

**BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ DOKTORA PROGRAMI**

**DERİN ÖĞRENME AĞLARI KULLANILARAK PROTEİN METAL
BAĞLANMA YERLERİNİN ANALİZİ**

**HAZIRLAYAN
İSMAİL HABERAL**

DOKTORA TEZİ

ANKARA - 2020

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ DOKTORA PROGRAMI

**DERİN ÖĞRENME AĞLARI KULLANILARAK PROTEİN METAL
BAĞLANMA YERLERİNİN ANALİZİ**

HAZIRLAYAN
İSMAİL HABERAL

DOKTORA TEZİ

TEZ DANIŞMANI
PROF. DR. HASAN OĞUL

ANKARA - 2020

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Doktora Programı çerçevesinde İsmail HABERAL tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Doktora Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 18/ 08/ 2020

Tez Adı: Derin Öğrenme Ağları Kullanılarak Protein Metal Bağlanma Yerlerinin Analizi

Tez Jüri Üyeleri

İmza

Prof. Dr. Hasan OĞUL, Başkent Üniversitesi

Prof. Dr. Mustafa DOĞAN, İstanbul Teknik Üniversitesi

Prof. Dr. Hamit ERDEM, Başkent Üniversitesi

Doç. Dr. Tunca DOĞAN, Hacettepe Üniversitesi

Dr. Öğr. Üyesi Mehmet DİKMEN, Başkent Üniversitesi

ONAY

Prof. Dr. Faruk ELALDI

Fen Bilimleri Enstitüsü Müdürü

Tarih : ... / ... /

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

Tarih: 25/ 08/ 2020

Öğrencinin Adı, Soyadı : İsmail HABERAL

Öğrencinin Numarası : 21220038

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Bilgisayar Mühendisliği Doktora Programı

Danışmanın Unvanı/Adı, Soyadı : Prof. Dr. Hasan OĞUL

Tez Başlığı : Derin Öğrenme Ağları Kullanılarak Protein Metal Bağlanma Yerlerinin Analizi

Yukarıda başlığı belirtilen Doktora tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 54 sayfalık kısmına ilişkin, 25/ 08/ 2020 tarihinde tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %5'dir. Uygulanan filtrelemeler:

1. Kaynakça hariç
2. Alıntılar hariç
3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

“Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını” inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:.....

ONAY

Tarih: 25/08/2020

Öğrenci Danışmanı Prof. Dr. Hasan OĞUL

TEŞEKKÜR

Yazar, bu çalışmanın gerçekleşmesinde katkılarından dolayı, aşağıda adı geçen kişi ve kuruluşlara içtenlikle teşekkür eder.

Sayın Prof. Dr. Hasan OĞUL'a (tez danışmanı), çalışmanın sonuca ulaştırılmasında ve karşılaşılan güçlüklerin aşılmasında her zaman yardımcı ve yol gösterici olduğu için...

Eğitim hayatım boyunca ve attığım her adım da hep yanımda olan babam merhum Recep HABERAL ve aileme...

Eşim Gözde KAHRAMAN HABERAL'a, bana çalışmam için zaman yarattığı, destek olduğu için...

Doktora yeterlik sınavı sabahı doğan oğlum Alptuğ HABERAL'a ve tez çalışmaları esnasında doğan ikinci oğlum Altuğ HABERAL'a, bana hayatı yeniden yorumlamayı öğrettikleri, onlara ayıracağım zamanın bir kısmını çalışmaya ayırma fırsatı verdikleri için...

ÖZET

İsmail HABERAL

DERİN ÖĞRENME AĞLARI KULLANILARAK PROTEİN METAL BAĞLANMA YERLERİNİN ANALİZİ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

2020

Proteinler çevrelerinde bulunan metal iyonlarıyla kuvvetli bağlar oluşturarak katlanırlar ve üç boyutlu yapılarına ulaşırlar. Proteinlerin üç boyutlu yapısı, hücre içerisinde hangi yaşamsal fonksiyonu yerine getirdiğini gösterir. Protein dizilimi kullanılarak proteinlerin metallere bağlanma durumunu tahmin etmek, proteinin yapısı, fonksiyonlarını tahmin etmek ve ilaç keşfi için önemlidir. Aminoasit dizilimlerinden elde edilen verilerden yola çıkarak ve makine öğrenmesi yöntemleri kullanılarak yapılan hesaplamalı tahminler çeşitli bioinformatik alanlarında yaygın olarak kullanılmaktadır.

Bu tez çalışmasında, protein dizilimlerinde bulunan Histidin (HIS) ve Sistein (CYS) aminoasitlerinin metallere bağlanma durumlarının tahmini için üç farklı derin öğrenme mimarisi önerilmektedir. Bu mimariler TensorFlow üzerinde çalışan Keras kullanılarak geliştirilmiştir. Bu mimariler sırasıyla evrimsel sinir ağı, uzun-kısa süreli hafıza ve kapalı tekrarlayan hücre modelleri üzerine inşa edilmiştir. Bu modeller doğrudan dizilim verileri üzerinde çalışmadığından, ilgili modelleri beslemek üzere PAM skorlama matrisi, protein kompozisyonları ve ikili temsil yöntemlerine dayalı sayısallaştırma teknikleri uygulanmıştır. Geliştirilen mimariler ve protein dizilimi sayısallaştırma yöntemleri 2727 proteinden oluşan kıyaslama veri kümesi üzerinde test edilmiştir. Elde edilen sonuçlar Naive Bayes, destek vektör makineleri (SVM), Adaboost ve Bagging makine öğrenme yöntemleri ile elde edilen sonuçlarla karşılaştırılmıştır. Protein metal bağlanma yeri tahmini için en iyi sonuçların evrimsel sinir ağı mimarisi ile elde edildiği görülmektedir. Bu sonuç, aynı veri kümesi ile literatürde var olan diğer çalışmalardan daha iyi başarımlar elde edildiğini göstermektedir. Elde edilen bu sonuçlar kullanılarak, bir metal iyonunun koordinasyonunda hangi kalıntıların birlikte yer aldığına karar vermek için metal bağlanma yerlerinin geometrisi değerlendirilmiştir.

ANAHTAR SÖZCÜKLER: derin öğrenme, evrimsel sinir ağları, metal bağlanma, metal bağlanma yeri geometrisi, protein analizi.

ABSTRACT

İsmail HABERAL

ANALYSIS OF PROTEIN METAL-BINDING SITES USING DEEP NEURAL NETWORKS

Başkent University Institute of Science and Engineering

Department of Computer Engineering

2020

Proteins fold by forming strong bonds with the metal ions in their environment and reach their three-dimensional structure. The three-dimensional structure of proteins shows which critical function it performs in the cell. Prediction of protein metal binding sites using protein sequence is important for predicting protein structure, functions, and drug discovery. Computational estimates using machine learning methods based on data from amino acid sequences are widely used in various bioinformatics fields.

In this thesis, three different deep learning architectures are proposed for the prediction of metal binding status of Histidine (HIS) and Cysteine (CYS) amino acids in protein sequences. These architectures are built on convolutional neural network (CNN), long-short term memory (LSTM) and gated recurrent unit (GRU) models, respectively. These architectures are developed using Keras with Tensorflow backend. Since these models cannot work directly on sequence data, digitization techniques based on PAM scoring matrix, protein compositions and binary representation methods have been applied to feed the relevant models. Developed architectures and protein sequence digitization methods have been tested on benchmark data set consisting of 2727 proteins. The results obtained were compared with the results obtained with Naïve Bayes, Support vector machines, Adaboost and Bagging machine learning methods. It seems that the best results for prediction of protein metal binding site are obtained with CNN architecture. This result shows that better performance was obtained with the same dataset than other studies in the literature. Using these results, the geometry of the metal binding sites was evaluated in order to decide which residues are involved in the coordination of a metal ion.

KEYWORDS: deep learning, convolutional neural networks, metal-binding, geometry of metal binding sites, analysis of protein.

İÇİNDEKİLER

TEŞEKKÜR.....	i
ÖZET.....	ii
ABSTRACT	iii
İÇİNDEKİLER.....	iv
TABLolar LİSTESİ.....	vi
ŞEKİLLER LİSTESİ.....	viii
SİMGELER VE KISALTMALAR LİSTESİ.....	x
1. GİRİŞ.....	1
1.1. Kapsam.....	1
1.2. Katkılar	2
1.3. Biyolojik Kavram ve Tanımlar	3
1.4. Metal Bağlanma Problemi.....	8
1.5. Önceki Çalışmalar.....	9
2. YÖNTEMLER.....	16
2.1. Veri Kümesi	16
2.2. Geleneksel Makine Öğrenmesi.....	16
2.2.1. Destek Vektör Makineleri (SVM).....	17
2.2.2. Naïve Bayes	17
2.2.3 AdaBoost	18
2.2.4. Bagging.....	19
2.3. Derin Öğrenme Ağları	19
2.3.1. Evrişimsel Sinir Ağları (CNN)	21
2.3.2. Kapılı Tekrarlayan Hücre (GRU)	25
2.3.3. Uzun-Kısa Süreli Bellek (LSTM).....	27

2.4. Öznitelik Çıkarma.....	28
2.4.1. Nokta Kabul Edilen Mutasyon (PAM) Skorlama Matrisi	29
2.4.2 Protein Kompozisyon Sunucusu (ProCos).....	31
2.4.3. İkili Temsil (BR).....	34
3. DENEYSEL BULGULAR.....	36
3.1. Derin Öğrenme Modelleri	36
3.1.1. CNN Modeli	37
3.1.2. GRU Modeli.....	39
3.1.3. LSTM Modeli	40
3.2. Metal Bağlanma Yeri Bulguları.....	41
3.2.1. PAM Matrisi ile Elde Edilen Veri Kümesi Bulguları	42
3.2.2. İkili Temsil ile Elde Edilen Veri Kümesi Bulguları	43
3.2.3. ProCos ile Elde Edilen Veri Kümesi Bulguları	44
3.2.4. Performan Karşılaştırması.....	45
3.3. Metal Bağlanma Yerlerinin Geometrik Tahimini	47
4. TARTIŞMA VE SONUÇ.....	52
KAYNAKLAR.....	55

EKLER

EK 1: Çalışmada Kullanılan Protein Listesi

EK 2: Çalışmada Kullanılan Metal Bağlı Protein Listesi

TABLÖLAR LİSTESİ

Sayfa

Tablo 1. 1. Sekans tabanlı metal bağlanma yeri çalışmaları.....	13
Tablo 1. 2. Yapı tabanlı metal bağlanma yeri çalışmaları.....	14
Tablo 1. 3. Metal bağlı protein veritabanları.....	15
Tablo 1. 4. Protein metal bağlanma yeri tahmini için geliştirilen yazılımlar.....	15
Tablo 2. 1. Veri Kümesi.....	16
Tablo 2. 2. Amino asitlerin ikili temsil olarak gösterilmesi.....	35
Tablo 3. 1. CNN modeli girdi kanalları sayısı ve çıktı sayıları.....	38
Tablo 3. 2. CNN mimarisinin devir sayısına bağlı çalışma süreleri ve sonuçları.....	38
Tablo 3. 3. CNN model parametreleri.....	39
Tablo 3. 4. GRU model parametreleri.....	39
Tablo 3. 5. LSTM model parametreleri.....	41
Tablo 3. 6. Derin öğrenme mimarilerinin 1 devir işlemi için ortalama çalışma süreleri.....	42
Tablo 3. 7. PAM öznitelikleri ve derin öğrenme modelleri ile elde edilen sonuçlar.....	43
Tablo 3. 8. PAM öznitelikleri ve makine öğrenme yöntemleri ile elde edilen sonuçlar.....	43
Tablo 3. 9. İkili temsil öznitelik kümesi ile elde edilen sonuçlar.....	43
Tablo 3. 10. PC öznitelik kümesi ile elde edilen sonuçlar.....	44
Tablo 3. 11. PAAC öznitelik kümesi ile elde edilen sonuçlar.....	44
Tablo 3. 12. 5FSS öznitelik kümesi ile elde edilen sonuçlar.....	45
Tablo 3. 13. Performans karşılaştırması.....	46
Tablo 3. 14. Veri kümesindeki Histidin ve Sistein bağlı metal sayıları.....	47
Tablo 3. 15. Tahmin sonucu bağlı metal sayıları.....	48

