asit yapılarına aittir. Bu veri tabanındaki verilerin büyük çoğunluğu *X*-

ışınları kristalografisi yöntemi ile elde edilmiştir. Verilerin elde edildiği diğer önemli iki deneysel yöntem ise *Nükleer Manyetik Rezonans Spektroskopisi (NMR)* ve *Elektron Mikroskopisi (EM)* yöntemleridir (Berman ve ark., 2000).

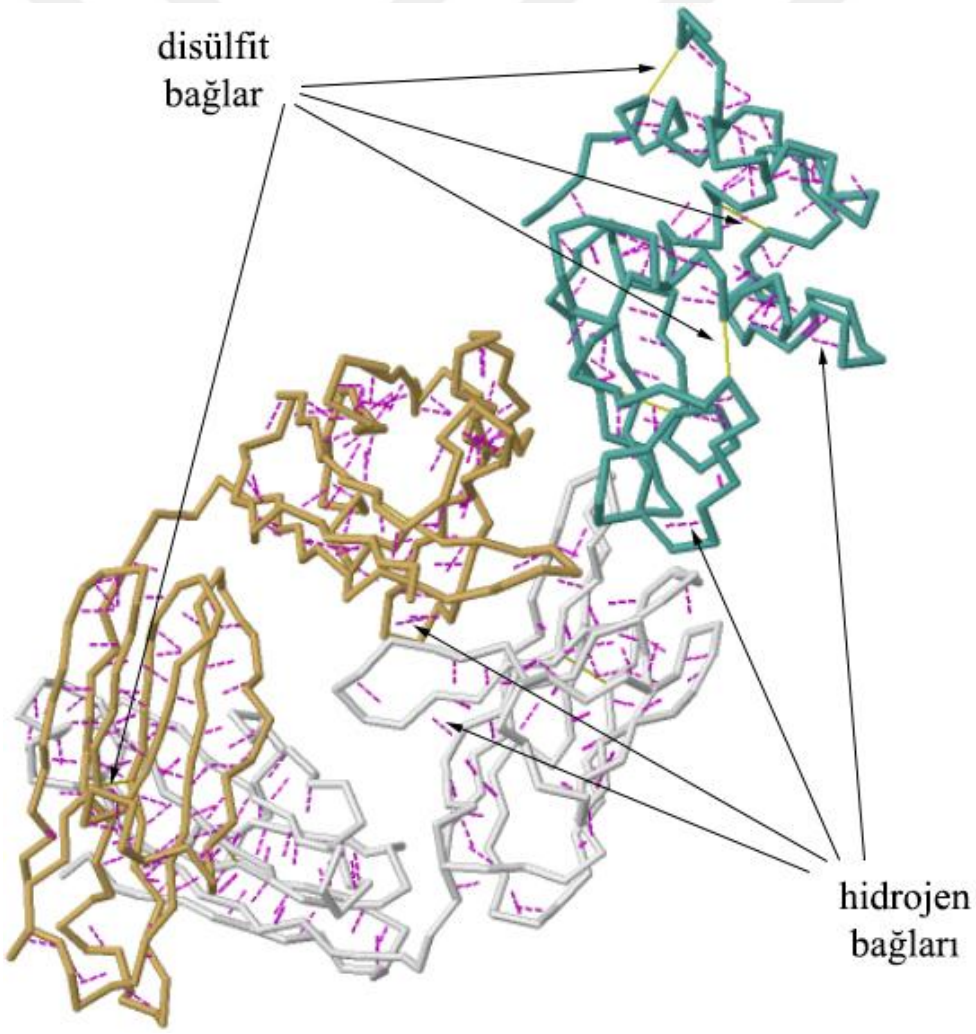
Örnek olarak Şekil 3.10'da RCSB PDB sisteminden elde edilen bir antikor-protein etkileşimine ilişkin bir komplekse ait gösterim sunulmaktadır. Bu gösterim bağışıklık sistemi ile alakalı bir protein-protein etkileşimi olup (PDB kodu: 1YQV) Hylel-5 antikoru ile tavuk yumurta beyazındaki lizozim enzimi arasında gerçekleşmektedir (Cohen ve ark., 2005). Şekilde, kompleksin zincire göre boyanmış ve boşluk doldurma uygulanmış hali görülmektedir. PDB web sayfasından (RCSBPDB, 2018) 3 boyutlu bir moleküler gösterim uygulaması olan NGL Viewer web ara yüzü kullanılarak elde edilmiştir (Rose ve ark., 2018).



Şekil 3.10. Hylel-5 antikoru ile lizozim enzimi arasında gerçekleşen etkileşimin sekans tabanlı gösterimi (PDB kodu: 1YQV) (RCSBPDB, 2018)

Moleküllerin zincir yapılarına göre renklendirilerek ve boşluk doldurma yöntemi uygulanarak elde edilen bu gösterimde beyaz kısım Hylel-5 antikorunun ağır zincir bölgesi iken mavi kısım ise hafif zincir bölgesidir. Aminoasit sayısı ve molekül ağırlığı fazla olan ağır zincir ile daha az sayıda aminoasit içeren ve molekül ağırlığı daha az olan hafif zincir bölgeleri bu molekül için antijen bağlama bölgesini oluşturmaktadır. Kırmızı kısım ise lizozim enzimidir.

Şekil 3.10'da verilen etkileşimi, içerdiği disülfit ve hidrojen bağlarıyla birlikte gösteren ikincil yapı omurga çiziminin ise yine PDB web sayfasından (RCSBPDB, 2018) elde edilmiş hali Şekil 3.11'de verilmektedir. Bu şekil, 3 boyutlu bir moleküler gösterim uygulaması olan açık kaynak kodlu Jmol java görüntüleyicisi kullanılarak (Jmol, 2018) elde edilmiştir.



Şekil 3.11. Hylel-5 antikoru ile lizozim enzimi arasında gerçekleşen etkileşimin omurga şeklinde gösterimi (PDB kodu: 1YQV) (RCSBPDB, 2018)

Bu şekildeki bir gösterim ile kompleks içinde oluşan hidrojen bağlarını ve protein zincirinin üstüne katlanması sırasında oluşan disülfit bağlarını görmek mümkün olmaktadır. Proteinin kararlı yapısının oluşumunu sağlayan disülfit bağları proteinin farklı bölgeleri arasında ya da farklı proteinler arasında 2 sistein (CYS) grubunun birbirine yaklaşmasıyla oluşan bir kovalent bağdır.

3.2.7. PPE'lerin sınıflandırılması

Proteinler farklı hücresel fonksiyonlar için çok sayıda muhtemel etkileşim ortağını tanırlar ve belirli fonksiyonel amaçlar doğrultusunda birbirleriyle etkileşirler. Bu etkileşimler farklı birleşme eğilimleriyle gerçekleşebilir. Bu nedenle PPE'ler çok geniş bir çeşitliliğe sahiptir. Bununla beraber, PPE'ler farklı faktörlere bağlı olarak bazı temel etkileşim türlerine ayrılırlar. Etkileşimlerin sınıflandırılması farklı özelliklere göre çeşitlendirilebilse de genel olarak protein kompleksinin bileşimine göre homo-oligomerik ve hetero-oligomerik kompleksler olarak, bağlanma eğilimlerine göre zorunlu ve zorunlu olmayan kompleksler olarak ve yaşam sürelerine göre de kalıcı ve geçici kompleksler olarak yapılmaktadır (Nooren ve Thornton, 2003).

3.2.7.1. Protein kompleksinin yapısına göre PPE'ler

Monomerler organik bileşiklerin yapı taşlarıdır ve diğer monomerlerle birleşerek daha büyük yapıları oluştururlar. Çok sayıdaki aminoasitten (monomer) oluşan yapılar *Oligomer* olarak isimlendirilir. PPE'ler yapılarına göre *homo-oligomer* ya da *hetero-oligomer* şeklinde iki türe ayrılır. Etkileşimlerin aynı proteinler arasında ya da farklı proteinler arasında olması şeklinde tanımlanabilir. Aynı alt birimlerden oluşan etkileşimler *homo-oligomer*, farklı türdeki alt birimlerden oluşan etkileşimler ise *hetero-oligomer* olarak sınıflandırılır.

Homo-oligomer kompleksler genellikle kararlı bir protein yapısı oluşturma eğilimindedir. Enzimler ve bu enzimlerin taşıyıcıları arasında kurulan etkileşimler *homo-oligomer* türündedir. *G-proteini bağlı reseptörler* (GPCR) ise *hetero-oligomer* etkileşim olarak tanımlanır. Hücre zarından 7 kez geçtikleri için "7 transmembran reseptörler" olarak ta bilinen GPCR'ler hücre içerisinde G proteinlerine bağlanmaktadır. Bu yapıların görevi hücre dışındaki molekülleri tespit etme ve buna bağlı olarak hücre

içine sinyal iletmektir. Bu sinyal iletimi ile hücre içinde gerekli düzenlemelerin yapılmasını sağlamaktadır.

3.2.7.2. Etkileşen birimlerin birbirlerine bağımlılığına göre PPE'ler

Proteinlerin etkileşim kurduğu eşlerinden hücre içerisinde bağımsız olarak bulunup bulunamadıklarına göre etkileşimin türü zorunlu ve zorunlu olmayan şeklinde sınıflandırılmaktadır. Zorunlu bir etkileşimde protein çiftini oluşturan birimler kendi başlarına (birbirlerinden ayrı olarak) kararlı yapılar halinde bulunmazlar. Antikor-antijen ve enzim-inhibitör gibi etkileşimleri kuran birimler kendi kararlı yapılarında da bulunabilirler.

Zorunlu olmayan etkileşim türü, aşağıda açıklandığı gibi, bağlanma eğilimlerine göre iki sınıfa ayrılabilir. Bu sınıflandırmayı yapmak zorunlu etkileşim türleri için mümkün değildir.

3.2.7.3. Bağlanma eğilimlerine göre PPE'ler

PPE'ler kararlılık durumlarına göre geçici ya da kalıcı etkileşimler olarak sınıflandırılır. Proteinlerin kalıcı şekilde bağlanması ve etkileşimden sonra proteinlerin ayrılmaması ile *kalıcı etkileşim* oluşur. Belirli bir aktiviteyi düzenlemek için birleşen ve aktivite gerçekleştikten sonra tekrar ayrılan proteinler *geçici etkileşim* yaparlar.

Geçici etkileşimler sinyal yollarının çalışmasını sağlarken kalıcı etkileşimlerin daha çok kararlı yapıdaki protein komplekslerinin oluşumunda rol aldığı söylenebilir (Rao ve ark., 2014). Hormonlar arasında kurulan ya da hemoglobin molekülü ile RNA polimeraz enzimleri arasında gerçekleşen etkileşimler kalıcı etkileşim türünde iken hücre sinyal yollarındaki çoğu protein *G-protein bağlı reseptörlerde* olduğu gibi geçici etkileşim kurarlar.

Bunula beraber, geçici etkileşimler de zayıf geçici iletişimler ve güçlü geçici iletişimler olarak iki gruba ayrılır. Zayıf geçici etkileşimde yapılar daha fazla korunmuş fakat daha az kararlıdırlar. Zayıf geçici etkileşimler çok düşük bağlanma eğilimine ve sadece birkaç saniye gibi kısa bağlı kalma sürelerine sahipken, güçlü geçici etkileşimler içinde bulundukları denge durumlarını bir molekül tetiklemesiyle değiştirirler. Zayıf geçici etkileşimler güçlü geçici etkileşimlere göre daha küçük yüzeyler arasında gerçekleşir.

3.2.7.4. Kimyasal bağ yapılarına göre PPE'ler

İçerdiği kimyasal bağ yapılarına göre etkileşimler kovalent ve kovalent olmayan olarak gruplandırılmaktadır. Disülfit bağları ya da elektron paylaşımı ile oluşan güçlü ilişkiler kovalent etkileşimleri oluşturmaktadır. Ubikitinasyon ve sumolasyon gibi posttranslasyonel modifikasyonlar kovalent etkileşimlerdir. Ubikitinasyon, ubikuitin proteini ekleyerek diğer bir proteinin indirgeme ya da yok etme amacıyla işaretlenmesidir. Bu işlem hatalı olarak katlanmış ya da hücrede kullanılmayacak olan proteinler için çalışmaktadır. Sumolasyon, SUMO grubu ubikuitin benzeri küçük düzenleyici proteinlerin bağlanması ile diğer proteinlerin fonksiyonlarının düzenlenmesi işlemidir. Kovalent olmayan etkileşimler ise çoğunlukla metabolik yollarda ve sinyal iletimi sürecinde görülen daha zayıf etkileşimlerdir. Hidrojen bağları, hidrofobik bağlar ve iyonik iletişimler gibi zayıf bağların oluşumu kovalent olmayan etkileşimlere örnek gösterilebilir.

3.3. PPE Tespit Yöntemleri

PPE'lerin tespit edilmesi; organizmalar için önemlerinin anlaşılması ve etkileşim sırasında ve sonrasında oluşan olayların incelenebilmesi için büyük önem taşımaktadır. Yüksek hacimli deneysel yöntemler kullanılarak proteinler arasındaki etkileşimlerin tahmin edilmesiyle elde edilen veri miktarının sürekli olarak artmasına rağmen bu verilerin laboratuvar teknikleriyle doğrulanmasının önünde önemli zorluklar bulunmaktadır. PPE verilerinin hesapsal analizi, deneysel yöntemlerin sahip olduğu zorlukları paylaşmadığı ve proteinlerin fonksiyonları hakkında bilgi sağladığı için yoğun olarak kullanılmaktadır.

PPE'lerin belirlenmesi için uygulanan yöntemler genel olarak üç temel tipte gruplanmaktadır. Bu gruplar; *in vivo* (canlı içinde gerçekleşen), *in vitro* (laboratuvar ortamında çalışılan) ve *in siliko* (bilgisayar ortamında kurulan) şeklinde isimlendirilir (Rao ve ark., 2014).

3.3.1. In vivo etkileşim tespit yöntemleri

PPE tespiti için yapılan işlemler *in vivo* tekniklerinde direkt olarak üzerinde çalışılan organizmanın içinde gerçekleştirilir. *Maya iki hibrit* (Y2H) (Fields ve Song, 1989), *maya üç hibrit* (Y3H) (Maruta ve ark., 2016) metotları iki protein arasında ya da bir protein ile bir DNA molekülü arasındaki olası etkileşimleri bulmaya çalışan tekniklerdir. Bu tekniklerle yapılan analizler proteinler arası oluşan etkileşimlerin direkt olarak belirlenmesini sağlar. Bu analizlerin, yoğun olarak yanlış pozitif ve yanlış negatif sonuçları üretme dezavantajı bulunmaktadır (An ve ark., 2016). *Sentetik letalite*, diğer bir *in vivo* yöntemidir. Bu yöntem; tek başlarına ortaya çıktıklarında hücre için önemli bir sorun teşkil etmeyen fakat belirli koşullar altında birleştiklerinde hücre ölümüne yol açan gen ekspresyonundaki mutasyon ve kopma gibi bozuklukları incelemektedir (Ooi ve ark., 2006).

3.3.2. In vitro etkileşim tespit yöntemleri

PPE tespiti için yapılan çalışmaların canlı organizmanın dışında, bir laboratuvar ortamında gerçekleştirildiği yöntemler *in vitro* olarak isimlendirilir. *İkili Benzerlik Saflaştırması (Puig ve ark., 2001), Nükleer Manyetik Rezonans Spektroskopisi (Wuthrich, 1989), X-ışınları kristalografisi* (Okada ve ark., 2002) ve *ko-immunopresipitasyon* (Ho ve ark., 2002) bu grupta yer alan yöntemlerdir.

İkili Benzerlik Saflaştırması (TAP) yöntemi, biyolojik komplekslerin tanımlanabilmesi için saflaştırılmasını temel alır. İlk olarak maya üzerindeki protein komplekslerinin saflaştırılmasıyla PPE'lerin tanımlanması çalışmalarıyla ortaya konan *TAP*, diğer organizmalar için de uygulanabilen bir araçtır. İmmünopresiptasyon (immün-çöktürme) temelli bir saflaştırma yöntemini kullanan *TAP*, etkileşime girdiği proteinler arasından yalnızca ilgilenilen proteinin çıkarılmasını sağlar (Puig ve ark., 2001).

Protein Mikrodizileri protein fonksiyonlarının tespiti için kullanılan yüksek hacimli bir yöntemdir. Bu yöntem ile, farklı protein moleküllerinin sıralanmasıyla oluşturulan mikrodizilerin kullanımıyla protein etkileşimlerinin belirlenmesi amaçlanmaktadır (Templin ve ark., 2002).

Diğer bir yöntem olan *Nükleer Manyetik Rezonans Spektroskopisi (NMR)* atomların çekirdeklerinin manyetik bir alana tabi tutulmasıyla oluşan durumun

izlenmesine dayanmaktadır (Wuthrich, 1989). Bu şekilde çekirdeğin etkileşimde olduğu atomlar belirlenmektedir.

X-ışınları kristalografisi (Okada ve ark., 2002) proteinlerin atomik seviyede incelenmesi yöntemidir. Protein fonksiyonlarının anlaşılmasını sağlayan bu yöntem ile proteinlerin yaptığı etkileşimler de incelenebilmektedir.

Protein etkileşimlerini analiz etmek için kullanılan bir yöntem olan *ko-immunopresipitasyon* (Ho ve ark., 2002) ise proteinlerin doğal hallerinde bulunduğu bir lizatın içinde etkileşimlerin doğrulanması mantığına dayanmaktadır.

3.3.3. In siliko etkileşim tespit yöntemleri

Etkileşim tahmini için bilgisayar ortamında yapılan çalışmalar *in siliko* grubundadır. Bu grupta geliştirilen yöntemlerin deneysel yöntemlerle elde edilen etkileşimleri destekleme özelliğine sahip olduğu söylenebilir. *Filogenetik ağaçlar* (Pazos ve Valencia, 2001), *yapı tabanlı yaklaşımlar* (Zhang ve ark., 2008), *gen komşuluğu* (Rao ve ark., 2014), *gen füzyonu* ve *sekans tabanlı yaklaşımlar* (Xia ve ark., 2010a) (Shoemaker ve Panchenko, 2007b; 2007a) *in siliko* grubundaki PPE tahmin yöntemlerinden bazılarıdır.

Yapı tabanlı yaklaşımlar; yapıları bilinen iki protein arasında tespit edilmiş olan bir etkileşimden yola çıkarak bu iki proteinin yapısına benzer proteinlerin de etkileşebileceği mantığını kullanmaktadır (Zhang ve ark., 2008). Proteinlerin çoğu için bir yapısal bilgiye sahip olmamamız bu yöntemler için bir dezavantaj olarak görülmektedir.

İşlevsel olarak benzer özelliklere sahip olan proteinlerin operonlar gibi belirli bölgelerde etkileşime geçtikleri fikrini temel alan *gen komşuluğu* yöntemleri diğer bir PPE tahmin metodu grubunu oluşturmaktadır (Rao ve ark., 2014). Proteinler arasındaki muhtemel bir etkileşimi işaret eden bu operonlar işlevsel olarak birbirine yakın olan ve tek bir birim olarak kodlanan gen topluluklarıdır (Dandekar ve ark., 1998; Shoemaker ve Panchenko, 2007a).

Biyolojik türler arasındaki evrimsel ilişkiler üzerinden yapılan etkileşim tahmin yöntemleri *Filogenetik ağaç* yöntemleri olarak bilinir. Bir filogenetik ağaç ise türler arasındaki evrimsel ilişkileri gösteren grafik yapısıdır. Etkileşen proteinlerin bu filogenetik ağaç üzerinde benzerlik göstereceği fikri öne sürülmektedir (Pellegrini ve ark., 1999).

Sekans tabanlı yaklaşımlar; protein etkileşimlerinin tahmininde birincil sekans bilgisinin kullanılması fikrini temel almaktadır (Xia ve ark., 2010a; You ve ark., 2014b). Bu yaklaşımda bir organizma için belirlenen bir etkileşimin diğer organizmalarda da oluşabileceği düşünülür. Bu fikirle sekanslar arasındaki homolojinin değerlendirilmesiyle benzer sekansa sahip proteinlerin aynı etkileşimi oluşturabileceği üzerinde çalışılır (Rao ve ark., 2014). Bu yaklaşım grubundaki bazı çalışmalar PPE'lerin oluşumunu sağlayan ve etkin alan (domain) olarak isimlendirilen sekans bölgelerinin belirlenmesini ve bu bölgeler arasındaki etkileşimleri temel alarak PPE'lerin tahmin edilmesini hedeflemektedir (Memisevic ve ark., 2013).

Gen füzyonu ya da *Rosetta Stone* metodu olarak bilinen yöntemler farklı genomlardaki tekil domain bölgelerinin çoklu domain bölgesine sahip proteini oluşturmak için birleşebileceğini dikkate alır. Rosetta stone proteini olarak bilinen ve bir protein zincirinde bu tür domain bölgelerini beraber olarak yapısında bulunduran sekansların varlığının tespit edilmesi gen füzyonu yönteminin çıkış noktasını oluşturmaktadır (Shoemaker ve Panchenko, 2007a).

3.4. Özellik Çıkarımı Yöntemleri

Özellik çıkarımı işlemi, etkileşim tahmini yapmak için geliştirilen bir metodun ilk adımı olarak görülebilir. Bu aşamanın amacı proteinlerin daha anlamlı özellik değerleriyle gösterilmesinin sağlanması olduğu için sistemin genel sınıflandırma başarısını önemli ölçüde etkilemektedir.

Tahmin sisteminin bu ilk aşamasında eldeki örneklerin verileri, sınıflandırma sisteminde kullanılabilmeleri için eşit boyutlu vektörlere dönüştürülür. Özellik çıkarımı aşamasında elde edilecek olan özellik tanımlayıcılarının etkinliği, önerilen sınıflandırma sisteminin başarısı için önem taşımaktadır (Wang ve ark., 2018).

Literatürde yoğun olarak başvurulan ve bu çalışmada önerilen etkileşim tahmin sistemlerinde de kullanılan bazı önemli özellik çıkarım metotları arasında; standart aminoasit kompozisyonu (Genfa ve ark., 1992), birleşik üçlü yöntemi (Shen ve ark., 2007), psödo aminoasit kompozisyonu (Chou, 2001) ve bi-gram gösterimleri (Sharma ve ark., 2013) sayılabilir.

3.4.1. Standart aminoasit kompozisyonu

Standart aminoasit kompozisyonu (AAC), proteinlerin sekans bilgisinden elde edilen en temel verilerden birisidir. Sekans üzerinden kolaylıkla elde edilebildiği için literatürde yoğun olarak kullanılmıştır. Standart aminoasit kompozisyonu verileri 20 aminoasidin her biri için protein sekansının içindeki frekans değerlerini gösterir (Genfa ve ark., 1992).

L uzunluklu bir $P = [R_1, R_2, R_3, R_4, \ldots, R_L]$ protein sekansı için $n_i (i = 1, 2, 3, \ldots, 20)$ değerleri sırasıyla *Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, İle, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr* ve *Val* aminoasitleri için bulunma frekansları olsun. Bu 20 aminoasit için normalize edilmiş frekans değerleri Denklem 3.1'deki formülle hesaplanır:

$$f_i = \frac{n_i}{\sum_{i=1}^{20} n_i} (i = 1,2, \ldots, 20) \tag{3.1}$$

Bu şekilde elde edilen normalize edilmiş frekanslar (*f*), *P* proteini için Denklem 3.2'de görüldüğü gibi birleştirilerek standart aminoasit kompozisyonu değerleri (*F*) oluşturulur:

$$F = [f_1, f_2, f_3, \ldots, f_{20}] \tag{3.2}$$

Veri kümesindeki her protein sekansı için elde edilen bu frekans değerlerinin toplamı 1 olacaktır.