Tablo 3. 16. Histidin ve Sistein 'ın aynı anda bağlandığı metal sayıları	48
--	----



ŞEKİLLER LİSTESİ

Sayfa

Şekil 1. 1. Amino asit yapısı.	3
Şekil 1. 2. Amino asit listesi.....	4
Şekil 1. 3. Histidin amino asidinin 2-boyutlu, 3-boyutlu ve biyolojik yapıları.....	5
Şekil 1. 4. Sistein amino asidinin 2-boyutlu, 3-boyutlu ve Kristal yapıları	5
Şekil 1. 5. Protein birincil yapı.....	6
Şekil 1. 6. Protein ikincil yapı.	6
Şekil 1. 7. 1AXC proteininin üçüncül yapısı [5].....	6
Şekil 1. 8. 1AXC proteininin dördüncül yapı [5].....	7
Şekil 1. 9. Çinko metali ile m1 ve m2 noktalarından bağlı üç boyutlu 5X6T proteini (üstte), protein diziliminin bağlantı yerlerinin grafiksel gösteriminden bir kesit (altta).	8
Şekil 2. 1. Yapay Sinir Ağı.	17
Şekil 2. 2. Örnek SVM sınıflandırmaları.	17
Şekil 2. 3. AdaBoost sınıflandırma örneği.	18
Şekil 2. 4. Bagging algoritması.	19
Şekil 2. 5. Yapay zeka, makine öğrenmesi ve derin öğrenme.	20
Şekil 2. 6. Derin öğrenme ağ örneği.....	21
Şekil 2. 7. MNIST veritabanı özellik hiyerarşısı. Aşağıdan yukarıya görüntü temsili [61].....	21
Şekil 2. 8. Evrişim katmanında çıktı hesaplama örneği.	23
Şekil 2. 9. Maksimum havuzlama örneği.	24
Şekil 2. 10. Tam bağlantılı katman.	24
Şekil 2. 11. ReLU aktivasyon fonksiyonu.....	25

Şekil 2. 12. Bırakma katmanı: siyah renkli nöronlar unutulmuştur.	25
Şekil 2. 13. GRU	26
Şekil 2. 14. LSTM.	28
Şekil 2. 15. Metal bağlanma yeri tahmini akış diyagramı. a) Çinko metali ile bağlı (m1,m2) üç boyutlu 5X6T proteini. b) protein diziliminden örnek bir kesit. c) metal bağlı Sistein ve Histidin aminoasitleri için çerçeve çıkarma ve öznitelik vektörü. d) derin öğrenme mimarileri.	30
Şekil 2. 16. PAM120 Skorlama Matrisi.	31
Şekil 2. 17. PAM ile elde edilen öznitelik kümesinden 280 özellik içeren iki örnek.	32
Şekil 2. 18. PAAC tarafından kullanılan parametre değerleri.	33
Şekil 2. 19. PAAC ile elde edilen veri kümesinden 100 özellik içeren bir örnek.	34
Şekil 3. 1. CNN Modeli.	37
Şekil 3. 2. GRU Model.	40
Şekil 3. 3. LSTM Modeli.	40
Şekil 3. 4. Bütün çerçeve boyutları için CNN ve PAM kullanılarak elde edilen sonuçlar.	41
Şekil 3. 5. Performans karşılaştırması.	46
Şekil 3. 6. 1RMD proteini dizilimi (üstte), 2 boyutlu (ortada) ve 3 boyutlu yapısı (altta).	49
Şekil 3. 7. 1RMD proteini için metal bağlanma yerlerinden bir kesit.	50
Şekil 3. 8. 1RMD proteini için metal bağlanma yerlerinden grafiksel bir kesit.	50
Şekil 3. 9. 1RMD proteininin ZN metali ile 4 farklı bağ yeri.	51

SİMGELER VE KISALTMALAR LİSTESİ

AUC	Area Under Curve
BR	Binary Representation
CNN	Convolutional Neural Networks
DNN	Deep Neural Network
DL	Deep Learning
DNA	Deoksiribonükleik Asit
FN	False Negative
FP	False Positive
GRU	Gated Recurrent Units
kNN	k-Nearest Neighbors
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficients
PAM	Point Accepted Mutation
ProCos	Protein Composition Server
PSSM	Position Specific Scoring Matrix
PSWM	Position Weighted Scoring Matrix
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

TERİMLER LİSTESİ

Derin Öğrenme	Deep Learning
Evrişimsel Sinir Ağları	Convolutional Neural Networks
Kapılı Tekrarlayan Hücre	Gated Recurrent Unit
Uzun-Kısa Süreli Bellek	Long Short-Term Memory
Nokta Kabul Edilen Mutasyon	Point Accepted Mutation
İkili Temsil	Binary Representation
Protein Bileşimi Sunucusu	Protein Composition Server
Düzeltilmiş Doğrusal Birim	Rectified Linear Unit
Adım Aralığı	Stride
Dolgu Değeri	Padding
Havuzlama	Pooling
Flatten	Düzleştirme

1. GİRİŞ

Proteinlerin fonsiyonel ve yapısal analizleri üzerine yapılan çalışmalara bağılı olarak büyük miktarlarda protein verisinin oluşması sağlanmıştır. Bu verilerin biyolojik deneylerle analiz edilmesi zordur. Bunun için proteinlerin yapıları ve fonksiyonlarının tespit edilmesinde hesaplamalı yöntemler ortaya çıkmıştır. Geleneksel makine öğrenme yöntemleri bu alanda etkin olarak kullanılmıştır. Son zamanlarda derin öğrenme ağıları kullanılarak bu alanda çok etkili sonuçlar alınmıştır.

1.1. Kapsam

Bu tez, protein metal bağlanma yerlerinin tahmin edilmesi için geliştirilen yöntemleri sunmaktadır. Bunun için geleneksel makine öğrenme yöntemlerinin yanında son yıllarda çok iyi performanslar sunan derin öğrenme teknikleri de kullanılmıştır. Elde edilen derin öğrenme modelleri protein analizine bağılı çeşitli problemlerde de kullanılabilirliği tartışılmıştır. Protein dizilimlerinin temsillerine ilişkin üç farklı öznitelik çıkarım metodu ele alınmıştır. Derin öğrenme teknikleri ve protein dizilim temsilleri metal bağlanma yeri tahmin problemine uygulanmış ve deneysel sonuçlar sunulmuştur. Birinci bölümün birinci kısmı tezin kapsamı, ikinci kısmı tezin katkısı, üçüncü kısmı ise tezde atıfta bulunulan biyolojik kavramlara ilişkin tanımlamalarla devam etmektedir. Dördüncü kısımda metal bağlanma problemi, son kısımda ise bu alanda daha önce yapılan çalışmalar yer almaktadır.

İkinci bölümde çalışmada kullanılan metodların tanımlanması yer almaktadır. Veri kümesi 2.1'de tanımlanmaktadır. Geleneksel makine öğrenme yöntemleri 2.2'de ve derin öğrenme teknikleri 2.3'te açıklanmaktadır. Protein dizilim temsil yöntemleri 2.4' te yer almaktadır. Üçüncü bölümde ise uygulanan modeller ve elde edilen deneysel bulgular yer almaktadır. Çalışmanın yapıldığı uygulama ortamının teknik alt yapı bilgileri burada açıklanmaktadır. Tasarlanan derin öğrenme model yapıları 3.1'de tanımlanmaktadır. Protein dizilim temsillerine göre metal bağlanma yeri tahmin sonuçları 3.2'de, geometrik yapı tahmini bulguları da 3.3'te verilmektedir. Bütün uygulamalar için kıyaslama veri kümeleri üzerinde yapılan deneyler sonucunda analizler sunulmaktadır. Son bölümde ise tartışma ve sonuç yer almaktadır.

1.2. Katkılar

Protein metal bağlanma yeri tahmini için hem sınıflandırıcı hem de öznitelik çıkarımına ilişkin literatürde birçok farklı metod geliştirilmiştir. Bu metodlar kullanılarak protein dizilimlerinde tek bir metalden başlayıp bütün metalleri kapsayan bağlanma yerleri tespit edilmiştir. Bu çalışmada ise Histidin ve Sistein amino asitlerinin bütün metallerle bağlanma yerleri incelenmiştir.

Tez çalışmasında protein metal bağlanma yerlerinin tahmini için yeni yöntemler önerilmektedir. Buna göre tezin katkıları şunlardır:

- Protein metal bağlanma yeri tahmini için derin öğrenme ağ mimarileri ilk defa uygulanmıştır.
- Protein dizilimlerini sayısallaştırmak için PAM skorlama matrisi, protein kompozisyonları ve ikili temsilden oluşan 5 farklı yöntem kullanılmıştır.
- Derin öğrenme ağ mimarileriyle protein dizilimi ve yapı analizinde kullanılabilecek etkin bir çözüm ortaya çıkmıştır.
- Metal bağlanma yeri tahmini için geleneksel makine öğrenme yöntemleri ve hesaplamalı yöntemler uygulanarak en son 2006 yılında en yüksek başarımla elde edilmiş olan referans veri kümesine evrimsel sinir ağları, uzun-kısa süreli bellek ve kapılı tekrarlayan hücre algoritmalarından oluşan üç farklı derin öğrenme ağ mimarisi uygulanarak önceki çalışmalardan daha iyi bir başarımla elde edilmiştir. En yüksek başarımla ise evrimsel sinir ağları ve PAM skorlama matrisi öznitelikleriyle elde edilmiştir.
- Geleneksel makine öğrenmeleri PAM skorlama matrisi özniteliklerine uygulanarak önceki çalışmalardan daha iyi sonuçlar elde edilmiştir.
- Genel olarak daha önceki çalışmalarda PSSM öznitelik kümeleri kullanılarak yapılan çalışmalar yanında bu çalışmada PAM skorlama matrisi ile elde edilen öznitelikler kullanılmasıyla başarımların yüksek olmasına etki etmiştir.
- PAM öznitelikleri ile kolektif sınıflandırma algoritmaları olan Bagging ve AdaBoost algoritmaları da daha önceki çalışmalardan daha iyi sonuçlar elde edilmiştir.
- Metal bağlanma yerlerinin geometrik tespiti için derin öğrenme ağları ilk defa kullanılmıştır.

- Bir protein dizilimindeki kalıntılardan hangisi kaç tane metale nereden bağlandığı geometrik olarak tahmin edilmiştir.

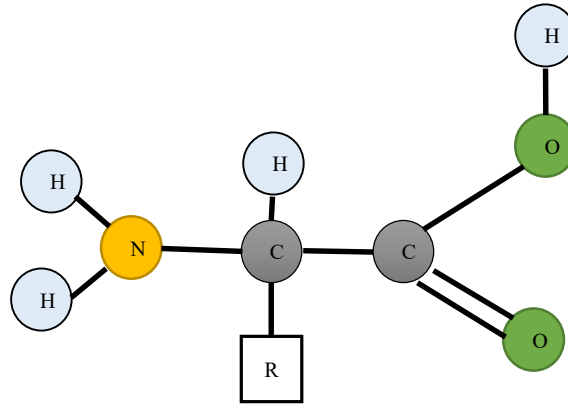
Bu çalışma sayesinde, elde edilecek yeni veri kümeleri aracılığı ile proteinlerin yapılarına ve fonksiyonlarına ilişkin yeni bulgulara ulaşılması olasıdır.

Çalışmada geliştirilen derin öğrenme ağ mimarileri protein analiz çalışmalarıyla ilgili farklı uygulamalarda kullanılabilecektir.

Bu tez çalışmasında ele alınan problemin çözümü için önerilen yöntemler, benzer problemler üzerinde çalışacak araştırmacılara ışık tutacaktır.

1.3. Biyolojik Kavram ve Tanımlar

Amino asitler proteinlerin temel yapı taşıdır ve metabolizma da birçok kritik rol oynarlar. Tüm amino asitler, Şekil 1.1’de gösterildiği gibi molekülün ortasında alfa-karbon olan bir karbon atomu içerir [1]. Bu atom üç kimyasal grupla çevrilidir. İlki bir amin grubu NH_2 'dir. İkincisi bir karboksil grubu $COOH$ 'dir. Üçüncü grup R ile gösterilir. Bu değişken radikal gruptur ve her amino asit için farklıdır. Bu R grubu amino asidi tekillik sağlar.



Şekil 1. 1. Amino asit yapısı.

Proteinler 20 farklı amino asitten meydana gelirler. Belirli bir proteinin amino asit içeriği ve bu amino asitlerin dizisi, bu proteini kodlayan gendeki bazların dizisi ile belirlenir. Amino asitler protein dizilimlerinde farklı sayılarda bulunurlar ve farklı yapıda ve fonksiyonda binlerce çeşit protein oluşumuna neden olurlar. Her bir amino asidin protein yapı ve fonksiyonlarında kendine özgü kimyasal özellikleri ve farklı rolü vardır. Suyla temas etme eğilimine bağlı olarak amino asitler hidrofobik (su ile temas etme eğilimi düşük), polar ve yüklü (su ile enerjisel olarak

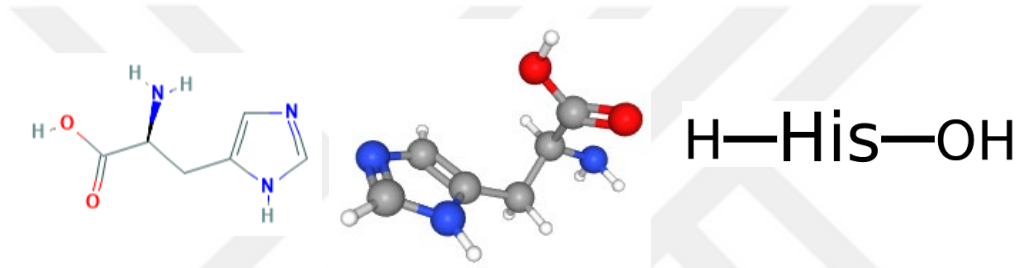
uygun temas) olarak sınıflandırılabilir [2]. Proteinler amino asit dizileri içinde, proteinin üç boyutlu bir yapıya nasıl katlanacağını belirlemek için gerekli bilgileri ve ortaya çıkan yapının kararlılığını içerir.

Amino asitler, yapılarına ve yan zincirlerinin yapısına, yani R zincirlerine göre sınıflandırılabilir. Polar olmayan amino asitler: bunlar hidrofobik olarak da bilinir. R grubu, alkil gruplarından veya aromatik gruplardan olabilir. Polar amino asitler: amino asidin yan zincirleri aminler, alkoller veya asitler gibi farklı polar gruplar içeriyorsa, bunlar doğada polardır. Bunlar hidrofilik asitler olarak da bilinir. Bunlar ayrıca üç kategoriye daha ayrılır. a) Asidik: Yan zincir fazladan bir karboksilik asit bileşeni içeriyorsa, bunlar asit-polar amino asitlerdir. Hidrojen atomlarını bağışlama eğilimindedirler. b) Bazik: Bunlar, bir hidrojen atomu çekme eğiliminde olan ekstra bir azot grubuna sahiptir. c) Nötr: Bunlar asidik veya bazik değildir. Eşit sayıda amino ve karboksil grubuna sahiptirler. Ayrıca, elektronegatif atomlara bağlı en az bir hidrojen bileşenine sahiptirler. Amino asit sınıfları ve listesi Şekil 1.2’de gösterilmiştir.

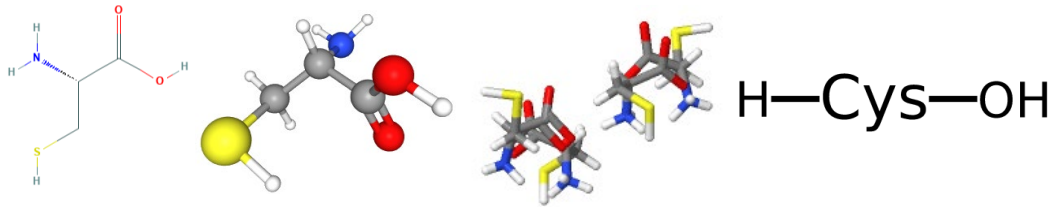
Aminoasit	3-Harf	1-Harf	Sınıf
Alanin	ALA	A	Polar değil
Arjinin	ARG	R	Polar-bazik
Asparajin	ASN	N	Polar-nötr
Aspartik asit	ASP	D	Polar-asidik
Sistein	CYS	C	Polar-nötr
Glütamik asit	GLU	E	Polar-asidik
Glütamin	GLN	Q	Polar-nötr
Glisin	GLY	G	Polar değil
Histidin	HIS	H	Polar-bazik
İzolösin	ILE	I	Polar değil
Lösin	LEU	L	Polar değil
Lizin	LYS	K	Polar-bazik
Metiyonin	MET	M	Polar değil
Fenilalanin	PHE	F	Polar değil
Prolin	PRO	P	Polar değil
Serin	SER	S	Polar-nötr
Treonin	THR	T	Polar-nötr
Triptofan	TRP	W	Polar değil
Tirozin	TYR	Y	Polar-nötr
Valin	VAL	V	Polar değil

Şekil 1. 2. Amino asit listesi.

Her bir amino asit metabolizma içerisinde farklı görevlere sahiptir. Çalışmamızda incelenen Histidin ve Sistein amino asitleri de bu kapsamda hayati süreçleri yerine getirmektedirler. Histidin, metabolizmada büyüme ve doku onarımı için kullanılması yanında, bağışıklık, alyuvar ve akyuvar yapımı gibi önemli fonksiyonlardan sorumludur [3]. Histidin, sinir hücrelerini koruyan miyelin kılıflarının bakımı için de önemlidir. Şekil 1.3’de Histidin’in iki boyutlu, üç boyutlu ve kristal yapılarına ilişkin görüntüleri verilmiştir. Sistein ise protein sentezi, zararlı toksik maddelerin temizlenmesi ve çeşitli metabolik fonksiyonlar için önemlidir. Tırnak, cilt ve saçtaki ana protein olan beta-keratin’de bulunan Sistein, kolajen üretiminde olduğu kadar cilt elastikiyeti ve dokusunda da önemlidir [4]. Sistein’in iki boyutlu, üç boyutlu ve kristal yapıları Şekil 1.4’te gösterilmiştir.



Şekil 1. 3. Histidin amino asidinin 2-boyutlu, 3-boyutlu ve biyolojik yapıları

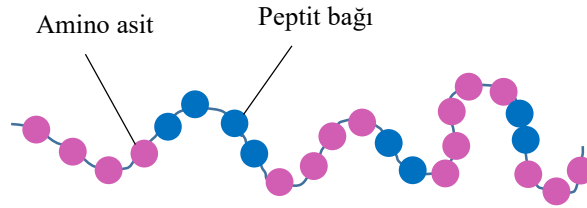


Şekil 1. 4. Sistein amino asidinin 2-boyutlu, 3-boyutlu ve Kristal yapıları

Proteinler yaşamın ana yapı taşlarından biridir ve hayati fonksiyonlara sahiptirler. Proteinler bütün hücrelerde ve hücrelerin bütün kısımlarında bulunurlar. Proteinlerin yapı taşları da amino asitlerdir. Proteinler, amino asitler olarak bilinen bu küçük birimlerden oluşan karmaşık biyomoleküllerdir. Amino asitler birbirlerine peptit bağı ile bağlanırlar. Bu bağ, bir amino asidin karboksil grubu (-COOH), bir su molekülünü (H₂O) salan başka bir amino asidin amino grubu (-NH₂) ile bağlandığında oluşur. Bir peptit, dipeptit, tripeptit ve polipeptit olabilir.

Proteinler yapısındaki amino asitlerle birlikte yapısal olarak dört türe ayrılır: birincil, ikincil, üçüncül ve dördüncül yapı. Her yapı belirli kurallar çerçevesinde meydana gelir. Şekil

1.5’de gösterildiği gibi birincil yapılar, amino asitlerin peptit dizisinde dizilimiyle oluşan seridir. Birincil yapıdaki amino asit değişiklikleri proteinlerin üç boyutlu yapılarına ve fonksiyonlarına etki edebilir. İkincil yapı Şekil 1.6’da ki gibi ya α -heliks bir yapıya ya da hidrojen bağları içeren β -tabakalı bir yapıya katlanır. Hidrojen bağlarının polipeptit omurgası boyunca düzenli aralıklarla kurulmasıyla bu katlanmalar oluşur.



Şekil 1. 5. Protein birincil yapı.



β -Sheet (3 strands)



α -helix

Şekil 1. 6. Protein ikincil yapı.

Proteinlerin özel ilmeklere ve kıvrımlara yerleştirilmesi ve birbirine bağlanması, üçüncül yapıları oluşturur (Şekil 1.7). Bu yapı polipeptit zincirinin üç boyutlu düzenlenmesi için hidrojen, iyonik ve disülfür bağları içerir. Üçüncül yapıdaki motifler, ikincil yapıların düzenli kombinasyonlarıdır. Yani ikincil yapı oluşuktan sonra polipeptit zincirinin daha ileri katlanmalar halinde düzenlenmesiyle oluşur.



Şekil 1. 7. 1AXC proteininin üçüncül yapısı [5].

Dördüncül yapı ise iki veya daha fazla polipeptidin çok alt birimli bir kompleks haline getirilmesi ile oluşur (Şekil 1.8). Yapısal, fonksiyonel veya topolojik alanlar ikincil ve üçüncül yapı modülleridir.



Şekil 1. 8. 1AXC proteininin dördüncül yapı [5].

Proteinler vücudun yapı taşları olarak adlandırılır. Hücre ve dokulardaki çeşitli yapıların önemli bir bileşenidir. Bu proteinleri ayrıca insan vücudundaki tüm hücrelerin dış zarında buluyoruz. Saç, cilt ve kaslarda da yapısal proteinler bulabiliriz.

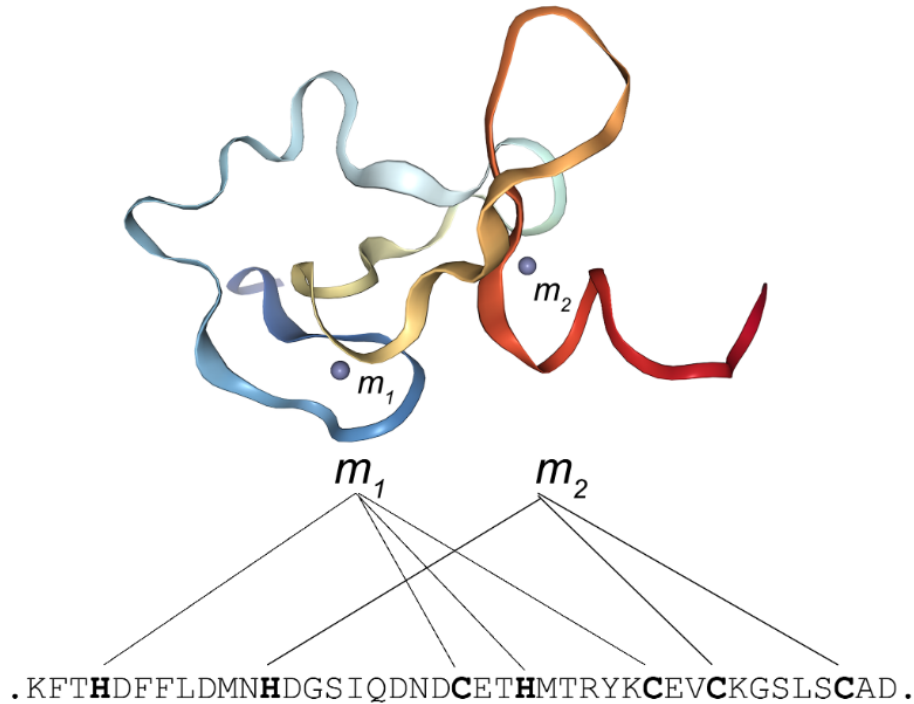
Metaloprotein, yapısında metal iyonu içeren proteinlerdir. Proteinlerin büyük bir kısmı bu gruba dâhildir. Proteinlerin yarısından fazlasının metal içerdiği tahmin edilmektedir. Metal iyonları, çok çeşitli hücresel süreçlerde önemli roller oynamaktadır. Alkali ve toprak alkali metaller yapısal bir rol oynarlarken, geçiş metalleri genellikle hem yapı stabilizasyonunda hem de katalizde yer alır. Metaller ayrıca anti-kanser ilaçları da dahil olmak üzere metal bazlı ilaçlar şeklinde tıpta önemlidir. Metaloproteinler hücreler içerisinde birçok görevi üstlenmektedir. Proteinlerin fonksiyonel ve biyolojik işlevlerinin keşfedilmesi için proteinlerin metallerle bağlarının tespit edilmesi ve metal-protein etkileşiminin bilinmesi önemlidir. Protein içerisinde bulunan metal iyonları proteinlerin fonksiyonel görevlerini yerine getirebilmesi, yapısı ve kararlılığı için önem arz etmektedir.