Geleneksel aminoasit kompozisyonu olarak ta bilinen bu yöntem birçok çalışmada protein sınıf tahmini amacıyla kullanılmıştır (Bahar ve ark., 1997; Garg ve ark., 2005). Yöntemin dezavantajı ise bir protein sekansına ilişkin sadece aminoasit frekansları bilgisini sunabilmesi ve bu bilginin PPE tahmin sistemlerinde yüksek doğruluk oranları elde etmek için yeterli olmamasıdır.

3.4.2. Birleşik üçlü

Bir protein sekansındaki bir aminoasit ve onun komşu aminoasitlerinin özellik değerlerinin kullanıldığı *Birleşik Üçlü* yönteminde her aminoasit 2 komşusuyla beraber bir birim olarak değerlendirilir (Shen ve ark., 2007). Protein sekansı 20 farklı

aminoasitten oluştuğundan dolayı sekans içerisinde 20^3=8000 farklı birim elde edilebilir. Bu boyuttaki bir özellik vektörünün ise işlenmesi hesapsal yöntemler açısından karmaşık ve uzun olacağından dolayı aminoasitler gruplara ayrılarak özellik vektörünün boyutu azaltılmıştır.

Aynı özellik değerlerine sahip aminoasitlerin ve bu aminoasitlerden oluşan birleşik üçlü birimlerinin, ASD ve GTE gibi, aynı şekilde davranacağı beklenmektedir. Bu sebeple 20 farklı aminoasit dipol ve yan zincir hacim değerlerine göre 7 gruba ayrılmıştır. Sistein (C) aminoasidi disülfit bağı yapabildiği için özellikleri aynı olmasına rağmen 3 nolu gruba dâhil edilmemiştir. 20 aminoasidin özelliklerine göre ayrıldığı 7 grup Çizelge 3.2'de gösterilmektedir.

Çizelge 3.2. Birleşik üçlü yöntemi ile aminoasitlerin gruplandırılması

Grup No	Aminoasitler	Dipol Ölçüsü	Yan Zincir Hacmi
1	A, G, V	-	-
2	I, L, F, P	-	+
3	Y, M, T, S	+	+
4	H, N, Q, W	++	+
5	R, K	+++	+
6	D, E	+'+'+'	+
7	C	+	+

Bir protein sekansı üzerinden ilk aminoasitten başlayarak dizi sonuna kadar oluşan tüm üçlü birimler sayılarak 7^3=343 boyutlu bir özellik verisi matrisi elde edilir. Bu özellik matrisi Çizelge 3.3'te görüldüğü gibi 343 olası birleşik üçlü birimlerinin frekans değerlerinden oluşacaktır.

Çizelge 3.3. Birleşik üçlü yöntemi ile oluşan frekanslar

Frekans No	Birleşik Üçlüler
f_1	111
f_2	211
f_3	311
…	…
f_7	711
f_8	121
f_9	221
…	…
f_{343}	777

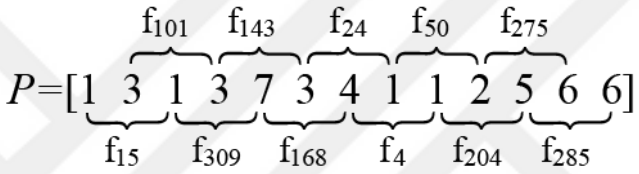
Birleşik üçlü özellik çıkarma yönteminde örnek bir protein sekansı şu şekilde gösterilir:

$$P = [G\ M\ A\ T\ C\ S\ N\ V\ A\ P\ R\ E\ D]$$

Protein örneğindeki harfler aminoasitlerin Çizelge 3.1'de verilen tek harf gösterimleridir. Bu sekanstaki tüm aminoasit tek harf gösterimlerinin Çizelge 3.2'de verilen grup numaralarıyla değiştirilerek oluşturulan yeni dizi aşağıda görüldüğü gibi olacaktır.

$$P = [1\ 3\ 1\ 3\ 7\ 3\ 4\ 1\ 1\ 2\ 5\ 6\ 6]$$

Yukarıda gösterilen örnek protein sekansı için birleşik üçlü birimlerinin frekansları Şekil 3.12'de gösterildiği gibi olacaktır.



Şekil 3.12. Birleşik üçlü frekanslarının çıkarılması

Şekil 3.12'de görülen sekans üzerinde elde edilen grup kombinasyonları ve frekans sayıları Çizelge 3.4'te görüldüğü gibi olacaktır. Bu örnek üzerinde 11 adet üçlü için frekans değerleri 1 olarak elde edilmiş olup kalan 332 adet üçlü için frekans değerleri 0 olarak kalmıştır.

Çizelge 3.4. Birleşik üçlü yöntemi ile hesaplanan frekanslar

Grup kombinasyonları	Frekans adı	Frekans değeri
411	f4	1
131	f15	1
341	f24	1
112	f50	1
313	f101	1
373	f143	1
734	f168	1
125	f204	1
256	f275	1
566	f285	1
137	f309	1

Elde edilen 343 adet frekans değeri girdi olarak kullanılmadan önce 0-1 aralığında normalize edilerek farklı uzunluklardaki proteinlere ilişkin değerlerin

oluşturacağı tutarsızlık giderilebilir. Bu amaç için Denklem 3.3'teki gibi bir normalizasyon fonksiyonu kullanılabilir:

$$d_i = \frac{f_i - min\{f_1, f_2, f_3, \dots, f_{343}\}}{max\{f_1, f_2, f_3, \dots, f_{343}\} - min\{f_1, f_2, f_3, \dots, f_{343}\}} \quad (3.3)$$

Burada *i* parametresi *(1,2,3,…,343)* değerlerini alır ve d_i tüm olası birleşik üçlü frekanslarının normalize edilmiş değerlerini verir.

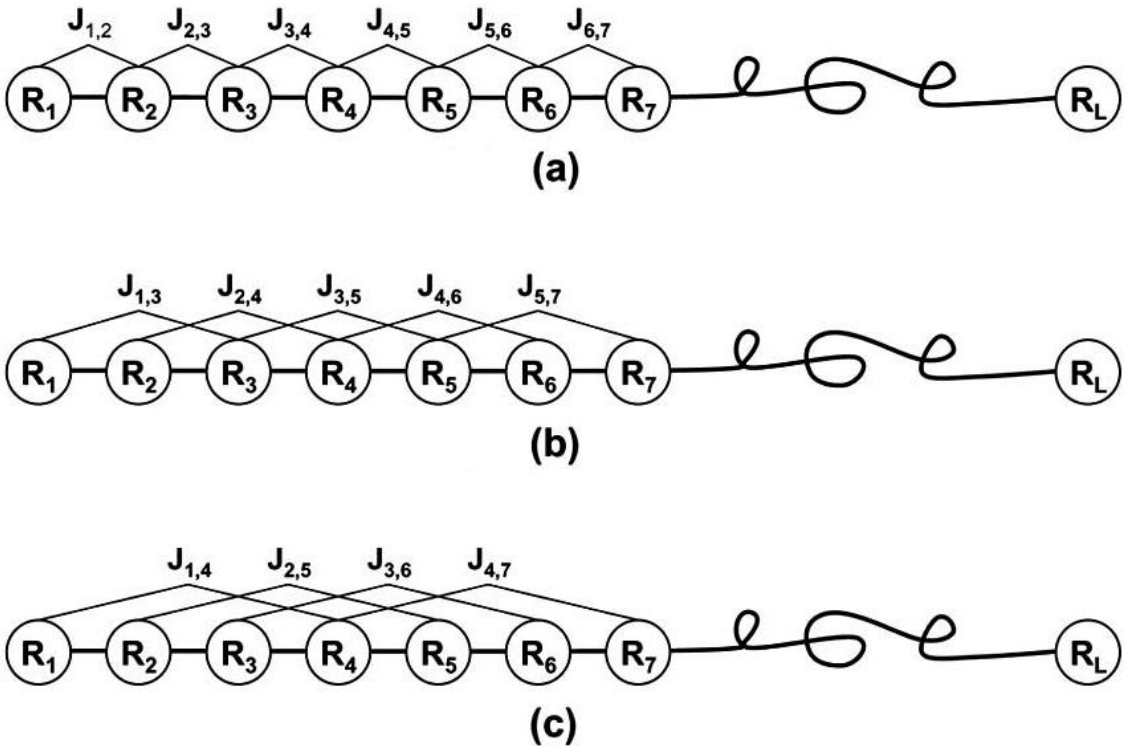
3.4.3. Chou'nun psödo aminoasit kompozisyonu

Standart aminosit kompozisyonu (AAC) özellik çıkarımı yönteminde protein sekansına ilişkin sadece 20 temel aminoasidin frekans değerlerinin elde edildiği ve sekans üzerindeki sırasal bilgilerin tamamen kaybedildiği fikri üzerine Chou tarafından (Chou, 2001) psödo aminoasit kompozisyonu yöntemi geliştirilmiştir. Bu yöntem bir protein sekansı için 20'den fazla özellik değeri üretir (20+λ). Bunların ilk 20 tanesi standart aminoasit kompozisyonu yönteminde olduğu gibi aminoasitlerin sekans içindeki frekanslarına dair bilgi verirken, kalan diğer özellikler (λ) sekans sırasına dair bilgileri gösterir.

Sekans üzerindeki aminoasit komşuluklarını göz önüne alan bu yöntemde proteinler için daha anlamlı özellikler çıkarılabilmektedir. Psödo aminoasit kompozisyonu belirlenen komşuluk kademelerine göre farklı eşleşme faktörleri elde etmektedir. Şekil 3.13, ilk 3 en yakın-komşuluk kademesi için (birinci, ikinci ve üçüncü kademeler) oluşan eşleşme faktörlerinin sekans üzerinden elde edilişini göstermektedir (Chou, 2009).

Psödo aminoasit kompozisyonu yöntemi hem protein sekansı üzerinde bulunan aminoasitlerin frekanslarını hem de sekans üzerindeki önemli sırasal bilgileri içerdiği için PPE tahmin sistemlerine daha anlamlı veriler sunmaktadır. Bu üstünlüğünden dolayı, literatürdeki birçok çalışma psödo aminoasit kompozisyonu yöntemini kullanmıştır (Chou, 2005; 2009). Protein yapı sınıfları ayrımı (Hayat ve Iqbal, 2014), protein dördüncül yapı tahmini (Zhang ve ark., 2008), antikanser peptitlerinin tahmini (Hajisharifi ve ark., 2014), bazı organizmalarda bulunan ve donma noktasını düşüren antifriz proteinlerinin tahmini (Mondal ve Pai, 2014) ve zar protein tipi tahmini (Han ve

ark., 2014) gibi proteomik alanındaki birçok çalışma psödo aminoasit kompozisyonu yöntemini kullanmıştır.



Şekil 3.13. Psödo aminoasit kompozisyonu için (a) Birinci kademe, (b) İkinci kademe ve (c) Üçüncü kademe sekans-sıra kombinasyonlarının şemasal gösterimi (Chou, 2009).

3.4.4. Konuma özel puanlama matrisi (PSSM)

Konuma özel puanlama matrisleri, sekans motiflerinin gösterimini sağlamak için kullanılan bir yöntemdir (Gribskov ve ark., 1987). Bu yöntem bir protein sekansında bulunan aminoasitler için hizalama tabloları oluşturarak korunan pozisyonları çıkarmayı sağlar. Protein sekansı için üretilen bir PSSM her bir aminoasit türü için bir kolon ve sekanstaki her aminoasit için bir satır içeren bir matristir.

Yöntemin ilk adımı olarak sekans verilerinden *konum frekans matrisleri* oluşturulur. Bu matris her bir aminoasidin her bir konumda bulunma frekanslarını içerir. *Konum frekans matrisleri* üzerinden her pozisyondaki aminoasit sayısını toplam sekans sayısına bölerek *konum olasılık matrisi* üretilir.

Bir *konuma özel puanlama matrisinin* elemanları (M), *n* elemanlı *l* uzunluklu sekanslardan oluşan bir *S* kümesi için Denklem 3.4 ve Denklem 3.5 ile hesaplanır;

$$M_{i,j} = \frac{1}{n} \sum_{k=1}^{n} I_i\,(S_{kj}) \tag{3.4}$$

$$I_i(x) = \begin{cases} 1 & i = x \\ 0 & değilse \end{cases} \tag{3.5}$$

Bu denklemlerde *i* aminoasit çeşitlerini ifade eder, *j* ise 1'den sekans uzunluğu *l*'ye kadar değer alır. $I_i(x)$ ise aranan aminoasit ile aranılan konumdaki aminoasidin karşılaştırılmasıdır.

Sonuç olarak elde edilen konuma özel puanlama matrisindeki her eleman belirli bir pozisyonda belirli bir aminoasidin gözlenme sayısını gösterir. Burada hesaplanan değerler aminoasitlerin mutlak frekanslarıdır. Ayrıca göreceli frekanslar ya da log-benzerlik değerleri de üretilebilir. Göreceli frekansların hesaplanması Denklem 3.6 ile yapılır.

$$f_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{20} n_{i,j}} \tag{3.6}$$

Burada *i* değişkeni 1'den aminoasit sayısı olan 20'ye kadar artmak suretiyle $n_{i,j}$, *i.* aminoasidin *j* pozisyonunda görülme sayısı, $f_{i,j}$, *i.* aminoasidin *j* pozisyonundaki göreceli frekansıdır. Bu yöntemle sekans içindeki aminoasitlere ait sekans motiflerine ilişkin bilgilerin ortaya çıkarılması sağlanır.

3.4.5. Bi-Gram gösterimleri

Bir *n-gram* gösterimi, verilen bir dizi üzerindeki ardışık durumdaki *n* adet elemanın frekansının bulunmasını içeren bir yaklaşımdır. *Bi-gram* daha genel boyutta *n-gram* olarak isimlendirilen gösterim yönteminin iki eleman için çalışan türüdür. Bir dizi üzerindeki iki ardışık elemanı beraber değerlendiren bir yöntem olan bi-gram hesapsal olarak sadedir ve ölçeklendirilebilir. Dil modelleri (Moore ve Lewis, 2010), konuşma tanıma (Hirsimaki ve ark., 2009), güvenlik ve şifreleme (Jiang ve Samanthula, 2011) ve biyoinformatik (Sharma ve ark., 2013) başta olmak üzere birçok alanda kullanılan bir yaklaşımdır.

Bi-gram yöntemi ile 20 adet aminoasit için 20^2 adet özellik değeri elde edilir. Bir çift aminoasidin ardışık olarak sekans içinde görülme sayıları olan bu değerler Denklem 3.7 ile elde edilir.

$$BG(m) = \left(\frac{f(i,j)}{n}\right), \qquad i,j \in [1,2,\ldots,20] \tag{3.7}$$

Burada *i* ve *j*, 20 aminoasidi işaret etmek üzere *f(i,j)* belirli bir aminoasit çiftinin sekans içinde görülme frekansını verir. Sekans uzunluğu *n* ile gösterilir. Üretilen *bi-gram* elemanının sıra numarası olan *m* değişkeni ise seçilen aminoasit numaralarına (*i* ve *j*) bağlı olarak Denklem 3.8 kullanılarak değer alır.

$$m = 20 * (i-1) + j \tag{3.8}$$

Sonuç olarak, bu yöntemle üretilen *bi-gram* elemanlarını içeren vektör 400 elemanlı olacaktır. Bir *P* = [MARECRVARGCRG] protein sekansı için elde edilecek *bi-gram* özellikleri Çizelge 3.5'te gösterilmektedir. Bu protein için çizelgede verilenlerin dışındaki tüm bi-gram elemanlarının frekans değeri 0 kalacaktır.

Çizelge 3.5. Örnek bir protein için çıkarılan bi-gram özellikleri.

Aminoasit Çifti	Bi-gram Eleman Numarası	Bi-gram Elemanı Frekans değeri
AR	BG_2	2
RE	BG_{27}	1
RG	BG_{28}	2
RV	BG_{40}	1
EC	BG_{136}	1
GC	BG_{156}	1
MA	BG_{241}	1
CR	BG_{302}	2
VA	BG_{381}	1

3.5. Temel Bileşen Analizi

İlk olarak Pearson (Pearson, 1901) tarafından geliştirilen bir yöntem olan *Temel Bileşen Analizinin* günümüzde kullanılan formu, "Temel Bileşen" isimlendirmesini de ilk olarak kullanan Hotelling (Hotelling, 1933) tarafından geliştirilmiştir. Bu yöntemle, veri kümesinden yeni bilgiler çıkartılır ve Temel Bileşenler olarak isimlendirilen yeni değişkenlerle ifade edilir.

Temel Bileşen Analizi, genellikle ilişkilendirilmiş değişkenlerden oluşan büyük bir veri kümesini, ilk halindeki bilgilerin çoğunu temsil etmeye devam edecek şekilde, birbiriyle ilişkili olmayan değişkenlerden oluşan daha küçük bir veri kümesi haline dönüştürmek amacıyla kullanılan bir yöntemdir. Boyut azaltma ya da veri sıkıştırma temelli olan bu yöntem sayesinde verileri araştırma ve temsil etme karmaşıklığı önemli ölçüde azaltılmaktadır.

PCA metodu genel anlamda aşağıda belirtilen adımların yerine getirilmesiyle çalışmaktadır.

- *n* boyutlu bir *X* veri kümesi için, kümeyi oluşturan değerlerin normalizasyonunun yapılması
- *X* veri kümesi için bir kovaryans matrisinin hesaplanması,
- Özvektörlerin ve onlara karşılık gelen özdeğerlerin bulunması,
- Özvektörlerin sıralanması ve yeni *k* adet boyutu oluşturacak ilk *k* adet özvektörün seçilmesi,
- Veri kümesinin *k* boyuta dönüştürülmesi,

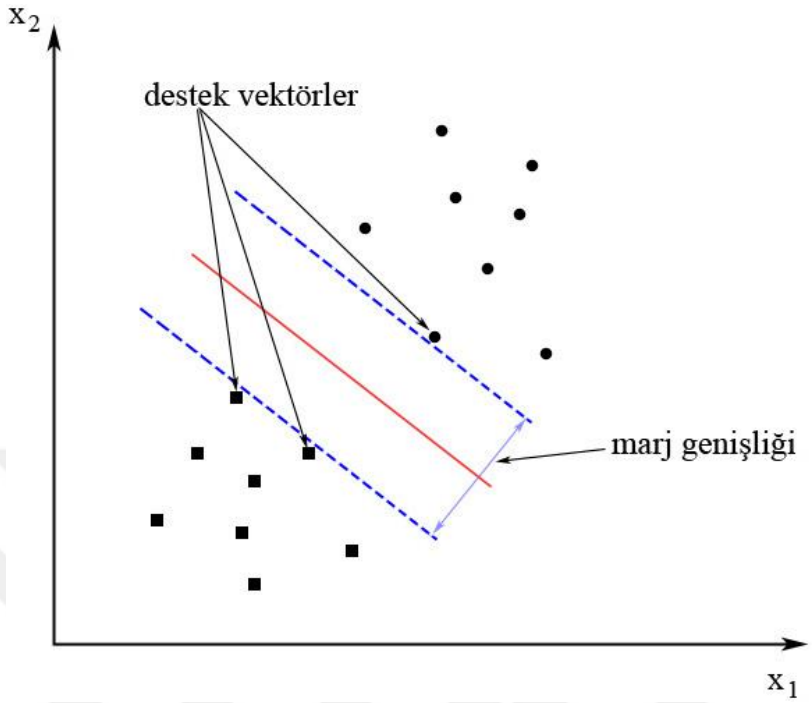
Hesapsal PPE tahmini ve proteinlerin yapısal analizi gibi konulardaki çalışmalarda sıkça kullanılan bir yöntemdir (Smith ve Sternberg, 2002; Du ve ark., 2006; Li ve ark., 2009; You ve ark., 2013).

3.6. Destek Vektör Makineleri Sınıflandırma Yöntemi

Danışmanlı bir makine öğrenmesi metodu olan DVM ilk olarak Vladimir Vapnik tarafından önerilmiş olup, standartlarının belirlenmesi ve bu günkü yumuşak marj formunun tanıtılması ise Vladimir Vapnik ve Corinna Cortes tarafından yapılmıştır (Cortes ve Vapnik, 1995; Vapnik, 2013). DVM yöntemi regresyon ve sınıflandırma problemleri için sıkça kullanılmaktadır. Örüntü tanıma (Pontil ve Verri, 1998), yazı sınıflandırma (Joachims, 1998) ve biyoinformatik (Brown ve ark., 2000) gibi birçok alanda yoğun olarak kullanılan bir yöntemdir. Değerleri arasında bir örüntü bulunmayan veri kümeleri için geliştirilecek sınıflandırma algoritmalarında kullanılabilecek bir yöntemdir.

DVM, sınıfı belli olan örneklerden oluşan bir veri kümesini en iyi şekilde farklı sınıflara ayıran en iyi ya da en uygun hiperdüzlemi bulma mantığı ile çalışır. Bu şekilde yeni örnekler de bu hiperdüzleme göre sınıflandırılır. Bu hiperdüzlemi bulurken Şekil

3.14'te görüldüğü gibi sınıflar arasındaki marj genişliğinin mümkün olan en büyük değere sahip olmasını amaçlar.

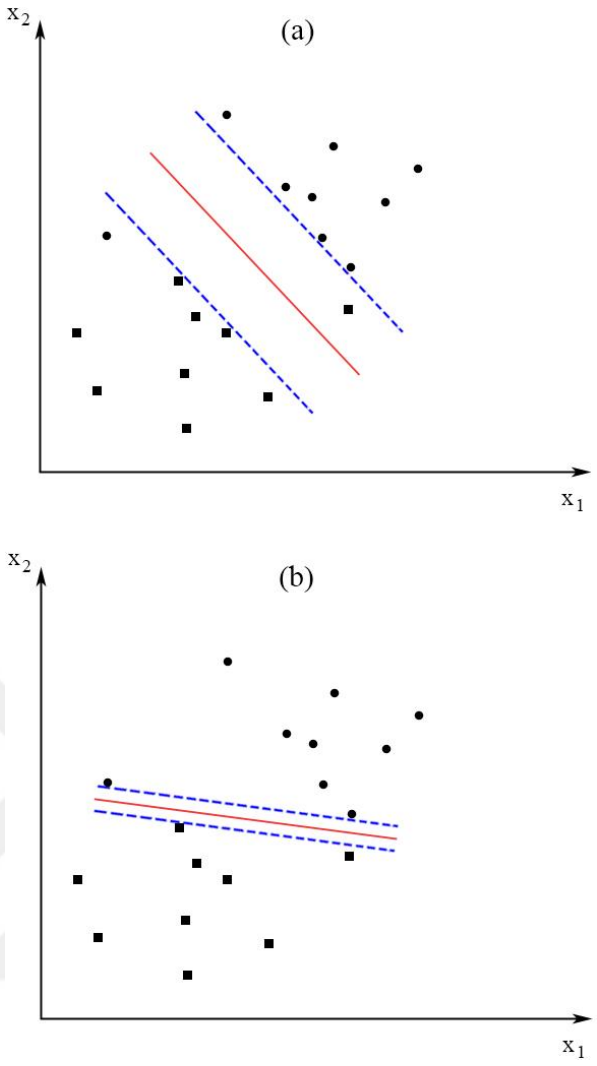


Şekil 3.14. DVM (hiperdüzlem ile örnekleri iki sınıfa ayırma)

DVM yönteminde hiperdüzleme en yakın olan örnekler veri kümesinin en önemli değerleridir. *Destek Vektör* olarak isimlendirilen bu örnekler hiperdüzlemin konumunu belirler. Öyle ki, bu örneklerin kaldırılması ya da yerinin değişmesi hiperdüzlemin de değişmesine sebep olur. Veri kümesindeki örnekler yerleştirildiğinde hiperdüzleme olan mesafelerinin fazla olması istenir çünkü bu mesafe ne kadar geniş olursa ilgili örneğe atanan sınıf etiketinin doğru olma ihtimali de o kadar yüksek olur.

DVM ile kurulan bir modelin performansı model parametrelerinin ayarlanması ile artırılabilir. Bunun için parametre optimizasyonları yapılmalıdır. DVM modellerinin en önemli üç parametresi; çekirdek fonksiyonu, *C* ve *gama* (γ) değerleridir.

Şekil 3.15'te aynı veri kümesi için (a) küçük bir C parametresi seçilerek geniş bir marj ve 2 adet hatalı sınıflandırma elde edildiği, ve (b) büyük bir C parametresi seçilerek daha dar bir marj aralığının oluştuğu ve hatalı sınıflandırmaların ortadan kalktığı gösterilmektedir.



Şekil 3.15. C parametresinin değişimiyle elde edilen farklı marjlar (a) düşük C geniş marj aralığı ve (b) yüksek C dar marj aralığı

DVM ile geliştirilen bir metot bir yandan mümkün olduğu kadar geniş bir marj genişliği elde etmeye çalışırken diğer yandan da marjın genişliğinden doğan yanlış sınıflandırma hatasının oranını en az seviyede tutmaya çalışır. Çalışılan bir metotta bu iki en iyi durumu aynı anda elde etmek mümkün olmayacaktır. Birbirine zıt olan bu iki durum arasındaki seçim, yumuşak marj fonksiyonu için hata maliyeti olarak ta bilinen *C* parametresi ile belirlenir. *C* parametresi için seçilen küçük değerler büyük bir marj genişliğine imkân verip hata beklentisini artırırken, daha büyük değerler ise küçük marj genişliklerinin oluşmasına sebep olup hata beklentisini azaltır. Cortes ve Vapnik karar kurallarının karmaşıklığı ile hata oranı arasındaki bu ödünleşim *C* parametresinin değerini değiştirmekle yönetilebileceğini söyler (Cortes ve Vapnik, 1995).

Çekirdek parametresi, örnekler için belirlenen özellik uzayını, en büyük marj genişliğini verecek şeklide bir hiperdüzlemi elde edecek şekilde yeni bir özellik uzayına

dönüştürmek için kullanılan bir fonksiyondur. Çekirdek metodu genellikle doğrusal, polinom, sigmoid ya da radyal tabanlı fonksiyon kullanılarak oluşturulmaktadır (Ben-Hur ve Weston, 2010). Gama parametresi ise radyal tabanlı bir gauss çekirdek fonksiyonu seçildiğinde belirlenmesi gereken bir değerdir. Gama değeri ile bir veri kümesindeki bir örneğin örnek uzayındaki etki alanı belirlenir. Bu şekilde, uzaydaki iki örnek arasındaki benzerlik oranının nasıl belirleneceği hakkında bilgi verir. Çekirdek fonksiyonuna verilen düşük bir gama değeri ile uzaydaki birbirine uzak olan noktalar benzer olarak değerlendirilirken, daha büyük gama değerleri ile sadece birbirine çok yakın olan noktalar benzer kabul edilir. Bu yüzden geliştirilen bir model, çok küçük bir gama değerine sahipse, seçilen bir destek vektörünün etki bölgesi geniş bir veri seti bölgesini içereceğinden, veri kümesini temsil edemez (Ben-Hur ve Weston, 2010). Bu şekilde, gama parametresi elde edilecek hiperdüzlem sınırının eğriliğini belirler.

Sonuç olarak DVM yöntemi, etkileşim yapan ve yapmayan protein çiftleri arasında ayrım yapmak için kullanılan güçlü bir sınıflandırma algoritmasıdır.

3.7. Performans Değerlendirme Ölçütleri

Bir veri kümesini sınıflandırmak için kullanılan yöntemlerin sınıflandırma başarılarının ne kadar iyi olduğunu ölçmek için kullanılan farklı değerlendirme ölçütleri bulunmaktadır. En sık kullanılan sınıflayıcı değerlendirme ölçütleri arasında *Doğruluk* (Accuracy), *Hassaslık* (Sensitivity), *Özgüllük* (Specificity), *Kesinlik* (Precision), Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient) ve alıcı işletim karakteristiği (ROC) eğrisi altındaki alan (AUC) bulunmaktadır. Bir sınıflayıcı sistem için bu değerlendirme ölçütlerinin hesaplanması dört temel oranın bulunmasıyla yapılmaktadır. İki sınıflı bir veri kümesi için belirlenen bir sınıfın pozitif olarak etiketlendiği düşünülürse diğer sınıf negatif etiketlenecektir. Sınıflayıcının elde edeceği sonuçlar veri kümesindeki bu etiketlerle karşılaştırılarak Çizelge 3.6'da görülen ve hata matrisi ya da olasılık matrisi olarak ta bilinen değerlendirme ölçütlerinin dört temel oranı hesaplanır (Zhu ve ark., 2010).

Çizelge 3.6. Hata matrisi ile dört temel değerlendirme oranının bulunması (DP: doğru pozitif, DN: doğru negatif, YP: yanlış pozitif, YN: yanlış negatif)

		Tahmin Edilen Sınıf		
		Pozitif	Negatif	Satır Toplamı
Gözlenen Sınıf	Pozitif	Doğru Pozitif	Yanlış Negatif	Pozitif Sınıflı Örnekler
	Negatif	Yanlış Pozitif	Doğru Negatif	Negatif Sınıflı Örnekler
	Sütun Toplamı	Pozitif Tahmin Edilen Örnekler	Negatif Tahmin Edilen Örnekler	Toplam Örnek Sayısı

PPE tahmin sistemlerinde gerçekte etkileşim yapan bir protein çifti için sınıflayıcı sistem "etkileşim yapar" tahmini yapıyorsa bu tahmin "Doğru Pozitif" (DP) olarak, sistem "etkileşim yapmaz" tahmini üretiyorsa "Yanlış Negatif" (YN) olarak değerlendirilir. Gerçekte iletişim yapmayan bir protein çifti için ise sınıflayıcı sistemin üreteceği "etkileşim yapar" tahmini "Yanlış Pozitif" (YP) ve "etkileşim yapmaz" tahmini de "Doğru Negatif" (DN) olarak kabul edilir. Bu oranlar kullanılarak hesaplanan değerlendirme ölçütlerinden ilki ve en sık kullanılanı olan *Doğruluk* (Acc) ölçütü sınıflayıcının sınıflandırma doğruluğu olup yapılan doğru tahmin sayısının toplam tahmin sayısına oranı olarak ifade edilir ve Denklem 3.9 ile hesaplanır.

$$Acc = \frac{(DP + DN)}{(DP + DN + YP + YN)} \tag{3.9}$$

Doğru pozitif oranı olarak ta bilinen *Hassaslık* ölçütü doğru pozitif tahmin sayısının tüm pozitif örnek sayısına oranı şeklinde ifade edilir. Pozitif örnekleri tahmin etme başarısını ifade eden bu ölçüt Denklem 3.10 ile bulunur.

$$Sen = \frac{DP}{(DP + YN)} \tag{3.10}$$

Doğru negatif oranı olarak ta bilinen *Özgüllük* ölçütü yanlış pozitif tahmin sayısının tüm negatif örneklere oranıdır ve sınıflayıcının negatif örnekleri tahmin etme başarısını gösterir. Denklem 3.11 ile hesaplanır.

$$Spe = \frac{DN}{(DN + YP)} \tag{3.11}$$

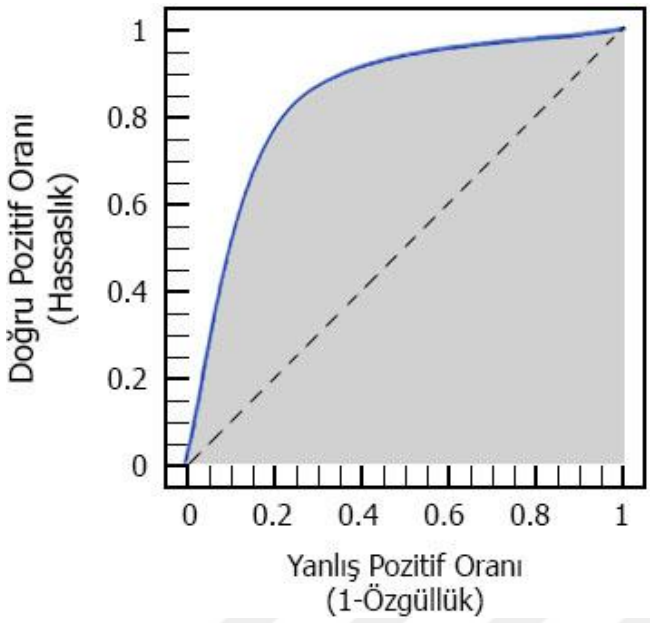
Kesinlik ölçütü yapılan pozitif tahminlerin ne kadarının doğru olduğunu gösteren bir orandır ve Denklem 3.12 ile ifade edilir. Pozitif tahmin değeri olarak ta bilinen ölçüttür.

$$Pre = \frac{DP}{(DP + YP)} \tag{3.12}$$

Pozitif örneklerin tüm örneklere oranı olarak ifade edilen *Yaygınlık* değeri bilinirse *Doğruluk* ölçütünün değeri (*Hassaslık*)*(*Yaygınlık*)+(*Özgüllük*)*(*1-Yaygınlık*) şeklinde de bulunabilir. Diğer bir değerlendirme ölçütü olan *Matthews korelasyon katsayısı* veri kümesindeki sınıflara ait örnek sayıları arasında fark olan dengesiz veri kümelerinde de uygulanabilmektedir. İlk olarak protein ikincil yapılarının tahmininde performans değerlendirme amacıyla kullanılarak B. W. Matthews tarafından önerilmiş olan ölçüte biyoinformatik konulu çalışmalarda yoğun olarak başvurulmaktadır (Boughorbel ve ark., 2017). Ölçüt, olasılık matrisi üzerinden Denklem 3.13 ile hesaplanabilmektedir.

$$Mcc = \frac{DP * DN - YP * YN}{\sqrt{(DP + FP)(DP + YN)(DN + YP)(DN + YN)}} \tag{3.13}$$

Sınıflandırma sistemlerinin performansını ölçmek için kullanılan diğer bir değerlendirme yöntemi de ROC (receiver operating characteristics curve) analizidir. Makine öğrenmesi ve veri madenciliği alanındaki çalışmalarda yoğun olarak kullanılmaktadır. ROC eğrisi, iki sınıflı bir sınıflandırma modeli için, doğru pozitif oranına (DPO-*hassaslık*) karşı yanlış pozitif oranının (YPO) çizimi yapılarak elde edilir. Yanlış pozitif oranı, negatif örneklere yapılan pozitif tahminlerin oranı olarak hesaplanır ve yanlış uyarı olarak ta bilinir. Yanlış pozitif oranı (*1-özgüllük*) değeriyle de bulunabilir. Şekil 3.16'da görüldüğü gibi ROC eğrisi, y eksenindeki doğru pozitif oranının x eksenindeki yanlış pozitif oranına olan fonksiyonu şeklinde farklı karar eşik noktalarındaki çizimlerinden oluşmaktadır.



Şekil 3.16. ROC uzayı, ROC eğrisi ve eğri altındaki alan

Burada bir DPO (doğru pozitif oranı) değeri ile ona karşılık gelen bir YPO (yanlış pozitif oranı) değeri ROC uzayındaki bir noktayı gösterir. Farklı karar eşikleri için hesaplanan noktalar ROC eğrisini oluşturur. Sol üst köşe %100 hassaslık ve %100 özgüllük oranına tekabül ettiği için, ROC eğrisi bu köşeye ne kadar yakın olursa uygulanan modelin sınıflandırma başarısı o kadar yüksek olacaktır. Çizimde kesikli çizgilerle belirtilen köşegen ise hassaslık ve özgüllük değerlerinin %50 olduğu noktalardır ve rastgele sınıflandırmaya karşılık gelir. ROC eğrisi altında kalan alan (AUC) sınıflandırma sisteminin doğruluğunu ölçmek için kullanılır ve bu alan Şekil 3.16'te taralı olarak gösterilmiştir. Bir veri örneğinin hangi sınıfa ait olduğunun ne kadar doğru olarak sınıflandırılacağını gösteren bu alan ne kadar büyükse doğruluk ta o kadar yüksek olacaktır. AUC, dengelenmemiş veri tabanlarında kullanım için de uygun olan bir değerlendirme yöntemidir.

3.8. PPE Veri Tabanları

Son yıllarda çalışmalarla önemli miktarda protein etkileşim bilgisine ulaşılmıştır. Araştırmacıların farklı deneysel yöntemlerle ya da hesapsal metotlarla ortaya çıkardığı protein etkileşim verileri üniversite ve enstitülerce yönetilen veri tabanlarına eklenmektedir. Bu amaçla oluşturulan 100'ün üzerinde veri tabanı olmasına rağmen verilerin tamamına yakını en yaygın kullanılan birkaç veri tabanında bulunmaktadır.

2000'li yıllarda kurulan DIP (Etkileşen proteinler veri tabanı) ve BIND (Biyomoleküler etkileşim ağı veri tabanı) gibi bu kapsamdaki veri tabanları genellikle araştırmacıların kullanımına açıktır.

3.8.1. DIP

DIP (The Database of Interacting Proteins) araştırmacılar tarafından deneysel çalışmalarla elde edilen protein etkileşimlerini içeren bir veri tabanıdır (Xenarios ve ark., 2000). Bu veri tabanının amacı PPE'ler hakkında bilgileri toplayıp bütünleştirmek ve protein fonksiyonları ve etkileşimleri gibi çalışma konularına katkı sağlamak için bir kullanıcı arayüzü ile sunmaktır.

DIP veri tabanında bulunan etkileşim verileri PSI-MI (1.0 ve 2.0 sürümleri), MITAB2.5 ve DIP tarafından üretilen XIN dosya formatlarında saklanmaktadır. Etkileşimleri oluşturan protein sekansları ise FASTA formatında sağlanmaktadır. FASTA; protein veri tabanlarında sık kullanılan bir formattır. FASTA formatındaki *DIP-81N* kodlu protein sekansı örneği Şekil 3.17'de görülmektedir.

```
>dip:DIP-81N|refseq:NP_014741|uniprot:P20676
MSSNTSSVMSSPRVEKRSFSSTLKSFFTNPNKKRPSSKKVFSSNLSYANHLEESDVEDTLHVNKR
KRVSGTSQHSDSLTQNNNNAPIIIYGTENTERPPLLPILPIQRLRLLREKQRVRNMRELGLIQSTE
FPSITSSVILGSQSKSDEGGSYLCTSSTPSPIKNGSCTRQLAGKSGEDTNVGLPILKSLKNRSNRKR
FHSQSKGTVWSANFEYDLSEYDAIQKKDNKDKEGNAGGDQKTSENRNNIKSSISNGNLATGP
NLTSEIEDLRADINSNRLSNPQKNLLLKGPASTVAKTAPIQESFVPNSERSGTPTLKKNIEPKKDK
ESIVLPTVGFDFIKDNETPSKKTSPKATSSAGAVFKSSVEMGKTDKSTKTAEAPTLSFNFSQKAN
KTKAVDNTVPSTTLFNFGGKSDTVTSASQPFKFGKTSEKSENHTESDAPPKSTAPIFSFGKQEEN
GDEGDDENEPKRKRRLPVSEDTNTKPLFDFGKTGDQKETKKGESEKDASGKPSFVFGASDKQA
EGTPLFTFGKKADVTSNIDSSAQFTFGKAATAKETHTKPSETPATIVKKPTFTFGQSTSENKISEG
SAKPTFSFSKSEEERKSSPISNEAAKPSFSFPGKPVDVQAPTDDKTLKPTFSFTEPAQKDSSVVSE
PKKPSFTFASSKTSQPKPLFSFGKSDAAKEPPGSNTSFSFTKPPANETDKRPTPPSFTFGGSTTN
NTTTTSTKPSFSFGAPESMKSTASTAAANTEKLSNGFSFTKFNHNKEKSNSPTSFFDGSASSTPI
PVLGKPTDATGNTTSKSAFSFGTANTNGTNASANSTSFSFNAPATGNGTTTTSNTSGTNIAGT
FNVGKPDQSIASGNTNGAGSAFGFSSSGTAATGAASNQSSFNFGNNGAGGLNPFTSATSSTN
ANAGLFNKPPSTNAQNVNVPSAFNFTGNNSTPGGGSVFNMNGNTNANTVFAGSNNQPHQ
SQTPSFNTNSSFTPSTVPNINFSGLNGGITNTATNALRPSDIFGANAASGSNSNVTNPSSIFGGA
GGVPTTSFGQPQSAPNQMGMGTNNGMSMGGGVMANRKIARMRHSKR
```

Şekil 3.17. FASTA formatındaki DIP-81N kodlu protein örneği

FASTA formatın ilk satırı Şekil 3.17'de görüldüğü gibi "> " karakteriyle başlar ve sekansa ilişkin tanımlama bilgisini içerir. Bu satırdaki "dip:DIP-81N" ifadesi ilgili protein sekansının DIP veri tabanındaki kodudur. "refseq:NP_014741" ifadesi proteinin NCBI veri tabanındaki referans numarası, "uniprot:P20676" ise UniProt veri tabanındaki referans numarasıdır. Bu satırdan sonra sekans bilgisini içeren satırlar sıralanır.

Veri tabanı düğüm (protein) ve kenar (etkileşim) bilgilerinden oluşmaktadır. DIP veri tabanındaki bir etkileşimi oluşturan her protein <*DIP:nnnN*> formatında bir tanımlayıcı ile tutulur. Ayrıca proteine ilişkin ad, fonksiyon ve en az bir protein veri tabanına ilişkin (SWISS-PROT, PIR ve GenBank) tanımlayıcı kod bilgisi de sağlanır. Bu veri tabanındaki her etkileşim ise <*DIP:nnnE*> formatında bir tanımlayıcı ile etkileşimi içeren bölge, etkileşimi tanımlayan deneysel metotlar gibi bilgilerle birlikte tutulur.

Web üzerinden (http://dip.doe-mbi.ucla.edu) araştırmacıların kullanımına açık olan veri tabanı "dosyalar" bölümünden tüm DIP veri tabanına ya da biyolojik türlere göre sadece istenen etkileşimleri içeren kısmına ulaşılabilmektedir (https://dip.doe-mbi.ucla.edu/dip/Download.cgi).

3.8.2. HPRD

HPRD (Human Protein Reference Database-İnsan Protein Referans Veri tabanı) insan proteinlerine ilişkin yayınlanmış literatür bilgilerinden uzmanlar tarafından yorumlanıp analiz edilerek elde edilen verileri içeren geniş çaplı bir veri tabanıdır (Peri ve ark., 2003; Mishra ve ark., 2006; Prasad ve ark., 2009). HPRD veri tabanının en güncel hali olan Sürüm 9'un içerdiği verilere ilişkin istatistiki bilgileri gösteren bir döküm Çizelge 3.7'de verilmiştir.

Çizelge 3.7. HPRD Sürüm 9'a ilişkin istatistiki bilgiler.

Veri Türü	Veri Sayısı
Protein sekansları	30,047
Protein-protein etkileşimleri	41,327
Posttranslasyonel modifikasyonlar	93,710
Protein ifadeleri	112,158
Subselüler lokalizasyon	22,490
Domainler	470
Pubmed Bağlantıları	453,521

Açık kaynaklı bir web sunucusunda hizmete sunulan HPRD, çok yönlü sorgu fonksiyonlarına imkân tanıyan ve verilerin dinamik olarak görüntülenmesini sağlayan bir veri tabanıdır. HPRD veri tabanında bulunan tüm içerik genel kullanıma açıktır ve Xml ya da sekme ile ayrılmış formata sahip düz dosyalar (Flat) şeklinde indirilebilmektedir (HPRD, 2019).

3.8.3. MINT

MINT (The Molecular Interaction Database), literatürden elde edilen deneysel olarak doğrulanmış protein etkileşimlerini içeren bir veri tabanıdır (Licata ve ark., 2012). Web üzerinden (http://mint.bio.uniroma2.it/) ulaşılabilen ve genel kullanıma açık olan veri tabanı 5,398 çalışmadan elde edilen 112,501 etkileşim içermektedir. 543 organizmadan 23,913 adet etkileşim yapan proteinden elde edilen bu etkileşimler *Homo sapiens* (insan), *Mus musculus* (fare), *Rattus norvegicus* (kahverengi keme), *Drosophila melanogaster* (sirke sineği), *Caenorhabditis elegans* (ipliksi solucan), *Saccharomyces cerevisiae* (maya), *Escherichia coli* (koli basili) ve *Helicobacter pylori* (mide bakterisi) de dâhil 30 farklı türe aittir. Etkileşim verilerini içeren dosyalar MITAB formatında tutulmaktadır.

3.8.4. UniProt

UniProt (The Universal Protein Resource) veri tabanı Protein sekansları ve protein fonksiyonları konusunda kapsamlı bir bilgi kaynağıdır. Araştırmacıların kullanımına açık olan bu kaynağın proteinler hakkında kapsamlı bilgiler içeren bölümü UniProtKB (The UniProt Knowledgebase) veri tabanıdır (Consortium, 2018). Proteinlere ilişkin taksonomi, sekans, hücre içi konum, fonksiyon, etkileşim, 3 boyutlu yapı ve benzeyen proteinler gibi ayrıntılı bilgileri içeren bir kaynaktır. Proteine ilişkin diğer veri tabanlarına yapılmış çapraz başvurulara da ulaşılabilmektedir.

UniProtKB veri tabanı 2 önemli kısımdan oluşur. UniProtKB/Swiss-Prot olarak isimlendirilen ilk bölüm literatürden elde edilen manuel olarak elde edilmiş ve doğrulanmış kayıtlara ilişkin bilgileri içermektedir. Veri tabanının diğer bölümü ise hesaplamalı olarak elde edilen ve doğrulama gerektiren kayıtları içerir ve UniProt/TrEMBL olarak isimlendirilmiştir.

2019_02 sürümünde 559,228 adet sekans girdisi bulunan UniProtKB/Swiss-Prot veri tabanı 264,627 referanstan 200,905,869 adet aminoasit çıkarmıştır. Yeni verilerle sürekli güncellenmekte ve büyümekte olan veri tabanının yıllara göre gelişimini gösteren grafik Şekil 1.2'de verilmiştir. Toplamda 13,719 adet farklı türe ait sekans verisini içeren veri tabanının bu sürümünde 5000'den fazla sekansla temsil edilen türler Çizelge 3.8'de gösterilmiştir.

Çizelge 3.8. UniProtKB/Swiss-Prot veri tabanında en fazla sekansa sahip 7 tür (UniProt, 2019).

Tür Adı	Açıklama	Sekans Sayısı
Homo sapiens	İnsan	20,417
Mus musculus	Fare	17,009
Arabidopsis thaliana	Fare kulağı teresi	15,832
Rattus norvegicus	Kahverengi keme	8,060
Saccharomyces cerevisiae	Maya	6,721
Bos taurus	Sığır	6,006
Schizosaccharomyces pombe	Fisyon mayası	5,141

3.8.5. Helicobacter pylori veri kümesi

Helicobacter pylori; insan midesine ya da duodenum bölgesine yerleşip gelişen ve mide zarına saldırma eğilimi olan genel bir bakteri türüdür. Genellikle zararsız kalmalarına rağmen, gastrit, ülser ve mide kanseri gibi hastalıklara da sebep olabilmektedir.