Protein dizilimleri üzerinden metal ile bağlanma noktaların tespiti, proteinlerin üç boyutlu yapılarının belirlenmesi için oldukça önemli bir çalışma alanıdır. Proteinler üç boyutlu yapılarına, çevresinde bulunan metal iyonları ile yapmış olduğu kuvvetli bağlar sonucunda

erişmektedir. Bu sebeple, proteinler üzerinde metal ile bağlanma noktalarının yüksek performans ile tespiti çok önemlidir.

1.4. Metal Bağlanma Problemi

Proteinlerin fonksiyonlarının tespit edilmesinde hesaplamalı yöntemler önem arz etmektedir. Proteinler katlanırlar ve üç boyutlu bir yapıya dönüşürler. Katlanma olabilmesi için proteinin birden fazla yerden bir veya daha fazla metal iyonu ile etkileşime girmesi ve bu metaller ile bir bağ yapması gerekmektedir. Protein Veri Tabanı (PDB)'da ki metaloproteinlerin %66'sını geçiş metalleri oluşturur, en fazla miktarda demir ve çinko metali içermektedir. Geçiş metalleri diğer metallere kıyasla kurdukları kovalent bağlar ile protein kalıntılarına daha yüksek bağlanma etkileşimi gösterirler. Diğer taraftan geçiş metalleri ile bağ yapan en yaygın protein kalıntıları ise CYS, HIS, ASP ve GLU'dur. Sadece CYS ve HIS kalıntıları ile bağlı metaller geçiş metallerinin büyük çoğunluğunu kapsamaktadırlar. Metal bağlanma yerinin tespit edilmesi o proteinin hangi görevi olduğu, biyolojik olarak hangi öneme sahip olduğunu gösterebilir [6], [7]. Bunun için proteinlerin metal bağlanma yeri tahmini biyolojik olarak önemli bir problemdir ve halen çözülmesi için çalışmalar devam etmektedir. Teknolojik gelişmelere bağlı olarak bu problemin çözümü için yeni teknikler de sunulmaktadır. Şekil 1.9'da 5x6t proteininin çinko metali ile bağlantısı gösterilmektedir.



Şekil 1. 9. Çinko metali ile m1 ve m2 noktalarından bağlı üç boyutlu 5X6T proteini (üstte), protein diziliminin bağlantı yerlerinin grafiksel gösteriminden bir kesit (altta).

1.5. Önceki Çalışmalar

Proteinlerin metal bağlanma yerlerinin deneysel olarak tanımlanması zor olabilir. Buna karşı hesaplamalı yöntemlerin geliştirilmesi ile metal bağlanma yerlerinin kolay ve hızlı bir şekilde analiz edilmesi sağlanabilir. Protein metal bağlanma yerlerinin tespit edilmesi için çeşitli hesaplamalı yöntemler geliştirilmiştir. Matematiksel hesaplamaların yanında yapay sinir ağları, destek vektör makineleri, karar ağaçları, aç gözlü algoritma, grafik teorisi ve geometri algoritması gibi yöntemler bunlara örnek verilebilir. Bu yöntemler kullanılarak bir çok çalışma yapılmıştır. Proteinlerin metal bağlanma yerlerinin tespiti için ya protein dizilimlerinden elde edilen öznitelikler ya da protein yapılarından elde edilen öznitelikler kullanılmaktadır. Çalışmalar bütün metalleri kapsayacağı gibi, geçiş metalleri veya tekil bir metalin bağlanma durumu içinde yapıldığı görülmektedir.

Sodhi vd. [8] yapay sinir ağları ve konuma özgü skor matrisleri kullanarak geliştirdikleri Metsite yaklaşımı ile Ca, Cu, Mg, Fe, Mn ve Zn metal iyonlarının protein dizilimleri üzerlerindeki bağlanma yerlerini %94,5 doğruluk değeri ile tahmin etmişlerdir. Deng vd. [9] kalsiyum bağlanma yerlerini tespit etmek için graf ve geometri tabanlı bir metod geliştirmiş ve 123 kalsiyum bağlı protein için % 90 duyarlılığa ulaşmışlardır. Lin vd. [10] Ca, P, Mg ve Na metallerinin proteinlerle bağlanma tahmini için kayan pencere tabanlı öznitelik çıkarımı ve biyolojik öznitelik kodlama teknikleri ile Yapay sinir ağları kullanarak %90 duyarlılık elde etmişlerdir. Lin vd. [11] 10 metal iyonunun (Ca, Co, Cu, Fe, Mg, Mn, Ni, P, Na, Zn) bağlanma kalıntılarını tahmin etmek için 53.333 metal bağlı kalıntı içeren veri kümesi üzerinde eğitilmiş SVM metodu kullanmışlardır. Metod, 31.448 bağımsız metal bağlı kalıntı kümesi üzerinde değerlendirilmiş ve %74,9 doğruluk değeri elde edilmiştir.

Sadece protein sekans bilgisi kullanarak Sistein ve Histidin kalıntıları için geçiş metali bağlanmasını tahmin eden bir yöntemi Passerini vd. [12] geliştirmişlerdir. Yöntem, Sisteinleri üç durumdan birine sınıflandırmıştır: serbest (F), disülfid bağlı (D) ve metal bağlı (M), Histidinleri ise iki durumdan birine sınıflandırmıştır: metal bağlı (M) ve serbest (F). Yöntem sadece pozisyona özgü evrimsel profillerin yanı sıra protein uzunluğu ve amino asit bileşimi gibi daha küresel tanımlayıcıları kullanarak sekans bilgilerini kullanır. Yöntem iki aşamalı makine öğrenme yaklaşımına dayanmaktadır. İlk aşama, Histidinlerin ve Sisteinlerin bağlanma durumunu lokal olarak sınıflandırmak üzere eğitilmiş bir destek vektör makinesinden (SVM) oluşur. İkinci aşama, aynı protein içindeki kalıntılar arasındaki bağımlılıkları dikkate alarak

yerel tahminleri düzeltmek için eğitilmiş iki yönlü tekrarlayan bir sinir ağından (BRNN) oluşur. Geçiş metal bağlayıcı bölgelerindeki Histidinleri ve Sisteinleri %73 kesinlik ve %61 duyarlılık değeri ile tahmin etmişlerdir. Goyal ve Mande [13] protein yapılarında metal bağlanma yerlerini tespit etmek için üç boyutlu yapısal motifler kullanmışlardır. Bordner [14], rastgele orman sınıflandırıcı kullanarak Ca, Cu, Fe, Mg, Mn ve Zn metallerinin bağlanma yerlerini tespit etmek için SitePredict metodunu geliştirmişlerdir. Bu yöntemde, kalıntı türlerinin uzlamsal kümelenmesi gibi çeşitli kalıntı bazlı yer özellikleri kullanılmıştır. Babor vd. [15] CHED algoritması ile üç boyutlu yapıya dayanarak 349 apoprotein ve 82 holoprotein içindeki geçiş metal (Zn, Co, Ni, Fe, Cu ve Mn) bağlanma yerlerini tahmin etmişlerdir. Lippi vd. [16] geliştirdikleri MetalDetector isimli web sunucusu ile proteinlerdeki HIS kalıntılarını serbest veya metal bağlı olarak iki durumdan birine, CYS'leri de serbest, metal bağlı veya disülfür bağlı olmak üzere üç durumdan birine sınıflandırmaktadır.

Passerini vd. [17] geçiş metallerinin protein dizilimlerinde CYS ve HIS bağlanma yerlerini tanımlayan MetalDetector metodunu geliştirip MetalDetector2 ile hangi kalıntıların aynı metal iyonunun koordinasyonunda birlikte yer aldığını tahmin etmeye çalışmışlardır. Burada iki alt görev tanımlanmaktadır: ilk görev olarak CYS ve HIS kalıntılarının bağlanma yerlerinin tespiti, ikinci görev olarak metal bağlı CYS ve HIS'nin birlikte gruplandırılması ve bunlara geleneksel bir metal iyon tanımlayıcının atanmasıdır. Brylinski ve Skolnick [18] modellenmiş protein yapılarında destek vektör makineleri ve evrimsel bilgiler kullanarak metal bağlanma yerlerini tespit etmek için FINDSITE-metal metodunu tasarlamışlardır. Zhao vd. [19] yapı tabanlı TEMSP çalışması ile proteinlerin Zn metali ile bağlanma yerlerini tahmin etmişlerdir. TEMSP ile hedef protein içindeki potansiyel ligand kalıntılarının C α ve C β atomlarının nispi pozisyonları ile Zn bağlanma yerlerindeki olası pozisyonlarla karşılaştırılmaktadır. Lu vd. [20] protein dizilimi ve yapısal bilgilerini kullanarak fragman transformasyon yöntemi ile proteinlerdeki Ca, Mg, Cu, Fe, Mn ve Zn metal iyonlarının bağlanma yerlerini %94,6 doğruluk değeri ile tahmin etmişlerdir. Zheng vd. [21] proteinlerde Zn bağlanma yerlerinin tahmini için çoklu dizilim, yapısal özellikler ve grafik-teorik ağ özelliklerinden oluşan üç çeşit protein özniteliği entegre ederek Zn tanımlayıcı adı verilen iki aşamalı rastgele orman algoritması önermişlerdir.

Önceki çalışmalarda genellikle sekans bilgilerinden, seçilen kalıntıların bağlanma durumlarını tahmin eden yöntemler geliştirilmiş olmaları yanında motif tabanlı yaklaşımlar ile

de bağlanma yerleri tespit edilebilmektedir. Passerini vd. [22] Sistein ve Histidinler tarafından koordine edilen geçiş metali bağlanma yerlerini tespit etmek için yapılandırılmış çıktı öğrenmeye dayanan yeni algoritmik fikirler geliştirmişlerdir. Hiçbir kalıntının birden fazla metal iyonunu koordine edemeyeceği varsayımına dayanarak, metal bağlamanın bir matroidin cebirsel yapısına sahip olduğunu göstermişlerdir. Metal bağlı biyolojik makromoleküllerin üç boyutlu yapıları hakkındaki mevcut bilgileri aktarmak için MetalPDB veri tabanı sunulmuştur [23]. Burada, proteinler ve nükleik asitlerdeki metal bağlayıcı bölgeler minimal fonksiyonel bölgeler yoluyla sistematik ve otomatik temsili yoluyla elde edilmektedir. Chen vd. [24] SVM, küme ve şablon tabanlı sınıflandırma metodları kullanarak CYS, HIS, ASP ve GLU amino asitlerinin Zn metali ile bağlanma yerlerini tespit etmek için ZincExplorer hibrit metodunu geliştirmişlerdir.

Liu vd. [25] Zn bağlanma yerlerinin yapısal özelliklerinin sekans uzunluğu dağılımı, amino asit tercihleri, ikincil yapı ve geometrik mesafe gibi farklı ardışık ve yapısal özelliklere sahip olduğunu göstermişlerdir. 3 ve 4 kalıntı bağlanma için geometrik mesafe aralıkları kullanarak 'Geometrik Kısıtlama' olarak adlandırılan Zn bağlayıcı bir kalıntı tahmin edici geliştirmişlerdir. Zheng vd. [26] kullanıcı tarafından yüklenen yapıların veya PDB'nin metal bağlanma yerlerini doğrulamak için bir CheckMyMetal web sunucusunu geliştirmişlerdir. Zhou vd. [27] üç boyutlu protein yapılarından Ca bağlanma yerlerinin tespiti için makromoleküler yapılardaki fonksiyonel alanların modellenmesi ve tanınması için geliştirilen FEATURE [36] çalışmasını kullanmışlardır. He vd. [28] mFASD adı verilen ve üç boyutlu yapı verilerine dayanarak metal ile bağ yaptığı düşünülen bir dizi fonksiyonel atomun kimyasal ortamını araştırarak metal bağlanma yerlerini tespit etmişlerdir.

Hu vd. [29] Zn, Cu, Fe, Ca, Mg, Mn, Na ve K metal iyonlarının protein bağlanma yerlerini tahmin etmek için ab initio modeli ile çoklu iplik hizalamalarını birleştiren kompozit bir yöntem (IonCom) geliştirmişlerdir. Lin vd. [30] MIB çalışmasıyla yapısal motifler kullanarak metal bağlanma yeri tahmini yanında metal iyon yerleştirme için de kullanılan bir web sunucusu geliştirmişlerdir. Cao vd. [31] metal iyonu bağlanma yerlerini konum ağırlıklı skorlama matrisi (PWSM) ve SVM algoritması kullanarak % 79,9'luk bir doğruluk ve 0,6'luk bir MCC ile tanımlanmasını sağlamışlardır. Li vd. [32] konuma özgü skor matrisi (PSSM) öznelikleri kullanarak doğrusal regresyon yaklaşımı ile proteinlerde Zn bağlanma yerlerini tahmin etmişlerdir. Song vd. [33] rastgele orman algoritmasını kullanarak Ca, Co, Cu, Fe, Ni, Mg, Mn,

ve Zn metallerinin proteinlerle bağlanma yerlerini tespit etmek için MetalExplorer çalışmasını geliştirmişlerdir. Srivastava ve Kumar [34] protein dizilimlerinde konuma özgü skor matrisi özniteliklerini kullanarak SVM tabanlı Zincbinder metodunu önermişler, %85,37 duyarlılık ve %86,20 özgüllük elde etmişlerdir. Proteinlerde Zn bağlanma yerlerinin tahmini için Li vd. [35] destek vektör makinesi ve örnek-ağırlıklı olasılık sinir ağından oluşan SSWPNN sınıflandırma metodunu geliştirmişlerdir. Sekans tabanlı protein metal bağlanma yeri çalışmaları Tablo 1.1’de, yapı tabanlı protein metal bağlanma yeri çalışmaları ise Tablo 1.2’de gösterilmiştir.

Geleneksel makine öğrenmesi algoritmaları yanında son zamanlarda derin öğrenme algoritmaları da biyoinformatik çalışmalarında kullanılmaktadır. Derin öğrenme algoritmaları biyobilişim alanında yüksek başarımlar göstermiştir. Alipanahi vd. [37] geliştirdikleri DeepBind metodu ile DNA ve RNA bağlı proteinlerin dizilim özelliklerinin tahmininde derin öğrenme teknikleri kullanmışlardır. Taeho ve. [38] protein dizilimlerinden ve protein çiftlerinden çıkarılan yapısal özellikleri kullanarak protein katlanmalarının doğru şekilde tanınması için derin öğrenme ağ yöntemini (DN-Fold) geliştirilmişlerdir. DeepSEA çalışması ile Zhou ve Troyanskaya [39] dizilimlerden kodlanmamış değişken etkilerinin tahmini için derin öğrenme yöntemi kullanmışlardır.

Kelley vd. [40] Basset isimli çalışmaları ile genomik verilerden DNA dizilimlerinin fonksiyonel aktivitelerini öğrenmek için derin evrimsel sinir ağları kullanılmıştır. Hamanaka vd. [41] ile Tian vd. [42] derin öğrenme teknikleri ile bileşik protein etkileşimlerinin tahmininde derin öğrenme algoritmaları kullanmışlardır. Wang vd. [43] protein ikincil yapı tahmini için koşullu rastgele alanların (CRF) ve sıkı sinir ağlarının bir entegrasyonu olan koşullu sinir alanlarının (CNF) derin öğrenme uzantısı olan DeepCNF (Derin evrimsel nöral alanlar) metodunu geliştirmişlerdir. Wang vd. [44] ultra derin bir sinir ağı yoluyla hem evrimsel birleştirme hem de sekans koruma bilgilerini birleştirerek protein temaslarını tahmin eden bir derin öğrenme yöntemi sunmuşlardır. Armenteros vd. [45] protein dizisel lokalizasyonunu sekans bilgisine dayanarak tahmin etmek için derin sinir ağları kullanan bir tahmin algoritması sunmuşlardır. Pan ve Shen [46] global ve lokal evrimsel sinir ağlarını birleştirerek RNA dizilimlerinden RNA-protein bağlanma bölgelerini tahmin etmek için iDeepE yöntemini sunmuşlardır. Kulmanov vd. [47] derin öğrenme ağları kullanarak protein sekans ve etkileşimlerinden protein fonksiyonlarının tahmini için DeepGo metodunu geliştirmişlerdir.

Tablo 1. 1. Sekans tabanlı metal bağlanma yeri çalışmaları.

Yazar	Öznitelik	Bağlı metal	Amino asit	Metod
Sodhi vd. [8]	PSSM	Ca, Cu, Fe, Mg, Mn, Zn	hepsi	Yapay sinir ağları
Lin vd. [11]	amino asit kompozisyon	Hepsi	hepsi	SVM
Passerini vd. [12]	PSSM	geçiş metalleri	Histidin, Sistein	SVM, BRNN
Lippi vd. [16]	PSSM	geçiş metalleri	Histidin, Sistein	Web sunucu
Passerini vd. [17]	PSSM	geçiş metalleri	Histidin, Sistein	Web sunucu, SVM-RNN, SVM-HMM
Lu vd. [20]	hizalama skoru	Ca, Mg, Cu, Fe, Mn ve Zn	Histidin, Sistein, Aspartik asit, Glütamik asit	fragman transformasyon yöntemi
Zheng vd. [21]	PSSM, DSSP, Grafik teorik ağ özellikleri	Zn	Histidin, Sistein, Aspartik asit, Glütamik asit	Rastgele orman algoritması
Passerini vd. [22]	PSSM	geçiş metalleri	Histidin, Sistein	Greedy algoritması
Chen vd. [23]	PSSM	Zn	hepsi	SVM, küme ve şablon tabanlı sınıflandırma metodları
Hu vd. [28]	PSSM, lokal yapı özellikleri, pozisyona ve segmente özgü korunma skorları ve ligand-spesifik bağlanma eğilimleri	Zn, Cu, Fe, Ca, Mg, Mn, Na ve K	hepsi	AdaBoost, SVM
Cao vd. [31]	PSSM	Zn, Cu, Fe, Ca, Mg, Mn, Na, K, Co	hepsi	SVM
Li vd. [32]	PSSM	Zn	hepsi	doğrusal regresyon
Song vd. [33]	PSSM, DSSP, Grafik teorik ağ özellikleri	Ca, Co, Cu, Fe, Ni, Mg, Mn, Zn	hepsi	Random Forest
Srivastava ve Kumar [34]	PSSM	Zn	Histidin, Sistein, Aspartik asit, Glütamik asit	SVM
Li vd. [35]	PSSM	Zn	hepsi	SVM, Sinir ağı

Tablo 1. 2. Yapı tabanlı metal bağlanma yeri çalışmaları.

Yazar	Öznitelik	Bağlı metal	Amino asit	Metod
Deng vd. [9]	oksijen atomları arası mesafe	Ca	hepsi	graf ve geometri tabanlı metod
Lin vd. [10]	biyolojik öznitelik kodlama	Ca, P, Mg, Na	hepsi	Çok katmanlı algılayıcı
Goyal ve Mande [13]	yapısal motifler, uzak dizi homologları, geometrik parametreler	hepsi	hepsi	hesaplamalı yöntem
Bordner [14]	kalıntı türlerinin uzamsal kümelenmesi ve evrimsel koruma özellikleri	Ca, Cu, Fe, Mg, Mn, Zn	hepsi	Rastgele orman algoritması
Babor vd. [15]	amino asit bileşimi, korunma skoru, hidrojen bağı temas yüzey alanı, triad sayısı	hepsi	hepsi	Karar ağaçları, SVM
Brylinski ve Skolnick [18]	yapı / evrim bilgisi, TM-Score	geçiş metalleri	hepsi	SVM
Zhao vd. [19]	geometrik özellikler	hepsi	hepsi	matematiksel metod
Liu vd. [25]	yapısal özellikler	Zn	hepsi	matematiksel metod
Zheng vd. [26]	-	hepsi	hepsi	web sunucu
Zhou vd. [27]	fizikokimyasal özellikler	hepsi	hepsi	Naive Bayes
He vd. [28]	atomlar arası uzaklık	hepsi	hepsi	matematiksel metod
Lin vd. [30]	yapısal motifler	Ca, Mg, Cu, Fe, Mn, Zn, Cd, Ni, Hg, Co	hepsi	web sunucu, matematiksel metod

Jones ve Kandathil [48] evrimsel sinir ağları ve minimal sekans öznitelikleri kullanarak protein temas tahmininde yüksek performans elde etmişlerdir. Metal bağlı proteinler için bir çok veritabanı oluşturulmuştur. Bu veri veritabanları Tablo 1.3’de gösterilmektedir.

Tablo 1. 3. Metal bağı protein veritabanları

Veritabanı	Erişim bağlantısı
MetalPDB [23]	http://metalweb.cerm.unifi.it/
COMe [49]	http://www.flymine.org/come/
MESPEUS_10 [50]	http://mespeus.bch.ed.ac.uk/MESPEUS_10/
MetalMine [21]	http://metalmine.naist.jp/metalmine009/index.html
MIPS [52]	http://dicsoft2.physics.iisc.ernet.in/mips/
MetLigDB [53]	http://silver.sejong.ac.kr/MetLigDB

Proteinlerin metallerle bağlanma yerlerini tespit etmek için geliştirilen yazılımlar Tablo 1.4'te gösterilmektedir.

Tablo 1. 4. Protein metal bağlanma yeri tahmnini için geliştirilen yazılımlar

Yazılım	Kullanılan öznitelik türü
MetSite [8]	sekans ve yapı tabanlı
SVMProt [12]	sekans tabanlı
MetalDetector v2.0 [17]	sekans tabanlı
FINDSITE-metal [18]	sekans ve yapı tabanlı
TEMSP [19]	yapı tabanlı
Mfasd [28]	yapı tabanlı
IonCom [29]	sekans ve yapı tabanlı
MetalPredator [54]	sekans tabanlı
SeqCHED Server [55]	sekans tabanlı
Metals2 [56]	yapı tabanlı
3DLigandSite [57]	sekans ve yapı tabanlı
FindGeo [58]	yapı tabanlı

2. YÖNTEMLER

2.1. Veri Kümesi

Metal bağlanma yeri tahmini için en iyi başarımlar, Passerini vd. 2006 yılında yaptıkları SVM-BRNN çalışması ile elde edilmiştir [12]. SVM-BRNN’de kullanılan veri kümesi referans veri kümesi olarak bizim çalışmamızda da kullanılmıştır. Veri kümesi 2727 proteinden oluşmaktadır. Bu veri kümesi toplam 564400 kalıntı, 9183 Sistein, 13678 Histidin içermektedir. Tablo 2.1’de gösterildiği gibi veri kümesinde yer alan proteinlerden 365 protein en az bir metal ile bağ yapmakta (MBS), 687 protein disülfid bağı yapmakta (DB) ve 1727 protein ise hiçbir bağ yapmayan (FREE) zincirlerden oluşmaktadır. 47 zincir hem metale bağı hem de disülfid bağı yapmaktadır. Veri kümesine ait proteinlerin sekans bilgileri PDB’den (Protein Data Bank) indirilerek MySql veri tabanına kaydedilmiş ve sonraki işlemler bu veritabanı üzerinden yürütülmüştür.

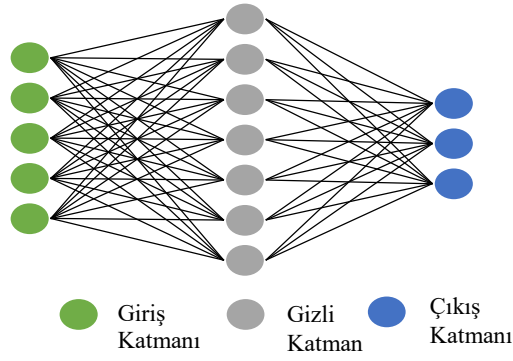
Tablo 2. 1. Veri Kümesi.