Rain ve ark. (Rain ve ark., 2001) araştırmalarında *maya iki-hibrit* sistemini kullanarak *Helicobacter pylori* proteinleri üzerinde proteinler arasında oluşan etkileşimleri belirlemişlerdir. Martin ve ark. (Martin ve ark., 2005) çalışmasında, literatürdeki birçok araştırmada yoğun olarak kullanılan 1458 tanesi etkileşen ve 1458 tanesi etkileşmeyen olmak üzere 2916 adet protein çiftini elde etmiştir. Bu veri kümesi dosyasında, tespit edilen her etkileşim için; etkileşimi kuran proteinlerin DIP, UniProt ve NCBI (The National Center for Biotechnology Information – Ulusal Biyoteknoloji Bilgi Merkezi) veri tabanları için tanımlama numaraları, bahsedilen yayın ve yazarları gibi bilgilerden oluşan kayıt satırları bulunmaktadır. NCBI ise NLM'nin (National Library of Medicine - Birleşik Devletler Ulusal Tıp Kütüphanesi) bir bölümüdür.

3.8.6. Human veri kümesi

Martin ve ark. (Martin ve ark., 2005) tarafından önerilen *Human* veri kümesi PPE tahmin sistemlerinin başarısını değerlendirmek için yoğun olarak kullanılmaktadır. Veri kümesinde 941 tanesi etkileşen ve 941 tanesi etkileşmeyen olmak üzere 1882 adet protein çifti bulunmaktadır (Martin ve ark., 2005).

Human veri kümesi için oluşturulan dosyada da *Helicobacter pylori* veri kümesi için kullanılan formatın aynısı kullanılmakta ve dosya içerisinde aynı tür bilgiler sunulmaktadır.

Helicobacter pylori ve *human* veri kümeleri iki sınıflı PPE verileri içermekte ve önerilen sistem için, sınıflandırma performansının literatürdeki birçok çalışma ile karşılaştırmasının yapılabilmesine imkân vermektedir.

3.8.7. Gram veri kümesi

Bakteri proteinleri ile ilgili olan bu veri kümesi Shen ve ark. tarafından (Shen ve Chou, 2007a) Swiss-Prot veri tabanından elde edilmiştir (Consortium, 2018). Hücre zarı, hücre duvarı, sitoplazma, hücre dışı ve periplazma alt bölgelerine ait gram-pozitif proteinleri içeren bir veri kümesidir. Aynı alt bölge için %25 oranında ya da daha fazla benzerlik gösteren protein sekansları veri kümesine dâhil edilmemiştir.

Verilerin alındığı dosya; proteinler için 5 alt hücre bölgesinden hangisine ait olduğu bilgisini, UniProt veri tabanına başvuru kodlarını ve aminoasit sekans verilerini içermektedir.

3.8.8. Gpcr veri kümesi

Hücre içi sinyal yollarının çalışmasında görev alan reseptörler olan G protein kenetli reseptörlere (GPCR-G-Protein Coupled Receptors) ilişkin bir veri kümesidir. Memeli genomundaki en büyük yapıya sahip olan ve en fazla çeşitlilik gösteren protein ailesi olarak bilinen bu yapıların ana görevi sinyalleri hücreye iletmektir (Xiao ve ark., 2008). Bu veri kümesi 2 sınıflı (GPCR ve GPCR olmayan) verilerden oluşmaktadır. Protein sekansının veri kümesine dâhil edilebilmesi için aynı kümedeki sekanslardan %40'dan daha az oranda benzerlik göstermesi ölçütü uygulanmıştır.

Verileri içeren dosya proteinler için sekans bilgileri ve UniProt veri tabanı için referans tanım kayıtlarını tutmaktadır.

3.8.9. Viral veri kümesi

Bu veri kümesi çekirdek, sitoplazma, hücre dışı ve hücre zarı alt bölgelerine ilişkin proteinleri içeren bir veri kümesidir. Virüs bulaşmış hücrelerde viral proteinlerin hücre içi alt bölgelerden hangisine ait olduğunu tespit etmeye çalışan araştırmalarda kullanılmaktadır (Shen ve Chou, 2007b; 2007a). Viral protein girdileri Swiss-Prot veri tabanından elde edilmiştir (Consortium, 2018). Aynı alt küme için %25'den daha fazla benzerlik gösteren sekanslar veri kümesine dâhil edilmemiştir.

Proteinler için 4 alt hücre bölgesinden hangisine ait oldukları, sekans bilgileri ve UniProt veri tabanı için referans tanım kayıtlarının tutulduğu bir veri tabanı dosya formatı kullanılmaktadır.

3.8.10. Membrane veri kümesi

8 farklı hücre zarı tipine ait zar proteinlerini içeren bir veri kümesidir. Bu proteinler şu türlere aittir (Chou ve Shen, 2007):

- Zar içi tek geçişli tip 1,
- Zar içi tek geçişli tip 2,
- Zar içi tek geçişli tip 3,
- Zar içi tek geçişli tip 4,
- Zar içi çok geçişli,
- Lipit zincir bağlı zar,
- GPI (Glycosylphosphatidylinositol) bağlı zar ve
- Zar çevresi.

Veri kümesindeki protein sekansları Swiss-Prot veri tabanından elde edilmiştir (Consortium, 2018). Proteinlerin biyolojik fonksiyonlarını belirlemeye çalışan araştırmalarda bir proteinin hangi zar protein türüne ait olduğunu belirlemek için kullanılan bir veri kümesidir.

Kullanılan veri dosyası; proteinler için 8 farklı hücre zarı tipinden hangisine ait olduğu bilgisini, UniProt veri tabanına başvuru kodlarını ve aminoasit sekans verilerini içermektedir.

3.8.11. AAindex veri tabanı

AAindex, aminoasitler ve aminoasit çiftleri için fizyokimyasal ve biyokimyasal özellikler hakkında sayısal değerler içeren ve genel kullanıma açık olan bir veri tabanıdır. AAindex veri tabanı literatürde yayınlanmış olan verilerin toplanmasıyla oluşturulmuştur (Nakai ve ark., 1988; Tomii ve Kanehisa, 1996; Kawashima ve ark., 1999; Kawashima ve Kanehisa, 2000).

Üç bölümden oluşan AAindex veri tabanının ilk bölümü olan AAindex1, hidrofobiklik endeksi, R grubu hacimleri ve yan zincirdeki atom sayıları gibi özellikler için nümerik değerden oluşan aminoasit indekslerinden oluşmaktadır. 20 adet nümerik değerden oluşan bir indeks, her bir aminoasit çeşidinin belirli bir fizyokimyasal ya da biyokimyasal özellik açısından sahip olduğu değeri gösteren bir dizidir. Veri tabanında bu alanda 566 adet indis dizisi bulunmaktadır.

İkinci bölüm olan AAindex2 aminoasit mutasyon matrislerinden oluşur. Bu matrisler aminoasit çift mesafeleri ve kimyasal benzerlik puanları gibi özellikler üzerinden aminoasitler arasındaki benzerlik hakkında değerleri içeren birer benzerlik matrisidir. 210 adet sayısal veriden oluşan bu matrisler sekanslar üzerindeki benzerlik çalışmalarında kullanılır. AAindex2 veri tabanında 94 adet mutasyon (benzerlik) matrisi bulunmaktadır.

Üçüncü ve son bölüm olan AAindex3 ise protein yapılarından türetilen yan zincirler arasındaki temas sayısı ve yan zincir temaslarından elde edilen etkileşim enerjileri gibi değerler üzerinden proteinler için çift yönlü ilişki potansiyellerini belirten istatistiksel değerleri içerir. Veri tabanının bu bölümünde 47 adet olası ilişki matrisi bulunmaktadır.

Veri tabanında verinin içerdiği özellikler ve saklama formatı AAindex1 bölümü için Çizelge 3.9'da gösterildiği şekildedir.

Çizelge 3.9. AAindex veri tabanının AAindex1 bölümünde tutulan özellikler için kullanılan bilgi saklama formatı

Kod	Özellik Açıklaması
H	Erişim numarası
D	Veri ismi/tanımlaması
R	PubMed sistemi tarafından verilen benzersiz kimlik numarası
A	Makalenin yazar bilgileri
T	Makalenin adı
J	Dergi bilgileri
C	Veri tabanındaki bu özelliğe benzeyen diğer özelliklerin erişim numaraları ve bu özelliklerin korelasyon değerleri
I	Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val sıralaması ile aminoasit indeks verileri

AAindex1 bölümünde bulunan "Hydrophobicity index" verisi için özellikler ve değerler Çizelge 3.10'da gösterildiği gibi tutulmaktadır. Hidrofobiklik özelliği sudan kaçma ya da suyu itme olarak bilinir ve suyu çekme ya da suya yönelme şeklinde ifade edilen hidrofiliklik özelliğinin tersidir. Kutupsuz olarak ta tanımlanan hidrofobik aminoasitler hidrokarbon içerikli yan zincir atomlarına sahip olup sudan uzak durmak için proteinin iç kısmında toplanma eğilimindedir. Bununla beraber hidrofilik aminoasitler ise suyla temas edebilmek için proteinin daha çok dış kısımlarında bulunmaktadır.

Çizelge 3.10. AAindex1 bölümünde "Hydrophobicity index" özelliği için tutulan bilgiler

Kod	Özellik Açıklaması									
H	ARGP820101									
D	Hydrophobicity index (Argos et al., 1982)									
R	PMID:7151796									
A	Argos, P., Rao, J.K.M. and Hargrave, P.A.									
T	Structural prediction of membrane-bound proteins									
J	Eur. J. Biochem. 128, 565-575 (1982)									
C	JOND750101	1.000	SIMZ760101	0.967	GOLD730101	0.936				
	TAKK010101	0.906	MEEJ810101	0.891	ROSM880104	0.872				
	CIDH920105	0.867	LEVM760106	0.865	CIDH920102	0.862				
	MEEJ800102	0.855	MEEJ810102	0.853	ZHOH040101	0.841				
	CIDH920103	0.827	PLIV810101	0.820	CIDH920104	0.819				
	LEVM760107	0.806	NOZY710101	0.800	GUYH850103	-0.808				
	PARJ860101	-0.835	WOLS870101	-0.838	BULH740101	-0.854				
I	A/L	R/K	N/M	D/F	C/P	Q/S	E/T	G/W	H/Y	I/V
	0.61	0.60	0.06	0.46	1.07	0.	0.47	0.07	0.61	2.22
	1.53	1.15	1.18	2.02	1.95	0.05	0.05	2.65	1.88	1.32

"Hydrophobicity index" özelliğine ilişkin tüm başvuru bilgilerine buradan ulaşılabileceği gibi, literatürde bu özelliğe benzeyen başka özellikler elde edilmişse bu özelliklerin erişim numaraları da Çizelge 3.10'da görüldüğü gibi "C" kodu ile belirtilmektedir. Böyle bir durumda bu özelliklerin benzerlik oranını belirten bir

korelasyon katsayısı da erişim numarasının yanında verilmektedir. Bu değerin 1'e ne kadar yakın olduğu ile özellikler arasındaki korelasyon oranı belirlenmektedir. "Hydrophobicity index" özelliği için "JOND750101" kodlu özelliğin korelasyon oranının 1 olduğu görülmektedir.

4. ÖNERİLEN SİSTEMLER

Bir hücredeki protein etkileşimlerinin tahmin edilmesi araştırmacılar tarafından bir ikili sınıflandırma makine öğrenmesi problemi olarak kabul edilmektedir ve bu problemin çözümü için farklı metotlar önerilmiştir (Reyes ve Gilbert, 2007). Fakat üretilmiş metotların çoğunda önemli kısıtlamalar bulunmaktadır. Birçok alanda PPE tahmin doğrulukları hala düşük seviyelerdedir. Bununla birlikte, önerilen metotların çoğu proteinler hakkında farklı karmaşık yapısal bilgilere ihtiyaç duymaktadır. Metotların çoğu ancak proteinler hakkında bazı ön bilgilere sahip olunursa çalışabilmektedir. Bu sebeplerden dolayı hesaplamalı PPE tahmin yöntemlerinin geliştirilmeye ihtiyacı bulunmaktadır (Roy ve ark., 2009; Ma ve Gu, 2010).

Bu tez çalışmasında, PPE verilerindeki protein sekanslarının çıkarılan farklı özelliklerle daha başarılı bir şekilde temsil edilebilmesini sağlayarak kurulan PPE tahmin sistemlerinin doğruluğu arttırılmaya çalışılmıştır. Bu amaçla "Protein-Protein Etkileşimlerinin Ağırlıklandırılmış Bir Psödo Aminoasit Kompozisyonu Tabanlı Yöntemle Tahmini" ve "Sekans Tabanlı Bir Birleşik Metot Kullanarak Protein-Protein Etkileşim Tahmini" yöntemleri geliştirilmiştir.

Geliştirilecek bir PPE tahmin sistemi için ele alınması gereken adımları; karşılaştırma veri tabanının seçimi, protein örneklerini iyi temsil edecek özelliklerin çıkarılması, tahmin algoritmasının geliştirilmesi ve önerilen sistemin doğruluğunun sınanması şeklinde sıralayan Chou (Chou, 2011) önerilen sistem için bir web arayüzünün de geliştirilebileceğini ifade etmiştir.

PPE tahmini için geliştirilen yöntemlerde kullanılan veriler, yöntemlerin literatürdeki çalışmalarla karşılaştırılabilmesi için sıklıkla tercih edilen veri kümelerinden alınmıştır.

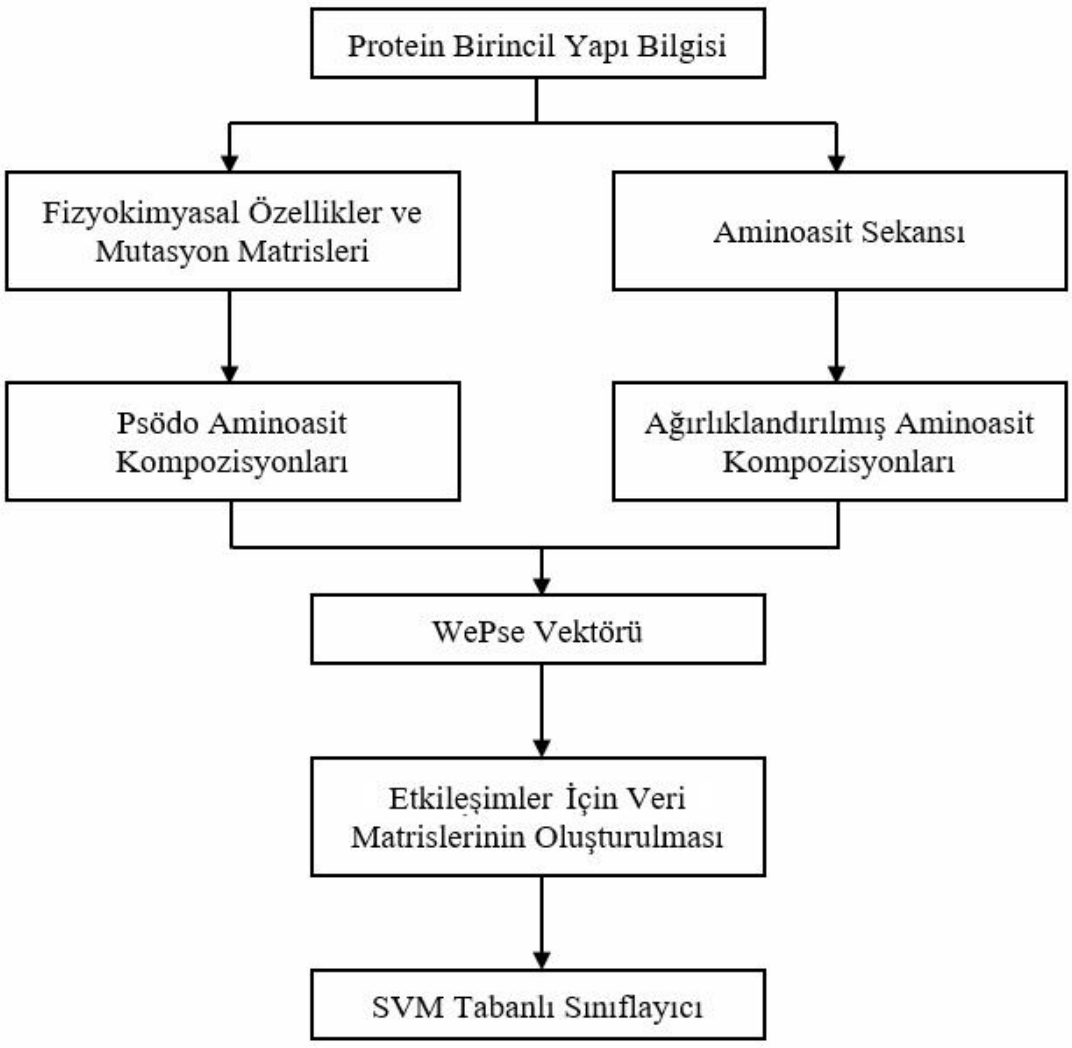
4.1. Protein-Protein Etkileşimlerinin Ağırlıklandırılmış Bir Psödo Aminoasit Kompozisyonu Tabanlı Yöntemle Tahmini

PPE'leri hesapsal metotlarla tahmin edebilmek için üretilen sistemlerde proteinlerin etkili bir şekilde ifade edilebilmesi gerekmektedir. Bu şekilde etkileşime katkı yapan özelliklerinin mümkün olduğu kadar öne çıkarılması sağlanabilecektedir. Sınıflandırma başarısını belirleyen bu durum, etkileşim tahmini için gereken bilgilerin azaltılması ve eldeki örnek kümesinin doğru bir şekilde temsil edilebilmesini sağlamayı

hedefleyen yaklaşımlarla çözülebilir. Önerilen bu sistemde eldeki protein etkileşim verilerine ait daha iyi bir özellik matrisi elde edilmeye çalışılmıştır. Bu şekilde etkileşim tahmin sisteminin sınıflandırma başarısının artırılması hedeflenmiştir.

Önerilen sistemde ağırlıklandırılmış bir aminoasit kompozisyonu mantığını kullanan bir özellik çıkarımı yöntemi kullanılmaktadır (WePse). Mevcut yöntemlerden farklı olarak, bir protein sekansı boyunca daha fazla tekrar edilen aminoasit rezidülerinin bir etkileşimin sebebi olma ihtimallerinin diğerlerine göre daha yüksek olacağı fikri temel alınmıştır. Bu yüzden sınıflandırma sistemine katkılarının daha yüksek olması için sık tekrar eden rezidülere daha yüksek temsil skorları atanmıştır. Bu şekilde etkileşim tahmin sisteminin doğruluğunu arttırmak hedeflenmiştir.

Önerilen ağırlıklandırılmış psödo aminoasit kompozisyonu *(WePse)* PPE tahmin sisteminin akış diyagramı Şekil 4.1'de gösterildiği gibidir.



Şekil 4.1. Ağırlıklandırılmış psödo aminoasit kompozisyonu için akış diyagramı

Önerilen sistemin ilk aşaması olarak, PPE verileri veri tabanından okunur. Okunan PPE çiftlerindeki her bir protein için sekans bilgileri elde edilir. Sekansları oluşturan aminoasitlerin fizyokimyasal özellikleri *AAindex* veri tabanındaki aminoasit indeks verilerinden çıkarılır. Bir indeks ise aminoasitler için fizyokimyasal ve biyolojik özellikler içeren 20 elemanlı bir dizidir. Bu değerler veri tabanından çıkarılıp Denklem 4.1'deki matris elde edilir.

$$M = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,20} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,20} \\ \cdots & \cdots & \cdots & \cdots \\ v_{i,1} & v_{i,2} & \cdots & v_{i,20} \end{bmatrix} \tag{4.1}$$

Denklem 4.1'deki *i*, veri tabanından okunan fizyokimyasal ve biyolojik özellik sayısını, $v_{a,b}$ değerleri, *b*. aminoasidin *a*. özellik için sahip olduğu değeri gösterir. *M* ise *i**20 boyutlu bir matris olmaktadır.

Özellik çıkarımı için *M* özellik vektöründeki bir aminoasit indeks satırı seçilir ve bu satırdaki değerler kullanılarak veri tabanından elde edilen PPE verilerindeki protein çiftleri için ağırlıklandırılmış standart aminoasit kompozisyonları ve psödo aminoasit kompozisyonları hesaplanır. Bu hesaplama işlemi *n* adet aminoasit indeksi için tekrarlanır. Ağırlıklandırılmış kompozisyon değerleri hesaplanırken aminoasitlerin sekanslardaki ilk gözlenmelerinde başlangıç skoru atanır. Bu skor sonraki her gözlem için belirlenen bir *c* ağırlık katsayısı kadar artırılır. Bu şekilde bir *x* proteini için çıkarılan ağırlıklandırılmış kompozisyon özelliklerini içeren matris Denklem 4.2'de görüldüğü gibi olmaktadır.

$$P(x) = \begin{bmatrix} p_1(x) \\ p_2(x) \\ \vdots \\ p_{20}(x) \end{bmatrix} \tag{4.2}$$

Burada $p_i(x)$, *P* proteini için çıkarılan bir özelliktir. Bu aşamadan sonra, protein sekansındaki aminoasit dizilim sırasına ait bilgilerin de sisteme dâhil edilebilmesi için psödo aminoasit kompozisyonları hesaplanır. Elde edilen PseAAC değerleri ile ağırlıklandırılmış kompozisyon değerleri birleştirilerek *P(x)* proteini 20+λ elemandan oluşacak şekilde Denklem 4.3 ile ifade edilir.

$$P(x) = \begin{bmatrix} p_1(x) \\ \vdots \\ p_{20}(x) \\ p_{20+1}(x) \\ \vdots \\ p_{20+\lambda}(x) \end{bmatrix} \tag{4.3}$$

Sekansta *k* komşuluğundaki aminoasit rezidüleri arasındaki korelasyon hakkında bilgi verecek olan bu değerler Denklem 4.4'e göre hesaplanmaktadır.

$$p_m = \begin{cases} \dfrac{f_m}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k} & 1 \le m \le 20 \\ | & \\ \dfrac{w\tau_{m-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda} \tau_k} & 20+1 \le m \le 20+\lambda \end{cases} \tag{4.4}$$

Burada *w* değeri ağırlıklandırılmış kompozisyon özellikleri ile PseAAC özellikleri arasındaki dengeyi sağlayacak olan katsayı faktörüdür. *k;* protein sekansı boyunca Şekil 3.13'te görüldüğü şekilde belirli bir komşuluktaki rezidüleri taramak için kullanılan değişken olmak üzere τ_k; *k* komşuluğundaki tüm rezidülerin sekans sıra korelasyonlarını gösteren *k*. korelasyon faktörüdür. τ_k faktörü ise Denklem 4.5 ile hesaplanmaktadır.

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k}\,, (k < L) \tag{4.5}$$

Burada *L* protein sekansının uzunluğudur. $J_{i,i+k}$ değerleri ise Denklem 4.6 ile hesaplanan eşleşme faktörleridir.

$$J_{i,i+k} = \frac{1}{\Gamma}\sum_{q-1}^{\Gamma}[\Phi_q(R_{i+k}) - \Phi_q(R_i)]^2 \tag{4.6}$$

Burada $\Phi_q(R_i)$ ifadesi R_i aminoasidi için q numaralı özellik değeri ve Γ değerlendirilen toplam özellik sayısıdır. Önerilen bu çalışmada 20 en yakın komşuluğa

kadar olan eşleştirmeler değerlendirilerek sekans sıra korelasyonları hesaplanmıştır. Bu şekilde her bir protein sekansı için hesaplanan özellik sayısı 40 olmaktadır.