	FREE	DB	MBS
Sistein	3786	3896	1783
Histidin	8924	2578	2602
Zincir	1727	687	365

Eğitim veri kümesi %19 pozitif, %81 negatif veriden, test veri kümesi ise %21 pozitif, %79 negatif veriden oluşmaktadır.

2.2. Geleneksel Makine Öğrenmesi

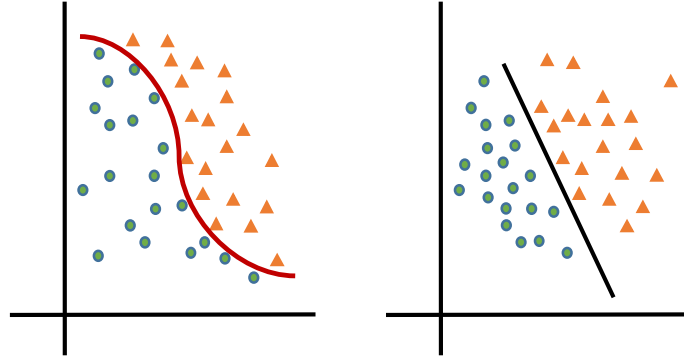
Makine öğrenimi, sistemlere deneyimlerden otomatik olarak öğrenme ve geliştirme yeteneği sağlayan yapay zekâ alanıdır. Bu alan, verilere erişebilen ve öğrenmeyi kendileri için kullanabilen bilgisayar programlarının geliştirilmesine odaklanır. Amaç, bilgisayarların insan müdahalesi veya yardımı olmadan otomatik olarak öğrenmelerini sağlamak ve eylemleri buna göre ayarlamaktır. Geleneksel makine öğrenme tekniklerinden biri olan Yapay Sinir Ağı mimarisi Şekil 2.1’de gösterildiği gibi giriş katmanı, gizli katman ve çıktıdan oluşmaktadır. Bu çalışmada dört makine öğrenmesi kullanılmıştır: Destek vektör makineleri, Naive Bayes, AdaBoost ve Bagging algoritması.



Şekil 2. 1. Yapay Sinir Ağı.

2.2.1. Destek Vektör Makineleri (SVM)

Farklı veri grubuna ait bir veri kümesini, gruplar arasına sınır çizgisi çekerek sınıflara ayırmak mümkündür. Burada gruplar arasındaki sınırın nasıl çizileceği sınıflandırmanın ne derece verimli olacağını gösterir [59]. SVM, sınıflandırma için oldukça kullanılan ve en etkili yöntemlerden birisidir. Diğer sınıflandırma algoritmalarında olduğu gibi SVM doğrusal olarak ayrılabilen verilerin sınıflandırmasını iyi bir şekilde yapabilmektedir. Şekil 2.2’de gösterildiği gibi SVM doğrusal olarak ayrılamayan veri kümeleri için de başarılı bir sınıflandırma işlemi yapmaktadır.



Şekil 2. 2. Örnek SVM sınıflandırmaları.

2.2.2. Naïve Bayes

Temelleri Bayes teoremine dayanır. Veri kümesi içerisinde ki bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olan değere göre sınıflandırma yapar. Aşağıdaki şekilde hesaplanmaktadır:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (1)$$

Burada;

$P(A|B)$, B olayı gerçekleştiğinde A olayının gerçekleşme olasılığını,

$P(A)$, A olayının gerçekleşme olasılığını,

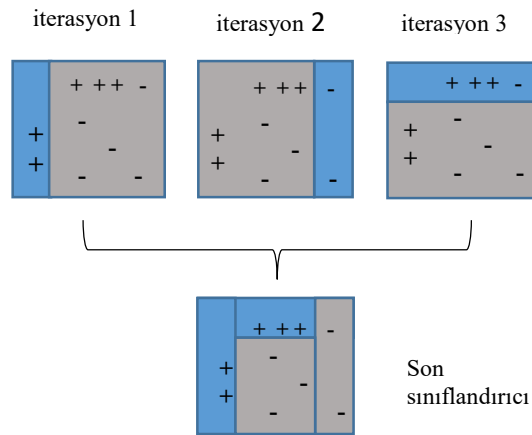
$P(B|A)$, A olayı gerçekleştiğinde B olayının gerçekleşme olasılığını,

$P(B)$, B olayının gerçekleşme olasılığını vermektedir.

2.2.3 AdaBoost

Sınıflandırma çalışmalarında genelde bilinen makine öğrenmesi veya derin öğrenme algoritmaları kullanılır. Hangi algoritmanın kullanılacağına ya deneme yanılma yöntemiyle ya da eğitim verisinin yapısına göre karar verilir. Kolektif öğrenme de ise tek bir temel öğrenici yerine birden fazla algoritma birlikte kullanılarak yeni bir model oluşturulur. Daha az hafıza kullanımı, tahmin hızı yüksekliği gibi özelliklerden dolayı Adaboost algoritması en çok kullanılan kolektif öğrenme algoritmalarından biridir. Adaptive Boosting' in kısaltması olarak 1996 yılında Freund ve Schapire tarafından önerilmiştir. Bu algoritmada öğrencilerin eğitileceği örneklerin seçiminde önceki temel öğrencilerin hata yaptıkları örnekler öncelik verilmektedir [60]. Örneklerin eğitim kümesine seçilme olasılıkları aynı değildir, her iterasyonda güncellenmektedir. Bu da yanlış verilen kararlar üzerine odaklanılmasını sağlamaktadır.

Adaboost algoritmasında en güçlü zayıf sınıflandırıcılar bir araya getirilir ve güçlü bir sınıflandırıcı oluşturulur. Şekil 2.3'de üç iterasyonlu AdaBoost sınıflandırma örneği gösterilmektedir.

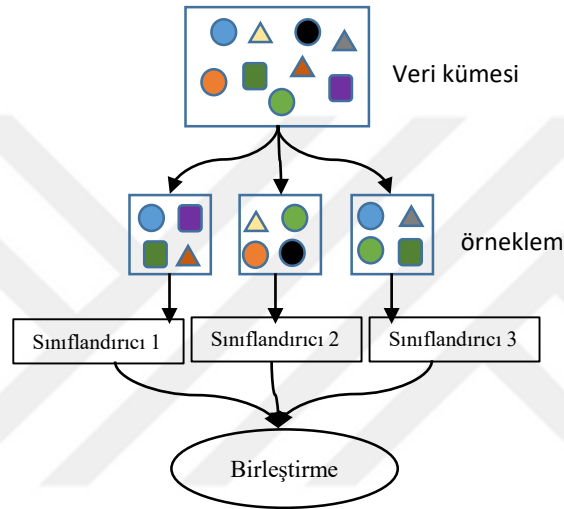


Şekil 2. 3. AdaBoost sınıflandırma örneği.

2.2.4. Bagging

Kolektif öğrenme yöntemlerinden biridir. Mevcut bir eğitim kümesinden yeni eğitim kümeleri türeterek temel öğreniciyi yeniden eğitmeyi sağlayan bir algoritmadır [61].

Eğitim kümesinden rastgele seçilen veri kümesi tekrar eğitim kümesine konularak yeni eğitim kümesi elde edilir. Yani rastgele seçilen her örnek tekrar geri eğitim kümesine konulur. Bu şekilde seçilen eğitim kümeleri ile başarılı temel öğreniciler eğitilir ve kolektif başarı elde edilir (Şekil 2.4). Burada temel öğrenici olarak karar ağaçları veya bir başka makine öğrenme yöntemi seçilebilir.



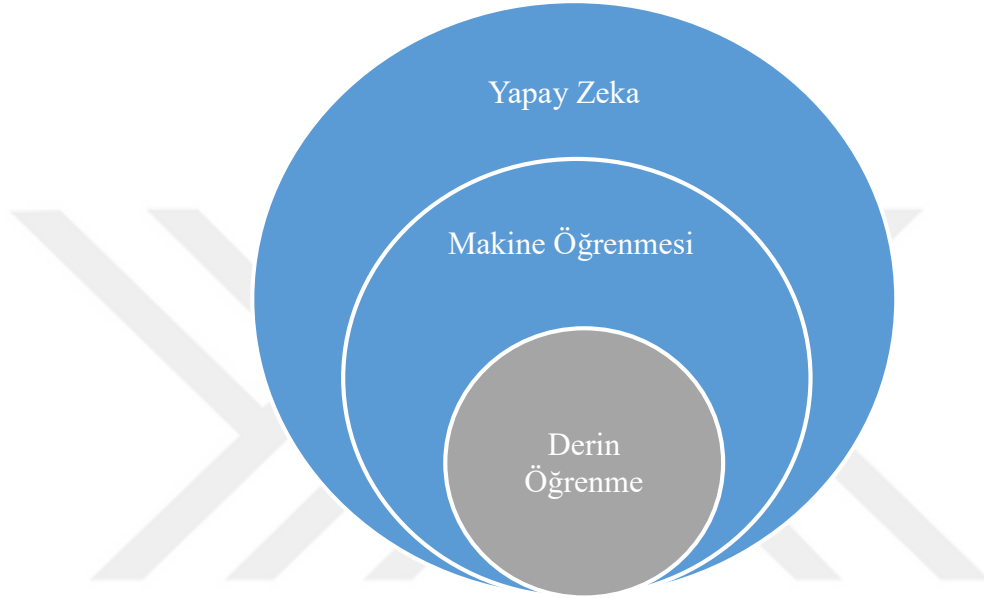
Şekil 2. 4. Bagging algoritması.

2.3. Derin Öğrenme Ağları

Derin öğrenme, hesaplamalı bir yöntem olarak çok sayıda katmandan oluşan makine öğrenmesinin bir alt sınıfıdır [62]. Makine öğrenmesi yapay zekanın bir alt alanı olarak ifade edilirse, derin öğrenme de makine öğrenmenin alt alanı olarak adlandırılabilir. Yapay zeka, makine öğrenmesi ve derin öğrenme arasındaki ilişki Şekil 2.5’de gösterilmiştir.

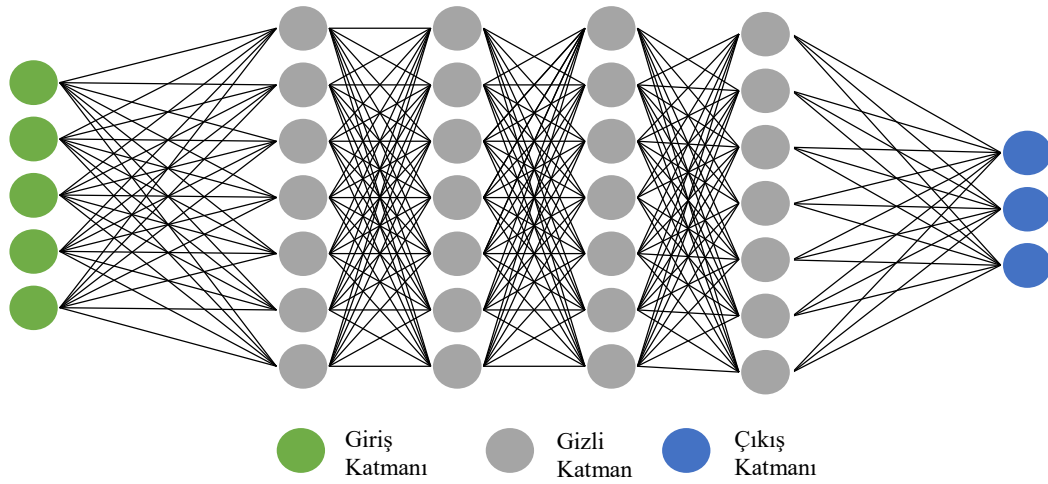
Derin öğrenme ağlarının ön işleme, boyut azaltma ve sınıflandırma aşamalarını tek bir modelde birleştirerek modeli tekillğe indirgermesi onlara çok önemli bir avantaj sağlamaktadır. Bu özelliğiyle derin öğrenme yaklaşımları üstün seviyede ayrıştırma performansı sağlamaktadır. Bu yaklaşımda her katmanın çıktısı, bir sonraki katmanın girdisi olarak kullanılmaktadır. Bilgi bu katmanlar boyunca birinden diğerine doğru akar ve bilgi adım adım öğrenilir ya da çıkarılır. Her seviyedeki nöronlar kendi tahminlerini ve en olası tahminlerini yaparlar ve daha sonra bu bilgiyi nihai sonuca kadar bir sonraki seviyeye aktarırlar. Karmaşık mimarilere sahip olmaları derin öğrenme ağlarına otomatik öz nitelik çıkarma

yeteneđi sađlar. Geleneksel makine öğrenme yaklaşımlarında ise bu işlem algoritmik aşama dışında gerçekleştirilir. Burada ham veriler kişiler veya veri bilimciler tarafından analiz edilerek anlamlı özniteliklere dönüştürölür. Derin öğrenmede ilk katmanlar daha ince kavramları öğrenmeye odaklanırken daha derin katmanlar daha soyut kavramları öğrenmek için daha önce öğrenmiş olduđu bilgileri kullanırlar. Bu prosedürlerden oluşun süreç öznitelik çıkarma işlemidir.



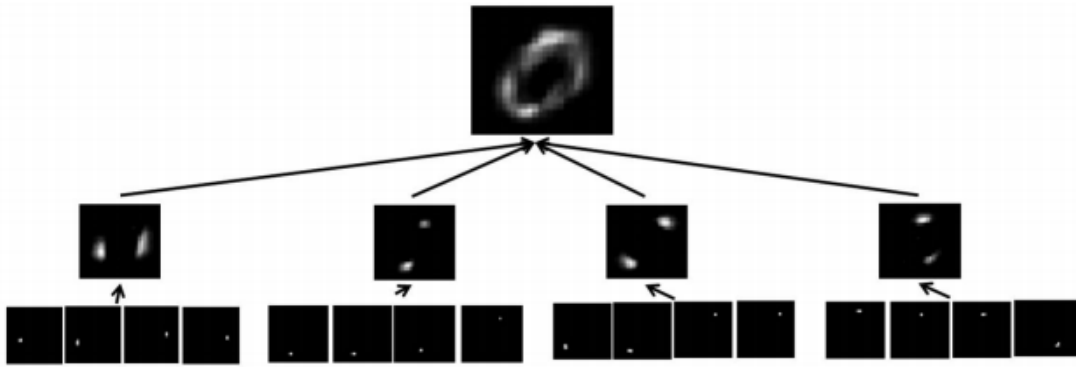
Şekil 2. 5. Yapay zeka, makine öğrenmesi ve derin öğrenme.

Derin öğrenme mimarilerinin temel avantajı yapılandırılmamış veriler üzerinde eğitilebilmesidir. Bu güçlü durum onlara daha iyi öğrenme fırsatı sunmaktadır. Şekil 2.6'de dört katmanlı bir derin öğrenme mimarisi gösterilmiştir. Derin öğrenme, veri temsiliinin öğrenilmesine dayanan bir yapıya sahiptir. Yani, alt seviyedeki özelliklerden üst seviyedeki özellikler türetilmekte, bu da hiyerarşik bir temsili meydana getirmektedir. Bir görüntü için örnek verirse, piksel başına yoğunluk değeriinin bir vektörü veya kenar kümeleri gibi özellikler ile görüntü temsil edilebilir. Bu özelliklerin hepsi veriyi temsil edemeyebilir, veriyi daha iyi temsil eden özellik tercih edilmelidir. Bu noktada derin öğrenmenin hiyerarşik bir şekilde otomatik öznitelik çıkarma katmanları verinin temsili için bir avantaj olarak karşımıza çıkmaktadır.



Şekil 2. 6. Derin öğrenme ağı örneği.

Şekil 2.7’de MNIST veritabanı için örnek bir özellik hiyerarşisi gösterilmiştir [63]. Burada ilk katman, temel yapı taşları olarak görülebilen çok basit noktaya benzeyen özellikleri öğrenir. İkinci katmanda ise bu birinci katman özelliklerinin kombinasyonları dağıtılmış örüntüde öğrenilir. Derin öğrenme, son zamanlarda derin öğrenme ağları (DNN), evrişimsel sinir ağları (CNN), kapılı tekrarlayan hücre (GRU), uzun-kısa süreli bellek (LSTM) gibi yöntemleri kullanılarak, bilgisayarla görü, konuşma tanıma, doğal dil işleme, ilaç keşfi ve gen uygulamalarından oluşan çeşitli problemlere uygulanmış ve yüksek başarımlı performansı göstermiştir [64].



Şekil 2. 7. MNIST veritabanı özellik hiyerarşisi. Aşağıdan yukarıya görüntü temsili [63].

2.3.1. Evrişimsel Sinir Ağları (CNN)

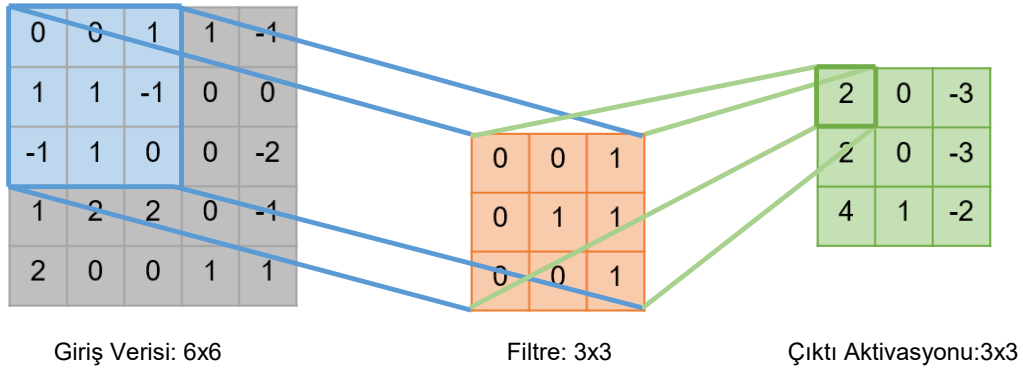
Evrişimsel sinir ağı (CNN), bir girdiyi alıp girdi içerisindeki çeşitli özellikleri birbirinden ayırabilen bir derin öğrenme algoritmasıdır. CNN, görme korteksinden esinlenerek ortaya çıkmıştır. Görme korteksi gözden gelen çok ayrıntılı duyuşal bilgileri alarak işler ve gerekli

işlevleri yürütür. Görsel korteks, görsel alanın belirli bölgelerine duyarlı küçük hücreli bölgelere sahiptir. Karmaşık hücreler ise daha geniş bölgelere veya tüm görsele yoğunlaşırlar. Evrişimsel sinir ağı, bir veya daha fazla evrişim katmanı, havuzlama katmanı ve bir veya daha fazla tam bağlantılı katmandan oluşan çok katmanlı bir yapay sinir ağıdır [62]. Evrişim katmaları giriş verisinden öznitelik çıkarmak için kullanılır. Her bir evrişim katmanı, öznitelik çıkarımında yardımcı olan birçok filtreye sahiptir. Burada her evrişim katmanı bir önceki katmandan çıkarılan öznitelikleri giriş verisi olarak alıp daha yüksek-seviyeli bir soyutlanma ile öznitelik çıkarımını sağlar. Evrişimsel sinir ağlarının öncül katmanlarında kenar bilgisi gibi daha temel öznitelikler elde edilirken, daha derin katmanlarında ise daha karmaşık öznitelikler elde edilmektedir.

Evrişimsel sinir ağlarında giriş ve çıkış verilerinin boyutları yapılan hesaplamalara bağlı olarak farklılıklar gösterebilmektedir. Şekil 2.8’de gösterilen örnekte olduğu gibi giriş verisi $(w \times w) = 5 \times 5$, evrişim filtresi $(f \times f) = 3 \times 3$ boyutlarında olması durumunda, yapılan evrişim işlemi ile çıkış verisinin boyutu $(w-f+1) \times (w-f+1) = 3 \times 3$ olur. Eğer giriş verisi boyutu ile çıkış verisi boyutunun aynı olması bekleniyorsa padding adı verilen piksel ekleme işlemi yapılır. CNN algoritmaları görüntü işleme, ses işleme, doğal dil işleme (NLP) ve biyobilişim gibi birçok farklı alanda başarıyla uygulanmaktadır. Özellikle görüntü işleme alanında en iyi başarımlar elde edilmiştir. CNN algoritmaları ilaç keşfi çalışmalarında da geliştirilen ilk derin öğrenme ağı olmuştur [65].

CNN mimarisinde girdi, 1 boyutlu, 2 boyutlu ve 3 boyutlu görüntü şeklinde olabilir. 1 boyutlu girdiler sekans veri tipleri için kullanılmaktadır. Bu çalışmada protein dizilimleri için ise CNN mimarisinin 2 boyutlu girdi yapısı kullanılmıştır. Her protein dizilimi öncelikle 2 boyutlu bir görüntüye dönüştürülmüştür.

Evrişim katmanı, evrişimsel sinir ağının ilk katmanıdır. Giriş verisinin sol üst köşesinden başlayarak verilen filtre boyutuna göre veriyi alıp bu verinin temsilini oluşturur. Burada temsil için ağırlıklı ortalama, en yüksek değer, orijinal veri ile çarpım gibi işlemler kullanılabilir. Bu işlem bütün girdi verisi için tekrarlanarak bir temsil verisi ortaya çıkarılır. Şekil 2.8’de evrişim katmanı için bir örnek gösterilmiştir. Burada 5×5 boyutlu giriş verisindeki 3×3 boyutlu mavi blok içerisindeki veri, 3×3 boyutlu filtre ile çarpılarak 3×3 boyutlu çıkış aktivasyonu elde edilmiştir.



Şekil 2. 8. Evrişim katmanında çıktı hesaplama örneği.

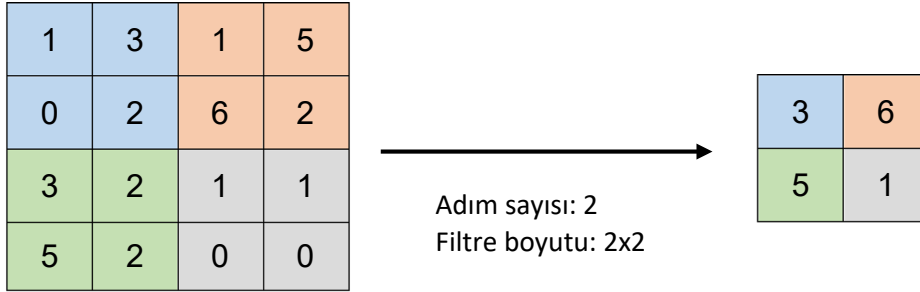
Adım aralığı (stride) sağa sola kaydırma yapılacak şekilde bir adım seçilmiştir. Çıktı aktivasyonunun ilk verisi olan 2 sonucu için yapılan örnek işlem: $0*0+0*0+1*1+1*0+1*1+(-1*1)+(-1*0)+1*0+0*1 = 2$

Çıkış hacminin uzaysal boyutu, giriş hacminin boyutu (W), filtre boyutu (F), uygulanan adım sayısı (S) ve kenarlarda kullanılan sıfır dolgu (P) ile hesaplanabilmektedir. Buna göre çıkış boyutu için en uygun formül $(W-F+2P)/S+1$ 'dir.

Örneğin Şekil 2.9'daki 5x5'lik giriş ve 3x3 filtre yanında adım sayısı 1 ve dolgu değeri 0 için 3x3 çıkış elde ederiz. Adım sayısı 2 seçilirse 2x2 çıktı elde edilir. Adım sayısı arttıkça çıkış boyutu da azalmaktadır. Bunu önlemek için dolgu (padding) kullanılır. Dolgu değeri de genellikle 0 alınır.