PPE veri kümesindeki protein çiftleri okunup her bir protein çifti için Denklem 4.3'te gösterilen özellik değerleri birleştirilir. Bu şekilde bir protein çiftindeki her iki protein için çıkarılmış olan özellik verileri yan yana eklenerek özellik matrisinin bir satırı oluşturulmuş olur. Bu işlem veri kümesindeki tüm protein çiftleri için tekrarlanarak tüm özellik matrisi elde edilir. Örnek olarak çalışmada kullanılan, 2916 adet protein sekansına sahip olan *Helicobacter pylori* PPE veri kümesi için bu özellik matrisinin boyutu 2916*40 olmaktadır. Hesaplanan bu veriler geliştirilen DVM tabanlı sınıflandırma sistemine girilerek önerilen model çalıştırılır.

Sınıflandırma işleminden önce eğitim ve test için kullanılacak olan özellik matrisi Denklem 4.7 kullanılarak [0-1] aralığına normalize edilir.

$$xn_i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{4.7}$$

Burada x_i *i.* özelliğin değerini ifade ederken, *max(X)* ve *min(X)* ise ilgili özellik için en büyük ve en küçük değerleri temsil etmektedir. Hesaplanan xn_i değeri ise *i.* özelliğin normalize edilmiş olan değeridir. Normalize edilen bu değerleri kullanan DVM tabanlı bir sınıflandırma modeli kullanılır. DVM, büyük veri tabanları üzerindeki yüksek hacimli işlemler için uygun olan bir tekniktir (Joachims, 1999).

Sınıflandırma problemleri için DVM tabanlı çözümlerde çekirdek fonksiyonlarının incelemesi çalışmasında da belirtildiği gibi radyal tabanlı çekirdek fonksiyonlarının daha iyi sonuçlar verdiği gösterilmektedir (Ayhan ve Erdoğmuş, 2014). Bu sebeple bu çalışmada DVM modeli için radyal tabanlı çekirdek fonksiyonunun kullanımı tercih edilmiştir.

Radyal tabanlı çekirdek fonksiyonu kullanılan DVM modellerinde marj genişliğinin büyüklüğü ile hatalı sınıflandırma oranı arasındaki dengenin en iyi şekilde kurulmasını sağlayabilmek için kapasite parametresi *C* ve veri örneklerinin etki alanını belirleyen γ (gamma) parametrelerinin belirlenmesi gerekir. Bu parametrelerin en uygun şekilde belirlenebilmesi için grid arama gibi yöntemlerin kullanılması gerekmektedir (Chang ve Lin, 2011).

4.2. Sekans Tabanlı Bir Birleşik Metot Kullanarak Protein-Protein Etkileşim Tahmini

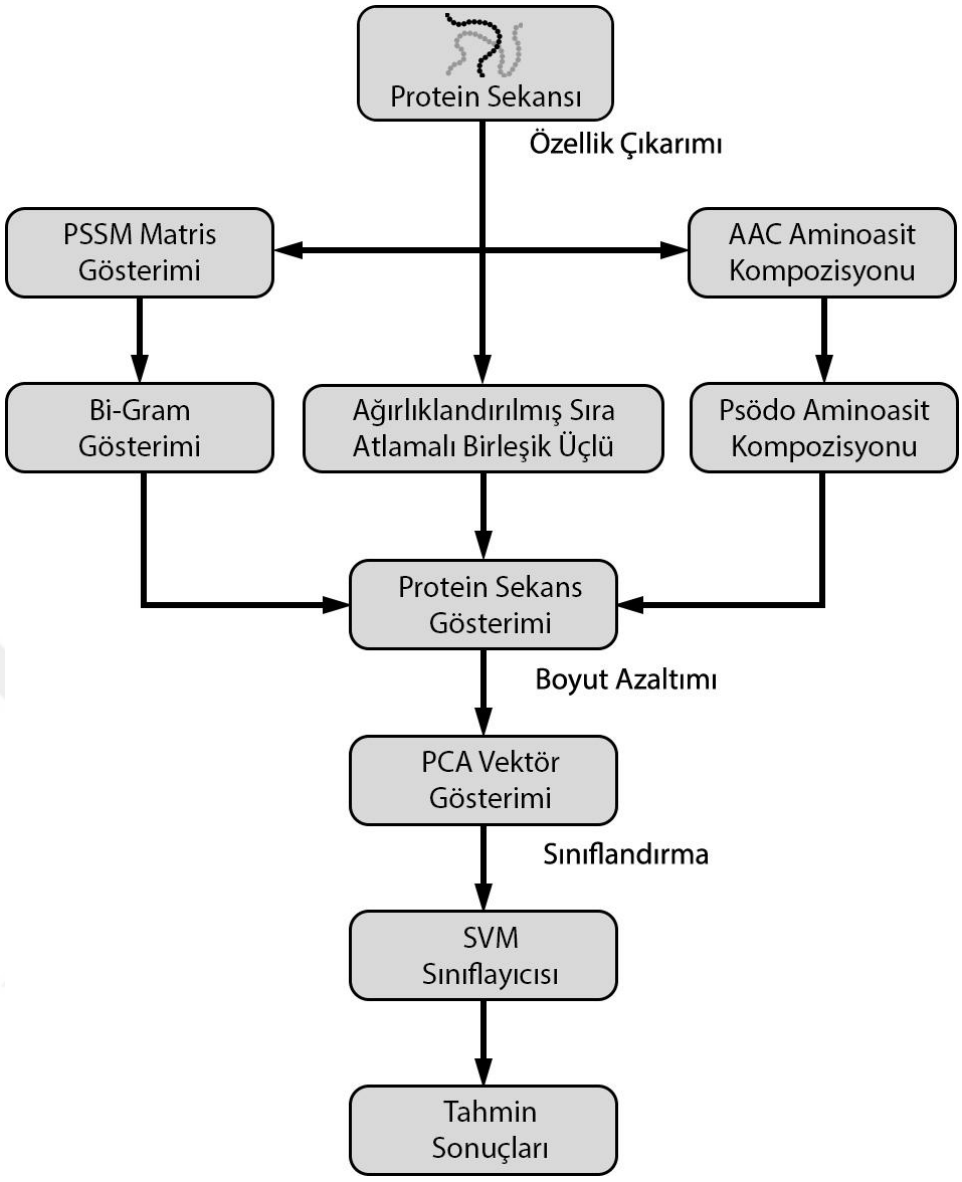
Önerilen bu ikinci sistemde proteinler arasındaki olası etkileşimlerin tahmini için yeni bir metot ortaya koyulmuştur (WeSeCT). Etkileşim tahmin sisteminin sınıflandırma başarısının arttırılması hedeflenmiştir. Bu metodun amacı etkili bir özellik çıkarımı yöntemi geliştirerek protein gösterimlerinin daha iyi yapılabilmesidir. Bu amaçla *Bi-Gram* gösterimleri, *Pseudo aminoasit kompozisyonu* ve *Birleşik üçlü* yöntemlerinden faydalanılmıştır. Elde edilen özellik matrisi üzerinde *Temel Bileşen Analizi* uygulanmıştır. DVM tabanlı bir sınıflandırma sistemi ile geliştirilen bir etkileşim tahmin algoritması kullanılmıştır.

Bu çalışmada özellik çıkarımı, boyut azaltma ve sınıflandırma aşamalarını içerecek şekilde, üç adımdan oluşan yeni bir metot önerilmiştir.

Önerilen metodun ilk adımında protein sekanslarından anlamlı özellik değerlerinin elde edilebilmesi için bir özellik çıkarım işlemi uygulanmaktadır. Bu aşamada proteinler için veri tabanlarından elde edilen aminoasit sekans verileri kullanılmaktadır. Bu veriler *Standart Aminoasit Kompozisyonu, Psödo Aminoasit Kompozisyonu* (Chou, 2001), *Bi-Gram* gösterimleri (Sharma ve ark., 2013) ve *Ağırlıklandırılmış Sıra-Atlamalı Birleşik Üçlü* özellik çıkarım metotlarının birlikte kullanımıyla işlenmektedir. Bu çalışmada önerilen *Ağırlıklandırılmış Sıra-Atlamalı Birleşik Üçlü* yöntemi (WeSeCT) ise standart *Birleşik Üçlü* (Shen ve ark., 2007) yönteminin değiştirilmiş bir formudur.

İkinci adımda yukarıda bahsedilen metotlarla elde edilen üç özellik matrisi birleştirilir. Veride değer kaybı oluşmadan özellik sayısını azaltmak için elde edilen matrise Temel Bileşen Analizi (Zhang ve ark., 2012b) uygulanarak özellik matrisinin boyutu düşürülür.

Son adımda ise Temel Bileşen Analizi ile elde edilen son özellik matrisinin eğitim ve test kümelerine ayrılır. Etkileşim tahmin sistemi, verilerin 10 kat çapraz geçerleme kullanılarak uygulanan DVM tabanlı bir sınıflandırma yöntemi ile işlenmesi ile tamamlanmaktadır. Önerilen metodun adımlarını gösteren akış diyagramı Şekil 4.2'de görülmektedir.



Şekil 4.2. Önerilen metodun akış diyagramı.

Veri tabanından çıkarılan protein sekans bilgileri özellik çıkarımının ilk aşamasında *Konuma Özel Puanlama Matrisi* yöntemi ile işlenmektedir. *Konuma Özel Puanlama Matrisi*, 20 aminoasit çeşidi için protein sekansının her noktasındaki bulunma oranları hakkında bilgi vermektedir. Bu matris üzerinden bi-gram özellikleri hesaplanmaktadır. *Konuma Özel Puanlama Matrisi* yöntemi ile L adet satırı ve 20 adet sütunu olan bir P matrisi i. satır ve j. sütunu ifade eden $p_{i,j}$ olasılık değerlerinden oluşmaktadır. Bu olasılık değerleri kullanılarak Bi-Gram bulunma frekansları Denklem 4.8 ile elde edilmektedir (Sharma ve ark., 2013).

$$B_{m,n} = \sum_{i=1}^{L-1} p_{i,m} \times p_{i+1,n} \tag{4.8}$$

Burada L sekans uzunluğunu göstermek üzere i değişkeni 1 ile L-1 arasında değerler alır. m ve n değişkenleri ise 1 ile 20 arasında değerler alır. $B_{m,n}$ toplamda 400 adet bi-gram özellik vektörünü ifade etmektedir. Bu şekilde *HPRD, Human* ve *Helicobacter Pylori* veri tabanlarından elde edilen protein sekansları için 400 kolonlu bir bi-gram özellik matrisi elde edilmektedir.

Özellik çıkarımının ikinci aşamasında, sekanslar için psödo aminoasit kombinasyonları hesaplanmaktadır. Chou (Chou, 2001) tarafından önerilen bu yöntem, standart aminoasit kombinasyonunun sahip olduğu sekanslar hakkında sıra-dizilim bilgisini içermeme eksikliğini gidermektedir. Bu şekilde sekans içinde gizli olan sırasal bilgiler etkileşim tahmininde kullanılabilmektedir.

Psödo aminoasit kombinasyonları 20+n adet bileşenden oluşur. Bunlardan ilk 20 tanesi standart aminoasit kombinasyonlarına ait iken sonraki n tanesi sekans-sıra ilişkilerine ait bilgileri içermektedir. Bahsedilen 20+n adet değeri Denklem 4.4, Denklem 4.5 ve Denklem 4.6 ile hesaplanmaktadır. Bu aşamada 40 kolonlu bir özellik matrisi elde edilmektedir.

Özellik çıkarımının üçüncü aşamasında ise etkileşim bölgelerini başarılı bir şekilde tespit edebilen *Ağırlıklandırılmış Sıra-Atlamalı Birleşik Üçlü* yöntemi ile sekans içindeki her bir aminoasidin, sahip olduğu komşuluklarla beraber değerlendirilmesi sağlanıyor.

Standart *Birleşik Üçlü* yöntemi sekans içindeki her bitişik üç aminoasidi bir birim olarak değerlendirmektedir (Shen ve ark., 2007). Belirlenen bu üçlü birimlerin frekansının etkileşim tahmininde kullanılabilecek önemli veriler olduğu öne sürülmektedir (Wang ve Wu, 2018). Standart *Birleşik Üçlü* yöntemine göre 20 aminoasit çeşidi dipol ölçüleri ve yan zincir hacimlerine göre Çizelge 4.1'de görüldüğü gibi 7 gruba ayrılır.

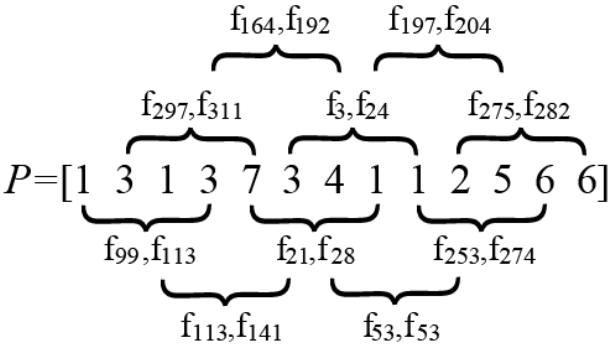
Çizelge 4.1. Birleşik üçlü yöntemine göre oluşturulan 7 grup

Grup No	Aminoasitler
1	A, G, V
2	I, L, F, P
3	Y, M, T, S
4	H, N, Q, W
5	R, K
6	D, E
7	C

Örnek bir protein sekansını oluşturan tüm aminoasitlerin isimleri aşağıdaki protein örneğinde görüldüğü gibi grup numaralarıyla değiştirilir.

$$P = [1\ 3\ 1\ 3\ 7\ 3\ 4\ 1\ 1\ 2\ 5\ 6\ 6]$$

Sekansın elde edilen yeni gösterimi üzerinde olabilecek 343 (7x7x7) adet üçlü grup kombinasyonlarının frekansları hesaplanır. Bu hesaplama adımı Şekil 3.12'de görüldüğü gibi yapılmasına rağmen bu çalışmada önerilen yöntemde hesaplama şekli değiştirilmiştir. Sekans içerisindeki üçlü birimleri ararken, bir üçlü birimin arasına fazladan bir aminoasit girebileceği varsayımı göz önüne alınarak frekans hesaplamaları ona göre yapılmaktadır. Diğer bir ifadeyle, üçlü gruplar aranırken bir aminoasit atlanarak oluşan üçlüler de hesaba katılmaktadır. Bu şekilde üçlü birim frekanslarının örnek bir sekans üzerinde hesaplanması Şekil 4.3'te görülmektedir.



Şekil 4.3. Ağırlıklandırılmış sıra-atlamalı yöntemle protein örneği için birleşik üçlü frekanslarının çıkarılması

Şekil 4.3'te görülen sekans üzerinde elde edilen grup kombinasyonları ve frekans sayıları Çizelge 4.2'de gösterilmektedir.

Çizelge 4.2. Ağırlıklandırılmış sıra-atlamalı birleşik üçlü yöntemi ile protein örneği için hesaplanan frekanslar

Grup kombinasyonları	Frekans adı	Frekans değeri
311	f3	1
731	f21	1
341	f24	1
741	f28	1
412	f53	2
113	f99	1
133	f113	2
173	f141	1
334	f164	1
374	f192	1
115	f197	1
125	f204	1
126	f253	1
156	f274	1
256	f275	1
266	f282	1
317	f297	1
337	f311	1

Uygulanan *Ağırlıklandırılmış Sıra-Atlamalı Birleşik Üçlü* yöntemi ile 343 kolonlu bir özellik vektörü elde edilmektedir. Bu boyut standart *Birleşik Üçlü* yönteminin kullanılması halinde üretilecek olan vektörün boyutuyla aynı olmasına rağmen sekans içindeki gizli bilgilerin de kullanılabilmesini sağlayarak etkileşim tahmin sisteminin daha başarılı çalışmasına imkân vermektedir.

Özellik çıkarımının üç aşaması da işletilince önerilen yöntemin ilk adımı tamamlanmış olur. Bu noktada 400 kolonlu *Bi-gram*, 40 kolonlu *PseAAC* ve 343 kolonlu *Birleşik Üçlü* özellik vektörleri elde edilmektedir.

Önerilen yöntemin ikinci adımında bu vektörler birleştirilerek 783 kolonlu tek bir özellik vektörüne dönüştürülmektedir. Oluşan bu toplam vektörü üzerinde *Temel Bileşen Analizi* yöntemi uygulanmaktadır. *Temel Bileşen Analizi*, veri kümelerinin sahip olduğu bilgileri kaybetmeden boyutunu azaltabilme imkânı veren bir tekniktir (You ve ark., 2013). Toplam özellik vektörü bu yöntemle 390 kolonlu bir vektöre indirilmektedir. Bu şekilde işlenen veri miktarı azaltılarak hesaplama süresi kısaltılması sağlanmaktadır.

Yöntemin son adımı sınıflandırma sisteminin kurulmasından ibarettir. Etkileşim tahmini için farklı makine öğrenmesi teknikleri kullanılmaktadır. DVM yöntemi, sahip olduğu gelişmiş genelleme performansı sayesinde biyoinformatik de dâhil olmak üzere birçok alanda en sık kullanılan sınıflandırma tekniklerinden birisidir (Brown ve ark., 2000).

Makine öğrenmesi literatüründe, sınıflandırma problemlerinin radyal tabanlı çekirdek fonksiyonlarını kullanan DVM tabanlı sistemlerle çözülmesi ile daha iyi sonuçların elde edildiği belirtilmektedir (Ayhan ve Erdoğmuş, 2014). Bu sebeple bu çalışmada DVM modeli için radyal tabanlı bir gauss çekirdek fonksiyonu seçilmiştir.

DVM modelinin sınıflandırma performansı çekirdek fonksiyonu (K), kapasite parametresi (C) ve gamma (γ) değerlerinin seçimine bağlıdır. C parametresi marj genişliğini mümkün olduğu kadar arttırmak ve hata oranını azaltmak arasındaki dengeyi kurmak için kullanılır. Gama parametresi ise tek bir örneğin, bulunduğu uzaydaki etki alanını belirlemektedir. Diğer bir deyişle gamma değeri uzaydaki iki örnek arasındaki benzerlik oranının belirlenmesi için kullanılır. Destek vektör sınıflandırması için geliştirilen bütünleşmiş bir yazılım olan LIBSVM araç kutusunun grid arama yöntemi, geliştirilen model için en iyi sınıflandırma sonuçlarını sunacak olan parametre (C ve γ) değerlerini belirlemeyi sağlamaktadır. Geliştirilen tahmin modelinde bu yöntem kullanılarak C ve gamma (γ) parametreleri sırasıyla 32 ve 0.04 değerlerine optimize edilmiştir.

5. DENEYSEL SONUÇLAR

PPE tahmin problemi için önerilen iki yöntemin elde ettiği deneysel sonuçlar burada verilmektedir. Bu yöntemlerin başarısı farklı değerlendirme ölçütlerine göre test edilmekte ve performansları değerlendirilip aynı veri kümeleri üzerinden önceki çalışmalarla karşılaştırılmaktadır.

Bu tez çalışmasında PPE tahmini için önerilen yöntemlerin kodlama işlemlerinin tamamı, Intel Core i5-8250U 1.6 Ghz CPU, 12 GB RAM'e sahip bir bilgisayarda Matlab 2016b (9.1) uygulama geliştirme aracı ile gerçekleştirilmiştir. Yüksek seviyeli bir programlama dili olan Matlab (Matris Laboratuvarı); program geliştirme, yüksek boyutlu matrisler üzerinde ileri sayısal hesaplamalar ve veri analizi gibi konulardaki üstün başarısından dolayı akademik ve endüstriyel alanlarda araştırmacılar tarafından sıklıkla tercih edilmektedir (Tsuda ve ark., 2005; Gu ve ark., 2010).

5.1. Ağırlıklandırılmış Bir Aminoasit Kompozisyonu Tabanlı Protein-Protein Etkileşimi Tahmini Yöntemi ile Elde Edilen Sonuçlar

Önerilen yöntemin sonuçlarını önceki çalışmalarla karşılaştırabilmek amacıyla özellik çıkarımından sonra elde edilen veri kümesi, eğitim ve test olmak üzere iki bölüme ayrılmıştır.

Sınıflandırma modelinin geliştirilmesi için Matlab uygulaması üzerinde OSU-SVM araç kutusu kullanılmıştır. Bu çalışmadaki etkileşim tahmin sistemi için DVM modelinin γ ve C parametreleri sırasıyla 3 ve 100 değerlerine optimize edilmiştir.

Önerilen etkileşim tahmin sistemi *Human, Helicobacter Pylori, Gram, Gpcr, Viral* ve *Membrane* veri kümeleri üzerinde uygulanmıştır. Sistemin performans ölçümünde ROC eğrisi yöntemi kullanılmıştır. Farklı veri kümeleri için alınan sonuçlar AUC (ROC eğrisi altında kalan alan) değerleri olup bu değerler Çizelge 5.1'de görülmektedir.

Çizelge 5.1. Önerilen sistemin test edilen veri kümeleri üzerinde elde ettiği eğri altındaki alan (AUC) değerleri

Veri Kümesi	AUC
Human	0.7348
Helicobacter pylori	0.9221
Gram	0.9290
Gpcr	0.9930
Viral	0.8110
Membrane	0.9440

AUC değerlendirme ölçütü için en ideal durum 1 değerini elde etmektir. Sınıflandırma sisteminin vereceği sonucun mümkün olduğu kadar 1'e yakın olması beklenir. Diğer bir ifadeyle ROC grafiğinin sol üst köşeye yakın olması istenir. En kötü durum ise 0.5 değeridir. Bu değer tamamıyla rastgele bir sınıflandırma yapılması durumunda ortaya çıkabilecek bir sonuçtur.

Sınıflandırma sistemlerinin performanslarının değerlendirilmesinde sıklıkla kullanılan ölçütlerden olan doğruluk (Acc), hassaslık (Sen), özgüllük (Spe) ve kesinlik (Pre) değerleri ile 1'e mümkün olduğu kadar yakın sonuçlar elde edilmeye çalışılır. Önerilen etkileşim tahmin sisteminin *Human, Helicobacter pylori, Gram, Gpcr, Viral* ve *Membrane* veri kümeleri üzerinde test edilmesiyle elde edilen değerler Çizelge 5.2'de verilmiştir.

Çizelge 5.2. Önerilen metot ile *Human, Helicobacter Pylori, Gram, Gpcr, Viral ve Membrane* veri kümeleri için elde edilen doğruluk (Acc), duyarlılık (Sen), özgüllük (Spe), kesinlik (Pre) ve Mcc sonuçları

Değerlendirme ölçütü	Human	Helicobacter pylori	Gram	Gpcr	Viral	Membrane
Acc.	0.6892	0.8083	0.9031	0.9490	0.8054	0.9090
Sen.	0.6840	0.7871	0.9250	0.9541	0.7964	0.9295
Spe.	0.6940	0.8274	0.8813	0.9439	0.8144	0.8885
Pre.	0.6818	0.8040	0.8862	0.9444	0.8110	0.8937
Mcc.	0.3690	0.6098	0.8070	0.8980	0.6109	0.8188

Elde edilen değerlendirme ölçüt sonuçları önerilen etkileşim tahmin sisteminin ilgili veri kümeleri için oldukça başarılı çalıştığını göstermektedir.

Bu çalışmada uygulanan metot, elde edilen sonuçların önceki çalışmalarla karşılaştırmalı olarak incelenebilmesi için *Human, Helicobacter Pylori, Gram, Gpcr, Viral* ve *Membrane* veri kümeleri üzerinde test edilmiştir. Bu veri kümelerini kullanan ve bu çalışmada karşılaştırma amaçlı başvurulacak olan özellik çıkarım metotlarının listesi Çizelge 5.3'te verilmiştir. Bu çizelgede metotların kullandığı özellik türleri ve boyutları da görülmektedir.

Çizelge 5.3. Literatürdeki bazı çalışmalarda kullanılan özellik çıkarım yöntemleri. (*: araştırmacı tarafından sabitlenen özellik boyutu, **: seçilen fizyokimyasal özellik sayısının iki katı kadar özellik boyutu)

Algoritma ismi	Kısa isim	Özellik boyutu	Seçilen özellikler	Referans
Aminoasit kompozisyonu	AC	20	Aminoasit frekansları	(Reyes ve Gilbert, 2007)
Bölünmüş aminoasit kompozisyonu	SAC	60	3 sekans bölümünden Aminoasit frekansları	(Kumar ve ark., 2006)
2-Gram	2G	400	İkili aminoasit frekansları	(Nanni, 2005a)
Otokovaryans	ACO	40	Aminoasit kompozisyonları ile fizyokimyasal özellikleri	(Li ve ark., 2012)
Tam dizi	FS	20	Aminoasit kompozisyonları	(Xia ve ark., 2010a)
Uyumsuzluk çekirdeği	MK	*	*k* sekans bölümünden Aminoasit frekansları	(Leslie ve ark., 2004)
Dalgacık tanımlayıcısı	WA	100	Fizyokimyasal özellikler ile dekompozisyon ölçekleri	(Li ve ark., 2009)
Benzer kalıntı çifti	RC	1200	Fizyokimyasal özellikler	(Guo ve Sun, 2005)
Aminoasit grup tabanlı fizyokimyasal kodlama	AAG	**	Aminoasit kompozisyonları ile fizyokimyasal özellikleri	(Hu ve Zhang, 2009)
AAIndexLoc	AA	65	Aminoasit kompozisyonları ile fizyokimyasal özellikleri	(Chou, 2001)
N-Gram	NG	64-400	Bir grup aminoasit için frekans değerleri	(Xia ve ark., 2010a)
Önerilen Metot (WePse)	WP	40	Aminoasit kompozisyonları ile fizyokimyasal özellikleri	(Goktepe ve ark., 2016)

Çizelge 5.3'te verilen özellik çıkarım yöntemlerinden ilki olan Aminoasit kompozisyonu (AC), 20 aminoasidin sekans içinde görülme sıklıklarını gösteren en temel yöntemlerden birisidir (Reyes ve Gilbert, 2007). 2-Gram (2G), protein içindeki belirli aminoasit çiftlerinin frekansını hesaplayan bir yöntemdir (Nanni, 2005a). N-Gram (NG) ise 2-Gram yönteminin geliştirilmiş bir formudur ve *n* adet aminoasit dizisinin görülme frekanslarını hesaplar (Xia ve ark., 2010a). Otokovaryans (ACO) yöntemi, standart amino asit kompozisyonu değerlerini ve dizi sırasının etkisini temsil eden değerleri içeren PseAAC özelliklerini bir proteinden elde etmektedir (Chou, 2001). Tam dizi (FS) ise amino asitlerin tüm fizikokimyasal özelliklerine dayanan bir özellik çıkarımıdır (Xia ve ark., 2010a). Benzer kalıntı çifti (RC), PseAAC modeline dayanan bir özellik çıkarımı yöntemidir (Guo ve Sun, 2005). AAIndexLoc (AA), proteinlerin ait oldukları hücre alt lokalizasyonlarını tahmin etmek amacıyla kullanılan aminoasit indekslerini kullanmaktadır (Tantoso ve Li, 2008). Dalgacık tanımlayıcısı (WA) yöntemi, protein sekansını aminoasit fizyokimyasal özelliklerini kullanarak numerik bir diziye dönüştürüp dalgacık güç spektrumu tekniğini uygulamaktadır (Li ve ark., 2009).

Uyumsuzluk çekirdeği (MK), protein sekansının *k* uzunluğundaki alt sekansları arasındaki benzerliği kullanan bir yöntemdir (Leslie ve ark., 2004). Bölünmüş aminoasit kompozisyonu (SAC) yöntemi protein sekanslarını amino-ucu, karboksil-ucu ve orta kısım olmak üzere üç bölüme ayırır ve bu üç bölümün kompozisyonlarını hesaplar. Uç bölümlerinin uzunluğunu 20 aminoasitle belirleyen yöntem sekansın kalan kısmını orta bölüm olarak değerlendirir (Kumar ve ark., 2006). Aminoasit grup tabanlı fizyokimyasal kodlama (AAG) ise aminoasitleri belirli bir fizyokimyasal özelliğe göre kümeleme mantığına dayanmaktadır (Hu ve Zhang, 2009).

Değerlendirme ölçütü olarak en sık kullanılan tekniklerden olan ROC eğrisi altında kalan alan (AUC-Area Under Roc Curve) yöntemi kullanılmıştır. Elde edilen sonuçlar karşılaştırmalı olarak Çizelge 5.4'te gösterilmiştir.

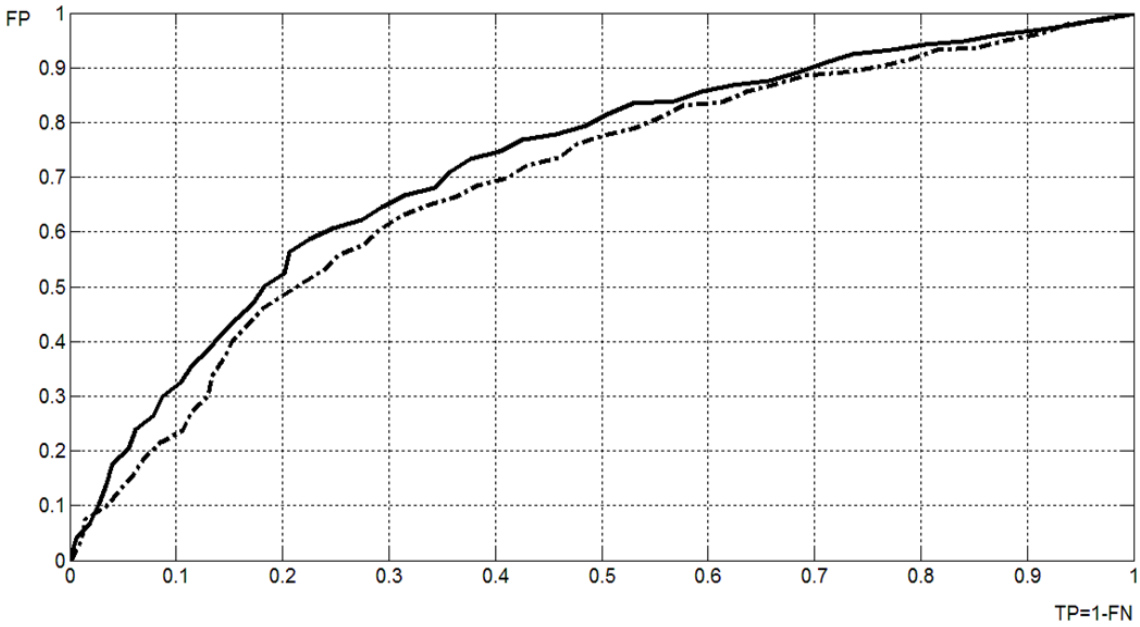
Çizelge 5.4. Önerilen metodun AUC sonuçlarının 6 farklı veri kümesi için literatürdeki çalışmalarla karşılaştırılması

Uygulanan Metot	Human	Helicobacter pylori	Gram	Gpcr	Viral	Membrane	Rank
AC (Reyes ve Gilbert, 2007)	0.613	0.780	0.872	0.960	0.615	0.889	10.8
SAC (Kumar ve ark., 2006)	0.679	0.824	0.870	0.959	0.685	0.917	9.2
2G (Nanni, 2005a)	0.687	0.918	0.899	0.978	0.647	0.940	7.3
ACO (Li ve ark., 2012)	0.704	0.901	**0.929**	0.992	0.754	0.926	4.2
FS (Xia ve ark., 2010a)	0.667	0.786	0.857	0.981	0.660	0.880	10.0
MK (Leslie ve ark., 2004)	0.665	0.768	0.702	0.988	0.718	0.936	9.1
WA (Li ve ark., 2009)	0.690	0.889	0.918	0.992	0.755	**0.953**	4.1
RC (Guo ve Sun, 2005)	0.717	**0.925**	0.880	0.991	0.608	**0.953**	5.0
AAG (Hu ve Zhang, 2009)	0.701	0.917	0.905	0.988	0.808	**0.953**	4.4
AA (Chou, 2001)	0.638	0.805	0.921	0.991	0.699	0.910	7.5
NG (Xia ve ark., 2010a)	0.693	0.920	0.906	0.991	0.735	0.943	4.7
Önerilen Metot (WePse) (Goktepe ve ark., 2016)	**0.735**	0.922	**0.929**	**0.993**	**0.811**	0.944	1.8

Çizelge 5.4'te görüldüğü gibi önerilen etkileşim tahmin metodunun *Human*, *Gpcr* ve *Viral* veri kümelerinde sırasıyla 0.735, 0.993 ve 0.811 AUC değerleri alınmıştır. Diğer metotların elde ettiği değerler ile kıyaslandığında bu sonuçların daha yüksek olduğu görülmektedir. Metodumuz, *Helicobacter pylori*, *Gram* ve *Membrane* veri kümelerinde elde ettiği AUC sonuçlarına göre de listedeki en iyi metotlar arasında

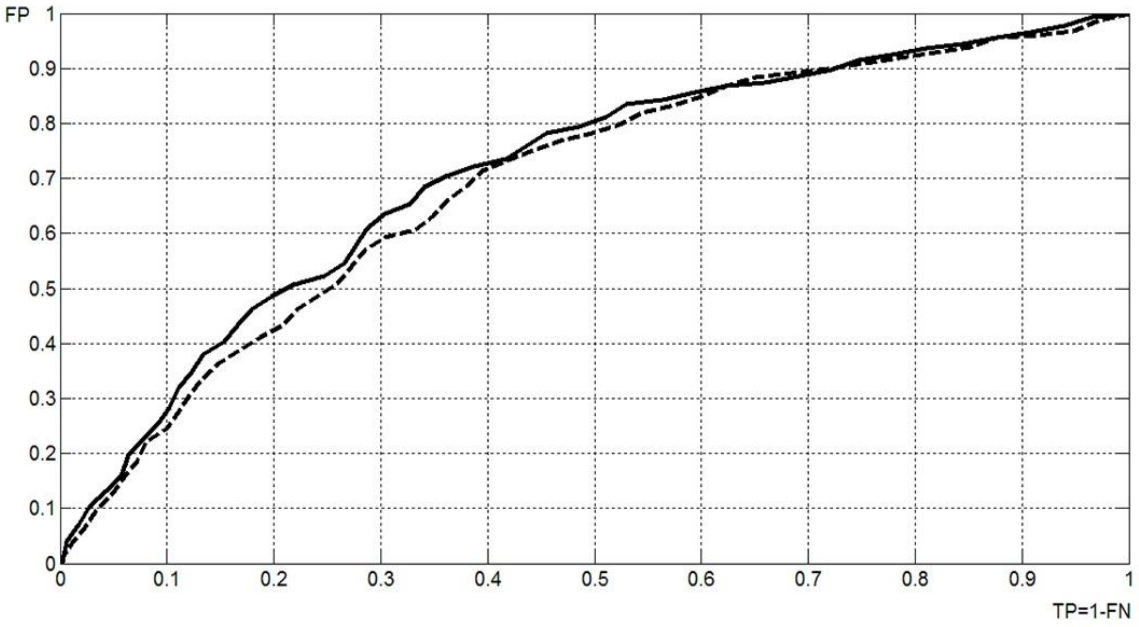
bulunmaktadır. Bu çizelgede verilen en son kolon tahmin metotlarının rank analiz sonuçlarını göstermektedir. Bu analize göre bakıldığında önerilen metodun tahmin başarısının diğer metotlardan daha yüksek olduğu görülmektedir.

Önerilen metot, PseAAC metodu (Chou, 2001) ile ROC eğrileri kullanarak karşılaştırılmıştır. Belirlenen bir fizyokimyasal özellik kullanılarak her iki metodun elde ettiği ROC eğrileri tek grafik üzerinde çizilmiştir. Hidrofobisite özelliği, proteinlerin kararlı hâle geçmek için katlanması ve uygun bir fonksiyonu yerine getirmek için açılması süreçlerinde önemli bir rol oynamaktadır. Bununla beraber hidrofobisite ve hidrofilisite özelliklerinin protein etkileşimlerinde önemi büyüktür (Cserhati ve Szogyi, 1995; Dyson ve ark., 2006; Holt ve ark., 2019). Bu yüzden "hidrofobik moment yönü" ve "hidrofilisite ölçeği" fizyokimyasal özellikleri için bu iki metodun elde ettiği ROC eğrileri çizilmiş ve bu eğrilerin kıyaslaması Şekil 5.1 ve Şekil 5.2'de sunulmuştur.



Şekil 5.1. Önerilen metot ile PseAAC yönteminin "hidrofobik moment yönü" özelliği için ROC eğrileri. (düz çizgi: Önerilen metot, kesikli çizgi: PseAAC yöntemi)

Bu grafikte elde edilen ROC eğrileri için altında kalan alanlar hesaplandığında, "hidrofobik moment yönü" özelliği için PseAAC yönteminin 0.6935 ve önerilen metodun 0.7255 AUC değerine ulaştığı görülmüştür.



Şekil 5.2. Önerilen metot ile PseAAC yönteminin "hidrofilisite ölçeği" özelliği için ROC eğrileri. (düz çizgi: Önerilen metot, kesikli çizgi: PseAAC yöntemi)

"Hidrofilisite ölçeği" özelliği için ROC eğrileri için altında kalan alanlar hesaplandığında, PseAAC yöntemi 0.6907 ve önerilen metot 0.7101 AUC değerine ulaşmıştır.

5.2. Sekans Tabanlı Bir Birleşik Metot Kullanımıyla Protein-Protein Etkileşim Tahmini Yöntemi ile Elde Edilen Sonuçlar

Bu çalışmada veri tabanı olarak Pan ve ark. (Pan ve ark., 2010) tarafından kullanılan veri kümesi üzerinde çalışılmıştır. Bu veri kümesindeki pozitif PPE verileri HPRD veri tabanından elde edilmiştir. Verinin doğru gösterimi için tekrar eden PPE verileri veri kümesinden çıkarılmıştır. Ayrıca uzunluğu 50 aminoasitten daha az olan sekanslar da göz ardı edilmiştir. Bu şekilde 2835 adet benzersiz protein sekansı elde edilmiştir. Etkileşim yapmayan çiftleri gösteren negatif örnekler farklı hücre içi bölgelerde bulunan protein sekanslarının eşleştirilmesi yöntemiyle elde edilmiştir. Sonuç olarak etkileşimlerin rastgele seçilmesi ile 3500 tanesi pozitif ve 3500 tanesi negatif örneklerden oluşan 7000 örnekli bir veri kümesi oluşturulmuştur.

Çalışmada, Martin ve ark. (Martin ve ark., 2005) tarafından tanımlanan ve literatürde yoğun olarak kullanılan *Human* ve *Helicobacter Pylori* veri kümeleri de kullanılmıştır. *Human* veri kümesi 941'i pozitif ve 941'i negatif olmak üzere 1882 adet

etkileşim verisi içermektedir. *Helicobacter Pylori* veri kümesi ise etkileşim yapan 1458 ve etkileşim yapmayan 1458 protein çiftinden oluşan toplam 2916 adet etkileşim verisini içermektedir.

Bu çalışmada önerilen sekans tabanlı etkileşim tahmini metodu bilgisayarda Matlab 2016b (9.1) uygulama geliştirme aracı kullanılarak hazırlanmıştır. DVM modelinin geliştirilmesi için ise Matlab uygulamasına LIBSVM araç kutusunun 3.22 sürümü eklenmiştir. Başta makine öğrenmesi olmak üzere birçok alanda geniş bir kullanım alanına sahip olduğu için LIBSVM paketi tercih edilmiştir. LIBSVM, DVM sınıflandırması, regresyon ve dağılım tahmini konuları için ara yüz sağlamaktadır (Chang ve Lin, 2011).

Bu çalışmada elde edilen sonuçlar doğru sınıflandırılan pozitif ve negatif çiftlerin tüm çiftlere oranı olan genel sınıflandırma doğruluğu (Acc), doğru sınıflandırılan pozitif çiftlerin tüm pozitif çiftlere oranı olan hassaslık (Sen), doğru sınıflandırılan negatif çiftlerin tüm negatif çiftlere oranı olan özgüllük (Spe), doğru sınıflandırılan pozitif çiftlerin yapılan tüm pozitif tahminlere oranı olan kesinlik (Pre) ve Matthews Korelasyon Katsayısı (MCC) ölçütleri üzerinden değerlendirilmiştir. Bunların dışında eğri altındaki alan (AUC) değerlendirme ölçütü de kullanılmıştır. Bu yöntem, doğru pozitif oranına karşı yanlış pozitif oranının çizilmesi ile elde edilen ROC eğrisi altında kalan alanın hesaplanmasına dayanmaktadır. Bu da en iyi performans ölçütleri arasında gösterilmektedir (Bradley, 1997).

Önerilen çalışma *helicobacter pylori* veri tabanı üzerinde çalıştırıldığında 0.8915 sınıflandırma doğruluğu, 0.8813 hassaslık, 0.8729 kesinlik ve 0.7721 MCC değerlerini üretmiştir. Metodun, *helicobacter pylori, human* ve *HPRD* veri kümeleri üzerinde sınıflandırma doğruluğu, hassaslık, kesinlik ve MCC ölçütleri ile elde ettiği sonuçların listesi Çizelge 5.5'te verilmiştir.

Çizelge 5.5. *Helicobacter pylori, Human* ve *HPRD* veri kümeleri için elde edilen doğruluk (Acc), duyarlılık (Sen), özgüllük (Spe), kesinlik (Pre) değerleri

Veri Kümesi	Acc.	Sen.	Spe.	Pre.
Helicobacter pylori	0.8915	0.8813	0.8996	0.8729
Human	0.7381	0.7324	0.7461	0.7411
HPRD	0.9345	0.8929	0.9532	0.8984

Önerilen metodun tahmin performansı AUC ölçütüyle de test edilmiştir. *helicobacter pylori, human* ve *HPRD* veri kümeleri için elde edilen test sonuçları Çizelge 5.6'da görülmektedir.

Çizelge 5.6. *Helicobacter pylori*, *Human* ve *HPRD* veri kümeleri için elde edilen eğri altındaki alan (AUC) değerleri

Veri Kümesi	AUC
Helicobacter pylori	0.9371
Human	0.8320
HPRD	0.9300

Önerilen çalışmanın *helicobacter pylori* veri kümesi üzerinde sınıflandırma doğruluğu, hassaslık, kesinlik ve MCC ölçütleri ile elde ettiği sonuçların listesi ve literatürdeki aynı veri tabanını kullanan çalışmalarla karşılaştırılması Çizelge 5.7'de verilmiştir. Çizelgede "-" ile belirtilen alanlar ölçüt değerlerinin ilgili çalışmalardan elde edilemediğini göstermektedir. Metotların rank analizi yapılmış olup çizelgenin en son kolonunda gösterilmiştir.

Çizelge 5.7. Önerilen metodun tahmin performansının, *helicobacter pylori* veri kümesi için doğruluk (Acc), duyarlılık (Sen), kesinlik (Pre) ve Mcc ölçütleri üzerinden önceki çalışmalarla karşılaştırılması

Tahmin Modeli	Acc	Sen	Pre	Mcc	Rank
Phylogenetic bootstrap (Bock ve Gough, 2003)	0.7580	0.6860	0.8020	-	11.0
HKNN (Nanni, 2005a)	0.8400	0.8600	0.8400	-	8.3
Signature products (Martin ve ark., 2005)	0.8340	0.7990	0.8570	-	9.3
Birleşik HKNN (Nanni ve Lumini, 2006)	0.8660	0.8670	0.8500	-	5.7
Birleşik ELM (You ve ark., 2013)	0.8750	**0.8895**	0.8615	**0.7813**	2.8
Ağırlıklı seyrek (Huang ve ark., 2015)	0.8674	0.8643	0.8701	0.7699	4.3
Çok ölçekli (You ve ark., 2014b)	0.8491	0.8324	0.8612	0.7440	7.3
MMI (Ding ve ark., 2016)	0.8542	0.8522	0.8770	0.7071	6.0
NMBAC (Ding ve ark., 2016)	0.8559	0.8333	**0.8953**	0.7135	5.3
MMI+NMBAC (Ding ve ark., 2016)	0.8759	0.8661	0.8823	0.7524	3.0
Önerilen Metot (WeSeCT) (Göktepe ve Kodaz, 2018)	**0.8915**	0.8813	0.8729	0.7721	2.3

Bu sonuçlar önerilen metodun sınıflandırma doğruluğu (Acc) değerlendirme ölçütüne göre listedeki en iyi değere sahip olduğunu göstermektedir. Hassaslık (Sen), özgüllük (Spe), kesinlik (Pre) ve MCC değerlerinde ise en iyi sonuçlara yakın değerler elde edilmiştir. Bu sonuç tüm ölçütlere göre yapılan rank analizinden de görülebilmektedir.

Modelin *human* veri kümesinde elde ettiği sonuçlar Çizelge 5.8'de listelenmiştir. Bu sonuçlar, kıyaslanan çalışmalardan daha başarılı sonuçların elde edildiğini göstermiştir. Ölçüt değerlerinin ilgili çalışmalardan elde edilemediği durumlar "-" ile belirtilmiştir.

Çizelge 5.8. Önerilen metodun tahmin performansının, *human* veri kümesi için doğruluk (Acc), duyarlılık (Sen) ve kesinlik (Pre) ölçütleri üzerinden önceki çalışmalarla karşılaştırılması

Tahmin Modeli	Acc	Sen	Pre	Rank
HKNN (Nanni, 2005a)	0.6300	0.6400	0.6300	4.3
Signature products (Martin ve ark., 2005)	0.7030	0.6620	0.7220	2.3
Sınıflayıcıların Birleşimi (Nanni, 2005b)	0.6000	0.5900	0.6000	5.3
Birleşik HKNN (Nanni ve Lumini, 2006)	0.7000	0.7080	0.6700	2.7
FUS1 (Nanni ve ark., 2014)	0.6500	-	-	4.0
Önerilen Metot (WeSeCT) (Göktepe ve Kodaz, 2018)	**0.7381**	**0.7324**	**0.7411**	1.0

Model *HPRD* veri kümesi üzerinde de çalıştırılmıştır. Elde edilen sonuçlar Çizelge 5.9'da verilmiştir. İlgili çalışmalardan elde edilemeyen ölçüt değerleri "-" ile belirtilmiştir.

Çizelge 5.9. Önerilen metodun tahmin performansının, *HPRD* veri kümesi için doğruluk (Acc), duyarlılık (Sen), kesinlik (Pre) ve Mcc ölçütleri üzerinden önceki çalışmalarla karşılaştırılması

Tahmin Modeli	Acc	Sen	Pre	Mcc
You ve ark. (You ve ark., 2014a)	0.8480	0.8408	0.8547	0.7422
Karma bayes modeli (Xu ve ark., 2011)	0.8000	-	0.8000	-
Genom ve korunum bilgisi (Emamjomeh ve ark., 2014)	0.8100	-	-	-
Önerilen Metot (WeSeCT) (Göktepe ve Kodaz, 2018)	**0.9345**	**0.8929**	**0.8984**	**0.8571**

Önerilen yöntemin elde ettiği sonuçlar AUC ölçütü üzerinden de değerlendirilmiştir. *Helicobacter pylori* veri kümesi için 0.9371 AUC değeri elde edilmiştir. Önceki metotların sonuçlarıyla karşılaştırılınca en iyi ikinci değerin elde edildiği Çizelge 5.10'da görülmektedir.

Çizelge 5.10. Önerilen metodun tahmin performansının, *Helicobacter pylori* veri kümesi için AUC ölçütü üzerinden önceki çalışmalarla karşılaştırılması

Tahmin Modeli	AUC
Ağırlıklı seyrek (Huang ve ark., 2015)	0.8985
WSRC + GE (Huang ve ark., 2016)	**0.9375**
Dalgacık (Nanni ve ark., 2012)	0.9160
HPS (Nanni ve ark., 2010)	0.9250
Önerilen Metot (WeSeCT) (Göktepe ve Kodaz, 2018)	0.9371

Human veri kümesi için elde edilen 0.8320 AUC değerinin, Çizelge 5.11'de listelendiği gibi önceki çalışmalardan daha iyi olduğu görülmektedir.

Çizelge 5.11. Önerilen metodun tahmin performansının, *Human* veri kümesi için AUC ölçütü üzerinden önceki çalışmalarla karşılaştırılması

Tahmin Modeli	AUC
PSSM(PP) (Nanni ve ark., 2014)	0.8120
FUS1 (Nanni ve ark., 2014)	0.8200
HPS (Nanni ve ark., 2010)	0.7170
Dalgacık (Nanni ve ark., 2012)	0.6750
Önerilen Metot (WeSeCT) (Göktepe ve Kodaz, 2018)	**0.8320**

Metot, *HPRD* veri tabanı için 0.9300 AUC değerine ulaşmıştır. Çizelge 5.12'de belirtildiği gibi bu sonuç önceki çalışmaların elde ettiği değerlerin üzerindedir.

Çizelge 5.12. Önerilen metodun tahmin performansının, *HPRD* veri kümesi için AUC ölçütü üzerinden önceki çalışmalarla karşılaştırılması

Tahmin Modeli	AUC
Eş-evrimsel ayrışma (Liu ve ark., 2013)	0.7000
PPlevo (Zahiri ve ark., 2013b)	0.7700
Karma bayes modeli (Xu ve ark., 2011)	0.8000
LocFuse (Zahiri ve ark., 2014)	0.8500
You ve ark. (You ve ark., 2014a)	0.9232
Önerilen Metot (WeSeCT) (Göktepe ve Kodaz, 2018)	**0.9300**

5.3. Elde Edilen Sonuçların Yorumlanması

Önerilen ilk yöntem (ağırlıklandırılmış psödo aminoasit kompozisyonu) ile geliştirilen etkileşim tahmin sistemi, *Human, Helicobacter Pylori, Gram, Gpcr, Viral* ve *Membrane* veri kümeleri üzerinde test edilmiştir. Yöntem, farklı özellik çıkarım metotlarını kullanan yöntemlerle ROC eğrisi altında kalan alan ölçütü kullanılarak kıyaslanmıştır. Geliştirilen sistem, Çizelge 5.4'te görüldüğü gibi önceki yöntemlerle kıyaslandığında oldukça başarılı sonuçlara ulaşılmıştır. Bu sonuçlar *human, Gpcr* ve *viral* veri kümeleri için sırasıyla 0.735, 0.993 ve 0.811 şeklinde oluşmuş olup elde edilen en iyi sonuçlardır. *Gram* veri kümesi için 0.929 değerine ulaşılmıştır. Bu sonuç listedeki en yüksek sonucu veren yöntemle aynı değerdedir. *Helicobacter pylori ve membrane* veri kümelerinde elde edilen 0.922 ve 0.944 değerleri ise listedeki en iyi sonucu alan yöntemlere oldukça yakındır.

Önerilen sistem, PseAAC yöntemini (Chou, 2001) temel aldığı için, performansı bir fizyokimyasal özellik kullanarak bu yöntemle karşılaştırılmıştır. Bu amaçla, *human* veri kümesi üzerinde, tek bir fizyokimyasal özellik seçilerek PseAAC yöntemi ile kıyaslama yapılmıştır. Bu kıyaslama ROC eğrilerinin çizimi ve eğri altında kalan alanın hesaplanması şeklinde olmuştur. Şekil 5.1 ve Şekil 5.2'de verildiği gibi, proteomik

alanındaki çalışmalarda sık kullanılan "hidrofobik moment yönü" ve "hidrofilisite ölçeği" fizyokimyasal özellikleri üzerinden yapılan ROC çizimleri, önerilen metodun PseAAC metodundan daha iyi grafikler sunduğunu göstermektedir. "Hidrofobik moment yönü" özelliği için elde edilen AUC değeri kıyaslanan metottan %4.6, "hidrofilisite ölçeği" özelliği için ise %2.8 daha büyüktür.

Önerilen ikinci yöntem (Sekans Tabanlı Bir Birleşik Metot Kullanarak Protein-Protein Etkileşim Tahmini) ile geliştirilen etkileşim tahmin sistemi *helicobacter pylori, human* ve *HPRD* veri kümeleri üzerinde sınıflandırma doğruluğu (Acc), hassaslık (Sen), özgüllük (Spe), kesinlik (Pre) ve MCC ölçütleri ile test edilmiştir. Çizelge 5.7'de görüldüğü gibi *helicobacter pylori* veri kümesi için 0.8915 sınıflandırma doğruluğu değeri elde edilmiştir. Bu, tablodaki en yüksek doğruluk değeridir. Hassaslık ölçütünde alınan 0.8813 kesinlik ölçütünde alınan 0.8729 ve MCC ölçütündeki 0.7721 oranları ise en iyi sonuca yakın değerlerdir. Aynı çizelgede yöntemlerin rank analizleri de yapılmıştır. Rank analizi sonucunda önerilen yöntemin en iyi sıralamaya sahip olduğu görülmüştür. Sistem *human* veri kümesi ile çalıştırıldığında Çizelge 5.8'de görüldüğü gibi sınıflandırma doğruluğu, hassaslık, kesinlik ve MCC ölçütleri için sırasıyla 0.7381, 0.7324, 0.7411 ve 0.7279 değerlerine ulaşılmıştır. Bu değerlerin önceki çalışmalarla kıyaslandığında oldukça iyi sonuçlar olduğu açıktır. *HPRD* veri kümesi için de aynı ölçüt türleri için önceki çalışmalardan daha başarılı olan 0.9345, 0.8929, 0.8984 ve 0.8571 değerleri elde edilmiştir. Bu değerler Çizelge 5.9'da gösterilmiştir.

Geliştirilen etkileşim tahmin sistemi ROC eğrisi altındaki alan (AUC) ölçütü ile de test edilmiştir. *Helicobacter pylori* veri kümesi için 0.9371 AUC değeri elde edilmiştir. Bu değer ise Çizelge 5.10'da verilen listedeki yöntemler arasına en iyi sonucu veren ikinci yöntem olmuştur. Çizelge 5.11'de görüldüğü gibi *human* veri kümesi için 0.8320 AUC değeri elde edilmiştir. Bu değer karşılaştırılan metotlardan daha yüksektir. 0.9300 AUC değerine ulaşılan *HPRD* veri kümesi ile yapılan test, Çizelge 5.12'de görüldüğü gibi diğer metotlardan daha başarılıdır.

Bu çalışmada önerilen iki etkileşim tahmin metodunun performanslarının karşılaştırılması Çizelge 5.13 ve Çizelge 5.14'te verilmiştir. Bu sonuçlara bakıldığında, tüm değerlendirme ölçütlerinde, önerilen ikinci yöntemin (WeSeCT-Sekans Tabanlı Bir Birleşik Metot Kullanarak Protein-Protein Etkileşim Tahmini) ilk yöntemden (WePse-Protein-Protein Etkileşimlerinin Ağırlıklandırılmış Bir Psödo Aminoasit Kompozisyonu Tabanlı Yöntemle Tahmini) daha başarılı olduğu görülmektedir. Bu sonuçlara bakarak,

ikinci yöntemde uygulanan protein sekanslarından özellik çıkarımı adımının ilk yöntemde uygulanan adımdan daha anlamlı veriler çıkardığı sonucuna ulaşılabilir.

Çizelge 5.13. Önerilen iki metodun *Human* veri kümesi için elde edilen doğruluk (Acc), duyarlılık (Sen), özgüllük (Spe), kesinlik (Pre) ve eğri altındaki alan (AUC) değerleri

	Değerlendirme Ölçütleri				
Önerilen Metotlar	Acc	Sen	Spe	Pre	AUC
WePse (Goktepe ve ark., 2016)	0.6900	0.6800	0.6900	0.6800	0.7348
WeSeCT (Göktepe ve Kodaz, 2018)	**0.7381**	**0.7324**	**0.7463**	**0.7411**	**0.8320**

Çizelge 5.14. Önerilen iki metodun *Helicobacter Pylori* veri kümesi için elde edilen doğruluk (Acc), duyarlılık (Sen), özgüllük (Spe), kesinlik (Pre) ve eğri altındaki alan (AUC) değerleri

	Değerlendirme Ölçütleri				
Önerilen Metotlar	Acc	Sen	Spe	Pre	AUC
WePse (Goktepe ve ark., 2016)	0.8100	0.7900	0.8300	0.8000	0.9221
WeSeCT (Göktepe ve Kodaz, 2018)	**0.8915**	**0.8813**	**0.8996**	**0.8729**	**0.9371**

6. SONUÇLAR VE ÖNERİLER

6.1. Sonuçlar

Protein-protein etkileşimlerinin önemi her geçen gün daha fazla vurgulanmakta ve bu etkileşimlerin tahmin edilmesi zaman içinde daha fazla araştırmacının çalıştığı bir biyoinformatik konusu olmaya devam etmektedir. Bu tez çalışmasında protein-protein etkileşimlerinin tespit edilmesi için etkileşim tahmin sistemleri geliştirilmiştir. Önerilen istemlerinde tahmin sınıflandırma başarısını arttırmak amaçlanmıştır. Bu amaç için farklı özellik seçim aşamalarını kullanan DVM tabanlı etkileşim tahmin modelleri geliştirilmiştir.

Bu tez çalışmasında ilk olarak proteinlerin yapısı hakkında bilgi verilmiştir. Proteinlerin yapılarında bulunabilen kimyasal bağlar incelenmiştir. Diğer proteinlerle ya da moleküllerle yapılan etkileşimlerin iç yüzünün anlaşılması için etkileşim türleri incelenmiştir. Etkileşimlerin tespit yöntemleri hakkında bilgiler verilmiştir. Hesapsal etkileşim tahmin sistemlerinde kullanılan farklı özellik çıkarım yöntemleri incelenmiştir. Bu çalışmalarda yaygın olarak bilinen ve en çok kullanılan veri tabanlarının içerikleri, yapıları ve kullanımları anlatılmıştır.

Önerdiğimiz metotlarda, etkileşim tahmin sistemlerinin başarısının artırılabilmesinin etkili bir özellik çıkarımı aşamasıyla gerçekleştirilebileceğinin üzerinde durulmuştur. Veri tabanlarından elde edilen protein sekanslarına ait veriler üzerinde 2 farklı özellik çıkarımı adımı geliştirilmiştir. Bu şekilde protein sekanslarının etkileşim tahmin modelinde daha iyi temsil edilebilmesi sağlanmıştır.

Önerilen ilk metotta, ağırlıklandırılmış aminoasit kompozisyonu mantığını kullanan bir özellik çıkarımı yöntemi kullanılmaktadır (WePse). Bu özellik çıkarımının mevcut yöntemlerden farkı, protein sekansında daha çok tekrar eden aminoasit rezidülerinin bir etkileşimin sebebi olma ihtimalinin diğerlerine göre daha yüksek olacağını göz önünde bulundurmasıdır. Bu yüzden daha fazla tekrar eden rezidülere sınıflandırma sistemine katkılarının daha yüksek olması için daha yüksek temsil skorları atanmaktadır. Bu şekilde etkileşim tahmin sisteminin doğruluğu arttırılmıştır. Önerilen bu ilk yöntemin sonuçlarını önceki çalışmalarla karşılaştırabilmek amacıyla özellik çıkarımından sonra elde edilen veri kümesi, eğitim ve test olmak üzere iki bölüme ayrılmıştır.

Önerilen ikinci metotta farklı özellik çıkarımı metotlarının birlikte kullanımından oluşan bir özellik çıkarımı yöntemi geliştirilmiştir (WeSeCT). Bu yöntemin mevcut yöntemlerden farkı, birleşik üçlü metodunun sıra-atlamalı olarak değiştirilmesidir. Bu şekilde birleşik üçlü metodunda değerlendirilen aminoasit üçlü gruplarının frekansları bulunurken bir sıra atlayarak oluşan frekanslar da değerlendirilmektedir. Bu şekilde etkileşimin sebebi olabilecek grupların tespiti daha etkili bir şekilde yapılabilmektedir. Frekans değerlerinin hesaplanmasında sık tekrar eden gruplara daha yüksek temsil skorlarının verilmesiyle bir ağırlıklandırma işlemi de uygulanmaktadır.

Kullanılan özellik çıkarım aşamalarından elde edilen verileri işleyip etkileşim tahmini yapan bir sınıflandırma sistemi kurularak önerilen metotlar tamamlanmaktadır. Bu sınıflandırma sistemi biyoinformatik, sınıflandırma ve örüntü tanıma problemlerinde yoğun olarak kullanılan DVM tabanlı bir modeldir. Veri kümesindeki değerler arasında bir örüntü olmadığı durumlarda da etkili bir sınıflandırma imkânı sağladığı için DVM tabanlı bir model kullanılmıştır.

Önerilen metotlar ile etkileşim tahminlerine ilişkin başarılı sonuçlar elde edilmiştir. Değerlendirme ölçütleri olarak doğruluk (Acc), kesinlik (Pre), hassaslık (Sen), özgüllük (Spe), MCC katsayısı ve AUC (eğri altındaki alan) değerleri kullanılmıştır. Farklı veri tabanları üzerinde elde edilen sonuçlar önceki çalışmalarla karşılaştırmalı olarak gösterilmiştir. Bu sonuçlara göre önerilen yöntemlerin bazı veri tabanları için en iyi sonuçları verdiği ve bazı veri tabanları için en iyi sonuçlardan birisine ulaştığı görülmüştür. Önerilen modellerin rank analizleri yapılarak önceki metotlara göre başarı durumu ölçülmüştür. Elde edilen değerler, geliştirilen özellik çıkarım aşamalarının etkileşim tahmin sistemlerinin sınıflandırma başarısını arttırdığı sonucuna ulaşmamızı sağlamıştır.

6.2. Öneriler

Literatürdeki çalışmalarda, bilinen etkileşim sayılarının henüz sınırlı sayıda olduğu belirtilmekte ve bu konuda hala yeni çalışmalara ihtiyaç duyulduğu ifade edilmektedir (Stumpf ve ark., 2008). Örnek olarak insan PPE'lerinin henüz çok sınırlı bir kısmının bilindiği sanılmaktadır. Bir başka açıdan, bilinen PPE'lerin insan proteinlerinin üçte ikilik bir kısmı arasında gerçekleştiği ve proteinlerin yaklaşık üçte biri hakkında elimizde bilinen bir etkileşim bilgisinin bulunmadığı belirtilmektedir

(Kotlyar ve ark., 2015). Bu sebeplerle hala etkileşim tahmin sistemlerinin doğruluğunu artıracak çalışmaların yapılması gerekmektedir.

Bu çalışmaların en önemli hedeflerinden birisi sınıflandırma başarısını arttırmak olarak görülebilir. Diğer önemli bir hedef ise bu konudaki veri kümelerinin doğal olarak çok büyük hacimli olması ve proteinlere ilişkin çok farklı özelliklerden veri alınabilmesinin getirdiği karmaşıklığı azaltabilecek tahmin sistemlerinin üretilmesi olarak düşünülebilir.

Literatürdeki PPE tahmin çalışmalarının büyük bir kısmı proteinler hakkında her durumda elde edilemeyecek ya da elde edilmesi yoğun çaba gerektiren bir takım özelliklere ihtiyaç duymaktadır. Bu çalışmada önerilen PPE tahmin sistemleri protein sekansları üzerinden elde edilen bilgilerin kullanımıyla çalıştığı için herhangi bir veri kümesine uygulanabilme özelliğine sahiptir. Bu bilgilerin elde edilmesi için herhangi bir karmaşık fonksiyona ihtiyaç olmamakta ve protein sekansına ulaşılması yeterli olmaktadır. Bu açıdan önerilen sistemler, işlem karmaşıklığı ve sonuç verme hızı açılarından daha başarılıdır.

Bu çalışmada önerilen etkileşim tahmin sistemleri, protein sekanslarından bilgi elde etmek için farklı özellik çıkarım yöntemlerini kullanmaktadır. Bu özellik çıkarım yöntemleri ile eldeki PPE veri kümelerindeki etkileşim kuran proteinlere ilişkin daha anlamlı özellik değerleri üretilmektedir. Sekans üzerinden elde edilen bu veriler bir etkileşimin sebebi olabilecek özellikleri daha ön plana çıkarmaktadır. Bu şekilde PPE tahmin sistemlerinin doğruluğunun arttırılması sağlanmaktadır.

Gelecek çalışma olarak, derin öğrenme sistemlerinin kullanılması değerlendirilebilir. Derin öğrenme sistemleri makine öğrenmesi yöntemlerinin yapay sinir ağı tabanlı bir parçasıdır. Büyük miktardaki verilerin anlamlı ve daha küçük miktarlara başarıyla dönüştürülebilmesine imkân tanımaktadır. Bu sebeple, önerilen sistemlerinin derin öğrenme yöntemleri ile geliştirilmesi etkileşim tahmin sistemlerinin sınıflandırma başarılarını arttırabilir.

KAYNAKLAR

An, J. Y., Meng, F. R., You, Z. H., Chen, X., Yan, G. Y. ve Hu, J. P., 2016, Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model, *Protein Science*, 25 (10), 1825-1833.

Ayhan, S. ve Erdoğmuş, Ş. J. E. O. Ü. İ. v. İ. B. D., 2014, Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi, 9 (1), 175-201.

Bahar, I., Atilgan, A. R., Jernigan, R. L. ve Erman, B., 1997, Understanding the recognition of protein structural classes by amino acid composition, *Proteins-Structure Function and Genetics*, 29 (2), 172-185.

Ben-Hur, A. ve Noble, W. S., 2005, Kernel methods for predicting protein-protein interactions, *Bioinformatics*, 21, I38-I46.

Ben-Hur, A. ve Weston, J., 2010, A user's guide to support vector machines, *Methods Mol Biol*, 609, 223-239.

Berggard, T., Linse, S. ve James, P., 2007, Methods for the detection and analysis of protein-protein interactions, *Proteomics*, 7 (16), 2833-2842.

Berman, H., Henrick, K. ve Nakamura, H., 2003, Announcing the worldwide Protein Data Bank, *Nature Structural Biology*, 10 (12), 980-980.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. ve Bourne, P. E., 2000, The Protein Data Bank, *Nucleic Acids Research*, 28 (1), 235-242.

Bobroff, V., Chen, H. H., Javerzat, S. ve Petibois, C., 2016, What can infrared spectroscopy do for characterizing organic remnant in fossils?, *Trac-Trends in Analytical Chemistry*, 82, 443-456.

Bock, J. R. ve Gough, D. A., 2003, Whole-proteome interaction mining, *Bioinformatics*, 19 (1), 125-134.

Boughorbel, S., Jarray, F. ve El-Anbari, M., 2017, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, *Plos One*, 12 (6).

Bourne, P. E. ve Weissig, H., 2003, Structural bioinformatics, 44, *Hoboken, N.J.*, Wiley-Liss, p.

Bradley, A. P., 1997, The use of the area under the roc curve in the evaluation of machine learning algorithms, *Pattern Recognition*, 30 (7), 1145-1159.

Braun, P. ve Gingras, A. C., 2012, History of protein-protein interactions: From egg-white to complex networks, *Proteomics*, 12 (10), 1478-1498.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. ve Haussler, D., 2000, Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proceedings of the National Academy of Sciences of the United States of America*, 97 (1), 262-267.

Chang, C. C. ve Lin, C. J., 2011, LIBSVM: A Library for Support Vector Machines, *Acm Transactions on Intelligent Systems and Technology*, 2 (3).

Chaurasia, S., 2014, IN SILICO STUDY OF PROTEIN PROTEIN INTERACTION STABILIZATION AND MECHANICAL FORCE APPLICATION ON BIOMOLECULES.

Chen, X. W. ve Liu, M., 2005, Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics*, 21 (24), 4394-4400.

Chen, X. W. ve Jeong, J. C., 2009, Sequence-based prediction of protein interaction sites with an integrative method, *Bioinformatics*, 25 (5), 585-591.

Cho, K. I., Kim, D. ve Lee, D., 2009, A feature-based approach to modeling protein-protein interaction hot spots, *Nucleic Acids Res*, 37 (8), 2672-2687.

Chou, K. C., 2001, Prediction of protein cellular attributes using pseudo-amino acid composition, *Proteins-Structure Function and Genetics*, 43 (3), 246-255.

Chou, K. C., 2005, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, *Bioinformatics*, 21 (1), 10-19.

Chou, K. C. ve Shen, H. B., 2007, MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, *Biochemical and Biophysical Research Communications*, 360 (2), 339-345.

Chou, K. C., 2009, Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology, *Current Proteomics*, 6 (4), 262-274.

Chou, K. C., 2011, Some remarks on protein attribute prediction and pseudo amino acid composition, *Journal of Theoretical Biology*, 273 (1), 236-247.

Cohen, G. H., Silverton, E. W., Padlan, E. A., Dyda, F., Wibbenmeyer, J. A., Willson, R. C. ve Davies, D. R., 2005, Water molecules in the antibody-antigen interface of the structure of the Fab HyHEL-5-lysozyme complex at 1.7 A resolution: comparison with results from isothermal titration calorimetry, *Acta Crystallogr D Biol Crystallogr*, 61 (Pt 5), 628-633.

Consortium, T. U., 2018, UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research*, 47 (D1), D506-D515.

Cortes, C. ve Vapnik, V., 1995, Support-Vector Networks, *Machine Learning*, 20 (3), 273-297.

Crick, F. H. ve Orgel, L. E., 1964, The Theory of Inter-Allelic Complementation, *J Mol Biol*, 8, 161-165.

Cserhati, T. ve Szogyi, M., 1995, Role of Hydrophobic and Hydrophilic Forces in Peptide-Protein Interaction - New Advances, *Peptides*, 16 (1), 165-173.

Dandekar, T., Snel, B., Huynen, M. ve Bork, P., 1998, Conservation of gene order: a fingerprint of proteins that physically interact, *Trends in Biochemical Sciences*, 23 (9), 324-328.

Deng, L., Guan, J. H., Dong, Q. W. ve Zhou, S. G., 2009, Prediction of protein-protein interaction sites using an ensemble method, *Bmc Bioinformatics*, 10.

Ding, Y., Tang, J. ve Guo, F. J. B. b., 2016, Predicting protein-protein interactions via multivariate mutual information of protein sequences, 17 (1), 398.

Du, Q. S., Jiang, Z. Q., He, W. Z., Li, D. P. ve Chou, K. C., 2006, Amino Acid Principal Component Analysis (AAPCA) and its applications in protein structural class prediction, *Journal of Biomolecular Structure & Dynamics*, 23 (6), 635-640.

Dyson, H. J., Wright, P. E. ve Scheraga, H. A., 2006, The role of hydrophobic interactions in initiation and propagation of protein folding, *Proceedings of the National Academy of Sciences of the United States of America*, 103 (35), 13057-13061.

Emamjomeh, A., Goliaei, B., Torkamani, A., Ebrahimpour, R., Mohammadi, N. ve Parsian, A., 2014, Protein-protein interaction prediction by combined analysis of genomic and conservation information, *Genes & Genetic Systems*, 89 (6), 259-272.

Fedorov, A. N. ve Baldwin, T. O., 1997, Cotranslational protein folding, *Journal of Biological Chemistry*, 272 (52), 32715-32718.

Fields, S. ve Song, O. K., 1989, A Novel Genetic System to Detect Protein Protein Interactions, *Nature*, 340 (6230), 245-246.

Fossum, E., 2008, Herpesviral interactomics. Intraviral and virus-host protein-protein interaction network from different species of herpesviruses.

Freeman, S., 2005, Biological Science, *Upper Saddle River, NJ*, Pearson Prentice Hall, p.

Garg, A., Bhasin, M. ve Raghava, G. P. S., 2005, Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search, *Journal of Biological Chemistry*, 280 (15), 14427-14432.

Genfa, Z., Xinhua, X. ve Chun- Ting, Z., 1992, A weighting method for predicting protein structural class from amino acid composition, *European Journal of Biochemistry*, 210 (3), 747-749.

Goktepe, Y. E., Ilhan, I. ve Kahramanli, S., 2016, Predicting protein-protein interactions by weighted pseudo amino acid composition, *International Journal of Data Mining and Bioinformatics*, 15 (3), 272-290.

Göktepe, Y. E. ve Kodaz, H. J. N., 2018, Prediction of protein-protein interactions using an effective sequence based combined method, 303, 68-74.

Gribskov, M., Mclachlan, A. D. ve Eisenberg, D., 1987, Profile Analysis - Detection of Distantly Related Proteins, *Proceedings of the National Academy of Sciences of the United States of America*, 84 (13), 4355-4358.

Gu, Q., Ding, Y. S. ve Zhang, T. L., 2010, Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns, *Protein and Peptide Letters*, 17 (5), 559-567.

Guo, J. ve Sun, Z., 2005, Residue-couple Model for Protein Subcellular Loaclization Prediction, *Proceedings of the Third Asia-Pacific Bioinformatics Conference (APBC-2005), Singapore*, 117-129.

Guo, Y. Z., Yu, L. Z., Wen, Z. N. ve Li, M. L., 2008, Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences, *Nucleic Acids Research*, 36 (9), 3025-3030.

Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M. ve Mohabatkar, H., 2014, Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test, *Journal of Theoretical Biology*, 341, 34-40.

Han, G. S., Yu, Z. G. ve Anh, V., 2014, A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC, *Journal of Theoretical Biology*, 344, 31-39.

Hayat, M. ve Iqbal, N., 2014, Discriminating protein structure classes by incorporating Pseudo Average Chemical Shift to Chou's general PseAAC and Support Vector Machine, *Computer Methods and Programs in Biomedicine*, 116 (3), 184-192.

Hedin, S. G., 1906, Trypsin and Antitrypsin, *Biochem J*, 1 (10), 474-483.

Hirsimaki, T., Pylkkonen, J. ve Kurimo, M., 2009, Importance of High-Order N-Gram Models in Morph-Based Speech Recognition, *Ieee Transactions on Audio Speech and Language Processing*, 17 (4), 724-732.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L. Y., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D. ve Tyers, M., 2002, Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, *Nature*, 415 (6868), 180-183.

Holt, C., Raynes, J. K. ve Carver, J. A. J. B., 2019, Sequence characteristics responsible for protein- protein interactions in the intrinsically disordered regions of caseins, amelogenins, and small heat- shock proteins, e23319.

Hotelling, H., 1933, Analysis of a complex of statistical variables into principal components, *Journal of educational psychology*, 24 (6), 417-441.

HPRD, 2019, The Human Protein Reference Database, http://www.hprd.org/: [10.01.2019].

Hu, J. J. ve Zhang, F., 2009, Improving Protein Localization Prediction Using Amino Acid Group Based Physichemical Encoding, *Bioinformatics and Computational Biology, Proceedings*, 5462, 248-258.

Huang, Y. A., You, Z. H., Gao, X., Wong, L. ve Wang, L. R., 2015, Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to

Predict Protein-Protein Interactions from Protein Sequence, *Biomed Research International*.

Huang, Y. A., You, Z. H., Chen, X., Chan, K. ve Luo, X., 2016, Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding, *Bmc Bioinformatics*, 17.

Hue, M., Riffle, M., Vert, J. P. ve Noble, W. S., 2010, Large-scale prediction of protein-protein interactions from structures, *Bmc Bioinformatics*, 11.

Jeong, H., Mason, S. P., Barabasi, A. L. ve Oltvai, Z. N., 2001, Lethality and centrality in protein networks, *Nature*, 411 (6833), 41-42.

Jiang, W. ve Samanthula, B. K., 2011, N-Gram Based Secure Similar Document Detection, *Data and Applications Security and Privacy Xxv*, 6818, 239-246.

Jmol, 2018, Jmol: an open-source Java viewer for chemical structures in 3D, http://www.jmol.org: [29.10.2018].

Joachims, T., 1998, Text categorization with support vector machines: Learning with many relevant features, *European conference on machine learning*, 137-142.

Joachims, T., 1999, Making large-scale support vector machine learning practical, In: Advances in kernel methods, Eds, *Cambridge MA USA*: MIT Press, p. 169-184.

Kawashima, S., Ogata, H. ve Kanehisa, M., 1999, AAindex: Amino Acid Index Database, *Nucleic Acids Research*, 27 (1), 368-369.

Kawashima, S. ve Kanehisa, M., 2000, AAindex: Amino acid index database, *Nucleic Acids Research*, 28 (1), 374-374.

Kotlyar, M., Pastrello, C., Pivetta, F., Lo Sardo, A., Cumbaa, C., Li, H., Naranian, T., Ding, Z. Y., Vafaee, F., Broackes-Carter, F., Petschnigg, J., Mills, G. B., Jurisicova, A., Stagljar, I., Maestro, R. ve Jurisica, I., 2015, In silico prediction of physical protein interactions and characterization of interactome orphans, *Nature Methods*, 12 (1), 79-84.

Kumar, M., Verma, R. ve Raghava, G. P. S., 2006, Prediction of mitochondrial proteins using support vector machine and hidden Markov model, *Journal of Biological Chemistry*, 281 (9), 5357-5363.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J. ve Noble, W. S., 2004, Mismatch string kernels for discriminative protein classification, *Bioinformatics*, 20 (4), 467-476.

Li, B. Q., Huang, T., Liu, L., Cai, Y. D. ve Chou, K. C., 2012, Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network, *Plos One*, 7 (4).

Li, Z. C., Zhou, X. B., Dai, Z. ve Zou, X. Y., 2009, Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis, *Amino Acids*, 37 (2), 415-425.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L. ve Cesareni, G., 2012, MINT, the molecular interaction database: 2012 update, *Nucleic Acids Research*, 40 (D1), D857-D861.

Lin, C., Chen, W. Q., Qiu, C., Wu, Y. F., Krishnan, S. ve Zou, Q., 2014, LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy, *Neurocomputing*, 123, 424-435.

Liu, C. H., Li, K. C. ve Yuan, S. S., 2013, Human protein-protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence, *Bioinformatics*, 29 (1), 92-98.

Liu, M., Chen, X. W. ve Jothi, R., 2009, Knowledge-guided inference of domain-domain interactions from incomplete protein-protein interaction networks, *Bioinformatics*, 25 (19), 2492-2499.

Liu, P., Yang, L., Shi, D. M. ve Tang, X. L., 2015, Prediction of Protein-Protein Interactions Related to Protein Complexes Based on Protein Interaction Networks, *Biomed Research International*.

Ma, J. W. ve Gu, H., 2010, A novel method for predicting protein subcellular localization based on pseudo amino acid composition, *Bmb Reports*, 43 (10), 670-676.

Martin, S., Roe, D. ve Faulon, J. L., 2005, Predicting protein-protein interactions using signature products, *Bioinformatics*, 21 (2), 218-226.

Maruta, N., Trusov, Y. ve Botella, J. R., 2016, Yeast Three-Hybrid System for the Detection of Protein-Protein Interactions, *Methods Mol Biol*, 1363, 145-154.

Medline, 2019, ABD Ulusal Tıp Kütüphanesi, https://www.ncbi.nlm.nih.gov/pubmed:

Memisevic, V., Wallqvist, A. ve Reifman, J., 2013, Reconstituting protein interaction networks using parameter-dependent domain-domain interactions, *Bmc Bioinformatics*, 14.

Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H. G., Nagini, M., Kumar, G. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S. ve Pandey, A., 2006, Human protein reference database--2006 update, *Nucleic Acids Res*, 34 (Database issue), D411-414.

Mondal, S. ve Pai, P. P. J. J. o. t. b., 2014, Chouʹ s pseudo amino acid composition improves sequence-based antifreeze protein prediction, 356, 30-35.

Moore, R. C. ve Lewis, W., 2010, Intelligent selection of language model training data, *Proceedings of the ACL 2010 conference short papers*, 220-224.

Nakai, K., Kidera, A. ve Kanehisa, M., 1988, Cluster analysis of amino acid indices for prediction of protein structure and function, *Protein Eng*, 2 (2), 93-100.

Nanni, L., 2005a, Hyperplanes for predicting protein-protein interactions, *Neurocomputing*, 69 (1-3), 257-263.

Nanni, L., 2005b, Fusion of classifiers for predicting protein-protein interactions, *Neurocomputing*, 68, 289-296.

Nanni, L. ve Lumini, A., 2006, An ensemble of K-local hyperplanes for predicting protein-protein interactions, *Bioinformatics*, 22 (10), 1207-1210.

Nanni, L., Brahnam, S. ve Lumini, A., 2010, High performance set of PseAAC and sequence based descriptors for protein classification, *Journal of Theoretical Biology*, 266 (1), 1-10.

Nanni, L., Brahnam, S. ve Lumini, A., 2012, Wavelet images and Chou's pseudo amino acid composition for protein classification, *Amino Acids*, 43 (2), 657-665.

Nanni, L., Lumini, A. ve Brahnam, S., 2014, An Empirical Study of Different Approaches for Protein Classification, *Scientific World Journal*.

NHGRI, 2019, Protein, https://www.genome.gov/:

Nooren, I. M. A. ve Thornton, J. M., 2003, Diversity of protein-protein interactions, *Embo Journal*, 22 (14), 3486-3492.

Okada, T., Fujiyoshi, Y., Silow, M., Navarro, J., Landau, E. M. ve Shichida, Y., 2002, Functional role of internal water molecules in rhodopsin revealed by x-ray crystallography, *Proceedings of the National Academy of Sciences of the United States of America*, 99 (9), 5982-5987.

Ooi, S. L., Pan, X. W., Peyser, B. D., Ye, P., Meluh, P. B., Yuan, D. S., Irizarry, R. A., Bader, J. S., Spencer, F. A. ve Boeke, J. D., 2006, Global synthetic-lethality analysis and yeast functional profiling, *Trends in Genetics*, 22 (1), 56-63.

Oti, M., Snel, B., Huynen, M. A. ve Brunner, H. G., 2006, Predicting disease genes using protein-protein interactions, *Journal of Medical Genetics*, 43 (8).

Pan, X. Y., Zhang, Y. N. ve Shen, H. B., 2010, Large-Scale Prediction of Human Protein-Protein Interactions from Amino Acid Sequence Based on Latent Topic Features, *Journal of Proteome Research*, 9 (10), 4992-5001.

Pazos, F. ve Valencia, A., 2001, Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Engineering*, 14 (9), 609-614.

Pearson, 2018, Pearson, https://www.pearson.com: [08.12.2018].

Pearson, K., 1901, LIII. On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, Dublin Philosophical Magazine Journal of Science*, 2 (11), 559-572.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. ve Yeates, T. O., 1999, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc Natl Acad Sci U S A*, 96 (8), 4285-4288.

Pelletier, J. N., Arndt, K. M., Pluckthun, A. ve Michnick, S. W., 1999, An in vivo library-versus-library selection of optimized protein-protein interactions, *Nature Biotechnology*, 17 (7), 683-690.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K. B., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z. X., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K.,

Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L. L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G. N., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A. ve Pandey, A., 2003, Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Research*, 13 (10), 2363-2371.

Phizicky, E. M. ve Fields, S., 1995, Protein-Protein Interactions - Methods for Detection and Analysis, *Microbiological Reviews*, 59 (1), 94-123.

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B. H., Vreven, T. ve Weng, Z. P., 2014, ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers, *Bioinformatics*, 30 (12), 1771-1773.

Pontil, M. ve Verri, A., 1998, Support Vector Machines for 3D object recognition, *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 20 (6), 637-646.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. ve Pandey, A., 2009, Human Protein Reference Database-2009 update, *Nucleic Acids Research*, 37, D767-D772.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M. ve Seraphin, B., 2001, The tandem affinity purification (TAP) method: A general procedure of protein complex purification, *Methods*, 24 (3), 218-229.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A. ve Legrain, P., 2001, The protein-protein interaction map of Helicobacter pylori, *Nature*, 409 (6817), 211-215.

Rao, V. S., Srinivas, K., Sujini, G. N. ve Kumar, G. N., 2014, Protein-protein interaction detection: methods and analysis, *Int J Proteomics*, 2014, 147648.

RCSBPDB, 2018, Research Collaboratory for Structural Bioinformatics Protein Data Bank, http://www.rcsb.org: [19.10.2018].

Reyes, J. A. ve Gilbert, D., 2007, Prediction of protein-protein interactions using one-class classification methods and integrating diverse biological data, *Journal of Integrative Bioinformatics*, 4 (3), 208-223.

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlic, A. ve Rose, P. W., 2018, NGL viewer: web-based molecular graphics for large complexes, *Bioinformatics*, 34 (21), 3755-3758.

Roy, S., Martinez, D., Platero, H., Lane, T. ve Werner-Washburne, M., 2009, Exploiting amino acid composition for predicting protein-protein interactions, *Plos One*, 4 (11), e7813.

Schuster-Bockler, B. ve Bateman, A., 2008, Protein interactions in human genetic diseases, *Genome Biol*, 9 (1), R9.

Sharma, A., Lyons, J., Dehzangi, A. ve Paliwal, K. K., 2013, A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition, *Journal of Theoretical Biology*, 320, 41-46.

Shen, H. B. ve Chou, K. C., 2007a, Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins, *Protein Eng Des Sel*, 20 (1), 39-46.

Shen, H. B. ve Chou, K. C., 2007b, Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells, *Biopolymers*, 85 (3), 233-240.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. ve Jiang, H., 2007, Predicting protein-protein interactions based only on sequences information, *Proc Natl Acad Sci U S A*, 104 (11), 4337-4341.

Shoemaker, B. A. ve Panchenko, A. R., 2007a, Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners, *PLoS Comput Biol*, 3 (4), e43.

Shoemaker, B. A. ve Panchenko, A. R., 2007b, Deciphering protein-protein interactions. Part I. Experimental techniques and databases, *PLoS Comput Biol*, 3 (3), e42.

Smith, G. R. ve Sternberg, M. J., 2002, Prediction of protein-protein interactions by docking methods, *Curr Opin Struct Biol*, 12 (1), 28-35.

Stumpf, M. P., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M. ve Wiuf, C., 2008, Estimating the size of the human interactome, *Proc Natl Acad Sci U S A*, 105 (19), 6959-6964.

Sun, T. L., Zhou, B., Lai, L. H. ve Pei, J. F., 2017, Sequence-based prediction of protein protein interaction using a deep-learning algorithm, *Bmc Bioinformatics*, 18.

Tantoso, E. ve Li, K. B., 2008, AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices, *Amino Acids*, 35 (2), 345-353.

Templin, M. F., Stoll, D., Schrenk, M., Traub, P. C., Vohringer, C. F. ve Joos, T. O., 2002, Protein microarray technology, *Drug Discovery Today*, 7 (15), 815-822.

Tian, B., Wu, X., Chen, C., Qiu, W., Ma, Q. ve Yu, B., 2019, Predicting protein–protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach, *Journal of Theoretical Biology*, 462, 329-346.

Tomii, K. ve Kanehisa, M., 1996, Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, *Protein Eng*, 9 (1), 27-36.

Tsuda, K., Shin, H. J. ve Scholkopf, B., 2005, Fast protein classification with multiple networks, *Bioinformatics*, 21, 59-65.

UniProt, 2019, https://www.uniprot.org/statistics/Swiss-Prot: [01.10.2018].

Vapnik, V., 2013, The nature of statistical learning theory, Springer science & business media, p.

Wang, H. ve Wu, P. J. B., 2018, Prediction of RNA-protein interactions using conjoint triad feature and chaos game representation, 9 (1), 242-251.

Wang, J., Zhang, L., Jia, L. Y., Ren, Y. Z. ve Yu, G. X., 2017, Protein-Protein Interactions Prediction Using a Novel Local Conjoint Triad Descriptor of Amino Acid Sequences, *International Journal of Molecular Sciences*, 18 (11).

Wang, T., Li, L. P., Huang, Y. A., Zhang, H., Ma, Y. H. ve Zhou, X., 2018, Prediction of Protein-Protein Interactions from Amino Acid Sequences Based on Continuous and Discrete Wavelet Transform Features, *Molecules*, 23 (4).

Wei, L. Y., Xing, P. W., Zeng, J. C., Chen, J. X., Su, R. ve Guo, F., 2017, Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier, *Artificial Intelligence in Medicine*, 83, 67-74.

Wikipedia, 2019, https://tr.wikipedia.org/wiki/Protein_biyosentezi:

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J. H., Bantscheff, M., Gerstmair, A., Faerber, F. ve Kuster, B., 2014, Mass-spectrometry-based draft of the human proteome, *Nature*, 509 (7502), 582-+.

Wuthrich, K., 1989, Protein-Structure Determination in Solution by Nuclear Magnetic-Resonance Spectroscopy, *Science*, 243 (4887), 45-50.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M. ve Eisenberg, D., 2000, DIP: the Database of Interacting Proteins, *Nucleic Acids Research*, 28 (1), 289-291.

Xia, J. F., Han, K. ve Huang, D. S., 2010a, Sequence-Based Prediction of Protein-Protein Interactions by Means of Rotation Forest and Autocorrelation Descriptor, *Protein and Peptide Letters*, 17 (1), 137-145.

Xia, J. F., Zhao, X. M. ve Huang, D. S., 2010b, Predicting protein-protein interactions from protein sequences using meta predictor, *Amino Acids*, 39 (5), 1595-1599.

Xiao, X., Wang, P. ve Chou, K. C., 2008, Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image, *Journal of Theoretical Biology*, 254 (3), 691-696.

Xu, Y., Hu, W., Chang, Z. Q., DuanMu, H. Z., Zhang, S. Z., Li, Z. Q., Li, Z. H., Yu, L. L. ve Li, X., 2011, Prediction of human protein-protein interaction by a mixed Bayesian model and its application to exploring underlying cancer-related pathway crosstalk, *Journal of the Royal Society Interface*, 8 (57), 555-567.

Yao, Y., Du, X., Diao, Y. ve Zhu, H. J. P., 2019, An integration of deep learning with feature embedding for protein–protein interaction prediction, 7, e7126.

You, Z. H., Lei, Y. K., Zhu, L., Xia, J. F. ve Wang, B., 2013, Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis, *Bmc Bioinformatics*, 14.

You, Z. H., Li, S., Gao, X., Luo, X. ve Ji, Z., 2014a, Large-Scale Protein-Protein Interactions Detection by Integrating Big Biosensing Data with Computational Model, *Biomed Research International*.

You, Z. H., Zhu, L., Zheng, C. H., Yu, H. J., Deng, S. P. ve Ji, Z., 2014b, Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set, *Bmc Bioinformatics*, 15.

Yu, C. Y., Chou, L. C. ve Chang, D. T. H., 2010, Predicting protein-protein interactions in unbalanced data using the primary structure of proteins, *Bmc Bioinformatics*, 11.

Zahiri, J., Bozorgmehr, J. H. ve Masoudi-Nejad, A., 2013a, Computational Prediction of Protein-Protein Interaction Networks: Algorithms and Resources, *Current Genomics*, 14 (6), 397-414.

Zahiri, J., Yaghoubi, O., Mohammad-Noori, M., Ebrahimpour, R. ve Masoudi-Nejad, A., 2013b, PPIevo: protein-protein interaction prediction from PSSM based evolutionary information, *Genomics*, 102 (4), 237-242.

Zahiri, J., Mohammad-Noori, M., Ebrahimpour, R., Saadat, S., Bozorgmehr, J. H., Goldberg, T. ve Masoudi-Nejad, A., 2014, LocFuse: Human protein-protein interaction prediction via classifier fusion using protein localization information, *Genomics*, 104 (6), 496-503.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A. ve Honig, B., 2012a, Structure-based prediction of protein-protein interactions on a genome-wide scale, *Nature*, 490 (7421), 556-560.

Zhang, S. L., Ye, F. ve Yuan, X. G., 2012b, Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM, *Journal of Biomolecular Structure & Dynamics*, 29 (6), 634-642.

Zhang, S. W., Chen, W., Yang, F. ve Pan, Q., 2008, Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach, *Amino Acids*, 35 (3), 591-598.

Zhang, S. W., Hao, L. Y. ve Zhang, T. H., 2014, Prediction of Protein-Protein Interaction with Pairwise Kernel Support Vector Machine, *International Journal of Molecular Sciences*, 15 (2), 3220-3233.

Zhu, W., Zeng, N. ve Wang, N., 2010, Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations, *NESUG proceedings: health care life sciences, Baltimore, Maryland*, 19, 67.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	Yunus Emre GÖKTEPE
Uyruğu	**:**	TC
Doğum Yeri ve Tarihi	**:**	Seydişehir, 16.06.1979
Telefon	**:**	0533 655 8728
Faks	**:**	0332 582 7609
e-mail	**:**	ygoktepe@erbakan.edu.tr, yegoktepe@gmail.com

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	Seydişehir Enis Şanlıoğlu Lisesi	1997
Üniversite	:	Selçuk Üniversitesi Mühendislik Mimarlık Fakültesi Bilgisayar Mühendisliği	2001
Yüksek Lisans	:	Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği ABD	2005
Doktora	:	Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği ABD	Devam ediyor

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2001-2013	Selçuk Üniversitesi	Öğretim Görevlisi
2008-2013	Selçuk Üniversitesi Bilgisayar Teknolojileri Bölümü	Bölüm Başkanlığı
2009-2012	Selçuk Üniversitesi	Müdür Yardımcılığı
2013-2019	Necmettin Erbakan Üniversitesi	Öğretim Görevlisi
2013-2019	Necmettin Erbakan Üniversitesi Bilgisayar Teknolojileri Bölümü	Bölüm Başkanlığı

UZMANLIK ALANI

Protein-Protein Etkileşimleri, Makine Öğrenmesi, Veri Madenciliği

YABANCI DİLLER

İngilizce

BELİRTMEK İSTEĞİNİZ DİĞER ÖZELLİKLER

YAYINLAR

Uluslararası SCI, SCI-Exp İndekslerinde Taranan Dergilerdeki Yayınlar

Göktepe, Y. E., İlhan, İ., ve Kahramanlı, Ş. (2016). Predicting protein-protein interactions by weighted pseudo amino acid composition. International Journal of Data Mining and Bioinformatics, 15(3), 272-290.

Göktepe, Y. E., ve Kodaz, H. (2018). Prediction of protein-protein interactions using an effective sequence based combined method. Neurocomputing, 303, 68-74.

Uluslararası Diğer İndekslerde Taranan Dergilerdeki Yayınlar

Turhan, T., **Göktepe, Y. E.**, ve Ayyıldız, N. (2017). Matlab Applications for Skew-Symmetric Matrices and Integral Curves in Lorentzian Spaces. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 5(2), 611-621.