Evrişimsel sinir ağlarında boyut azaltma işlemi olarak kullanılan diğer yöntem havuzlama (pooling) işlemidir. Havuzlama işleminde boyutun azaltılması ile özellik kaybı olmasına rağmen sinir ağının parametre ve hesaplama ile ilgili eğitim süresi azalır, aşırı uyum (ezberleme) ihtimali de azalmış olur. Evrişimsel sinir ağlarında art arda her evrişim katmanı arasına genellikle havuzlama katmanı eklenmektedir. Havuzlama işlemi için en yaygın olarak maksimum havuzlama (max pooling) ve ortalama havuzlama (average pooling) yöntemleri kullanılmaktadır. Ortalama havuzlama yöntemi geçmişte sıklıkla kullanılmaktaydı fakat son zamanlarda maksimum havuzlama pratikte daha iyi çalıştığı için daha çok kullanılmaya başlandı. Maksimum havuzlama yönteminde, belirlenen adım sayısına göre havuzlama filtresi girdi üzerinde bir pencere gibi gezdirilir ve her pencerenin en yüksek değeri çıktıda karşılık

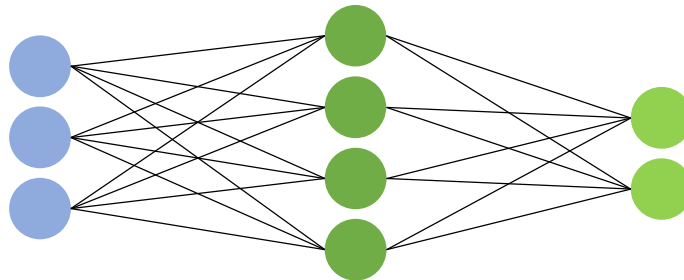
gelen alana yerleştirilmektedir. Şekil 2.9’da girdi boyutu kesiti 4x4, filtre boyutu 2x2 ve adım sayısı 2 olan maksimum havuzlama uygulaması gösterilmiştir.



Şekil 2. 9. Maksimum havuzlama örneği.

Tam Bağlantılı Katman, evrişim ve havuzlama katmanları ile öznelilik çıkarım işlemi tamamlanan evrişimsel sinir ağının son katmanı olan tam bağlantılı katmanda sınıflandırma işlemi yapılmaktadır. Şekil 2.10’da gösterildiği gibi tam bağlantılı katmandaki nöronlar, normal bir yapay sinir ağındaki gibi bir önceki katmandaki tüm nöronlarla tam bağlantılıdır. Tam bağlantılı katman ile evrişim katmanı arasındaki fark, evrişim katmanındaki nöronların sadece girişteki bir bölgeye bağlı olmasıdır.

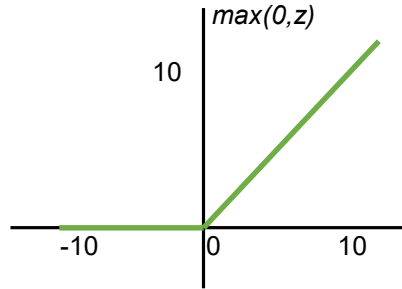
Evrişim ve havuzlama katmanları üç boyutlu çıktılar üretmesine karşın, tam bağlantılı katman girdi olarak tek boyutlu bir veri beklemektedir. Bunun için üç boyutlu çıktı verisi düzleştirme (Flatten) katmanı ile tek boyutlu bir yapıya indirgenerek tam bağlantılı katman için uygun bir hale getirilir.



Şekil 2. 10. Tam bağlantılı katman.

Doğrultulmuş Doğrusal Birim (ReLU) katmanı, derin öğrenme ve evrişimsel sinir ağlarında en çok kullanılan aktivasyon fonksiyonlarından biridir ve $\max(0,z)$ denklemi ile tanımlanmaktadır. Şekil 2.11’de gösterildiği gibi, z değerinin sıfırdan büyük olduğu durumlarda

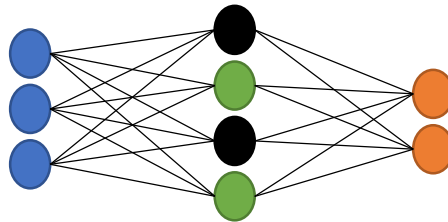
z 'yi, sıfır ve sıfırdan küçük olduğu durumlarda sıfır değerini üretir. Yani gelen girdileri sıfır ile sonsuz arasında çıktılara dönüştürür.



Şekil 2. 11. ReLU aktivasyon fonksiyonu.

ReLU ile ağdaki bütün nöronlar yerine bazı nöronların aktif olması sağlanır, böylece daha seyrek bir aktivasyon ve verimli bir hesaplama yükü ortaya çıkmış olmaktadır. Negatif değerlerde fonksiyonun sıfır değer alması da ağın hızlı çalışmasının bir göstergesidir.

Bırakma (Dropout) Katmanı, ağın büyük olmasına bağlı olarak uzun süre eğitim yapılıyorsa veya veri sayısı az ise aşırı uyum riski ortaya çıkabilir. Eğitim sırasında aşırı uyumu engellemek için bazı nöronların unutulması için bırakma katmanı kullanılır. Şekil 2.12'te gösterildiği gibi bırakma katmanı her katmandaki rastgele seçilen belirli nöron kümesinin eğitim aşamasında bazı nöronlarının unutulması ile daha sağlam öznetelikleri öğrenmeye zorlar. Her yinelemede bırakılacak nöronlar rastgele seçileceğinden öğrenme algoritmasının da hangi nöronların unutulacağı hakkında hiçbir fikri olmayacaktır.



Şekil 2. 12. Bırakma katmanı: siyah renkli nöronlar unutulmuştur.

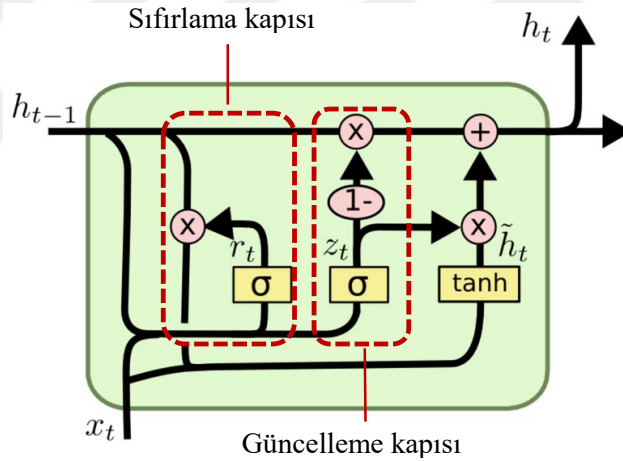
2.3.2. Kapılı Tekrarlayan Hücre (GRU)

Kapılı Tekrarlayan Hücre, bir tekrarlayan sinir ağı türüdür. Tekrarlayan sinir ağları (RNN) diğer sinir ağlarından farklı olarak sıralı bilgi kullanırlar. İleri beslemeli sinir ağlarının aksine tekrarlayan sinir ağları giriş dizilerini işlemek için dahili belleklerini kullanırlar. RNN'ler kendi kendilerine uzun süreli bağımlılıkları öğrenemezler ve genellikle kısa süreli hafızaları kullanırlar.

Uzun dizilerle, özellikle zaman serileri modellemesinde ve metin analizinde, RNN'ler geri yayılma sırasında yok olan gradyan problemine maruz kalırlar. Bu problemi çözmek için birçok varyasyon geliştirilmiştir. Bunlardan en önemlisi uzun-kısa süreli bellek (LSTM) ve diğeri kapılı tekrarlayan hücre (GRU) ağıdır. GRU'lar Şekil 2.13'te gösterildiği gibi dahili bellek özelliğini kullanarak, güncelleme ve sıfırlama kapılarını da kullanarak bilgileri saklamayı ve filtrelemeyi sağlar. Bu kapılar ileriki süreçlerde sonucu etkileyecek bilginin taşınmasını sağlar ve çıktıya hangi bilgilerin aktarılması gerektiğine karar verirler. Yani değer, gelecekteki bir zamanda güncellemek için mevcut durumuyla birlikte saklanır. Güncelleme kapısı (z_t) aşağıdaki formüle göre hesaplanır:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (2)$$

Güncelleme kapısı, bir modelin önceki zaman adımlarından yola çıkarak geçmiş bilgilerin ne kadarının geleceğe iletilmesi gerektiğini belirlemesine yardımcı olur.



Şekil 2. 13. GRU

Bu özellik ağı güçlü olmasını sağlar, çünkü model geçmişten gelen tüm bilgileri kopyalayıp iletmeye karar verebilir ve gradyan problemini ortadan kaldırma riskini azaltabilir.

Sıfırlama kapısı (r_t) ise geçmiş bilgilerin ne kadarının unutulacağına karar vermek için kullanılır. Güncelleme kapısının tersi bir işleve sahiptir. Sıfırlama kapısı aşağıdaki formül ile hesaplanmaktadır:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (3)$$

Mevcut hafıza içeriğini hesaplamak için;

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \quad (4)$$

Ve geçerli zamandaki son hafızayı hesaplamak için;

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (5)$$

formülleri kullanılmaktadır.

Burada x_t input vektörünü, h_t gizli katman vektörünü, W_r ve W_h ise ağırlığı ifade etmektedir. GRU güncelleme ve sıfırlama kapısı özellikleriyle dikkatli bir şekilde eğitilirse karmaşık senaryolarda bile iyi performans gösterebilir.

2.3.3. Uzun-Kısa Süreli Bellek (LSTM)

Uzun-kısa süreli bellek ağları, uzun süreli bağımlılıkları öğrenebilen bir RNN türüdür. LSTM yapısında geçmiş ve gelecekteki bilgiyi anlamlandıracak öznitelikler özyinelemeli olarak taşınmaktadır. LSTM üniteleri uzun süre bellekte bilgi tutabilen bir 'bellek hücresi' içerir [62]. Bu bellek ile bir önceki zamandan gelen bilgi alınabilir ve bir sonrakine iletilebilir. Hangi bilginin alınıp hangi bilginin alınmayacağına eğitim ile karar verilmektedir. Şekil 2.14'te gösterildiği gibi RNN'den farklı olarak bir tane sinir ağı yerine dört tane vardır.

Bu katmanlar aracılığıyla girdi, unutma, çıktı kapıları hücre durumundan ağın hafızasını oluşturmaktadır.

Unutma kapısı;

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (6)$$

girdi kapısı;

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (7)$$

çıktı kapısı;

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

hücre durum vektörü;

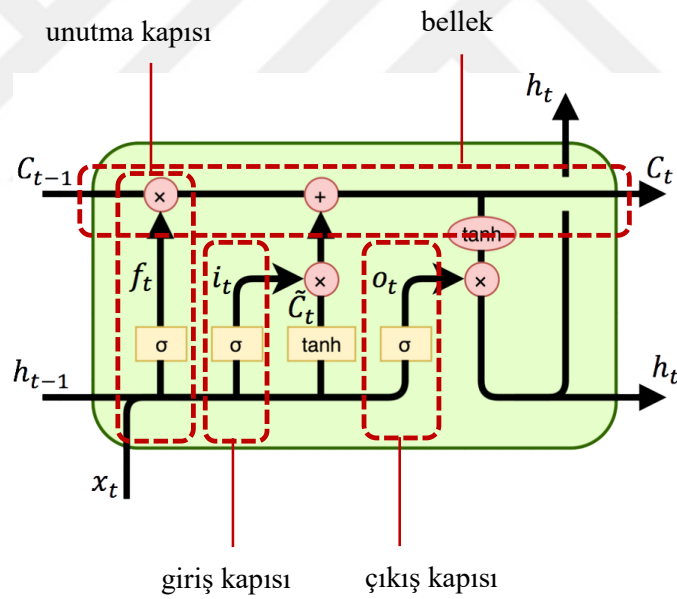
$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (9)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (10)$$

gizli durum vektörü;

$$h_t = o_t \odot \tanh(C_t) \quad (11)$$

formülleri ile hesaplanmaktadır. Burada x_t input vektörünü, h_t gizli katman vektörünü, W_i , W_f , W_o ve W_c ise ağırlığı ifade etmektedir.



Şekil 2. 14. LSTM.

2.4. Öznitelik Çıkarma

Proteinler farklı uzunlukta amino asit dizilimlerinden oluşmaktadır. Proteinlerin farklı uzunluklu amino asitlerden meydana gelmesi sınıflandırma da kullanılacak makine öğrenme metodları için tutarlı bir veri değildir. Bunun için proteinlerin sabit uzunluklu bir öznitelik vektörü ile temsil etme araçlarına ihtiyaç vardır. Bunun sağlanması ile dizilimler üzerinde

matematiksel işlem yapmak da kolay olacaktır. Bunun için ilk olarak bu dizilimlerin hesaplanabilecek şekilde temsil edilmeleri gerekmektedir.

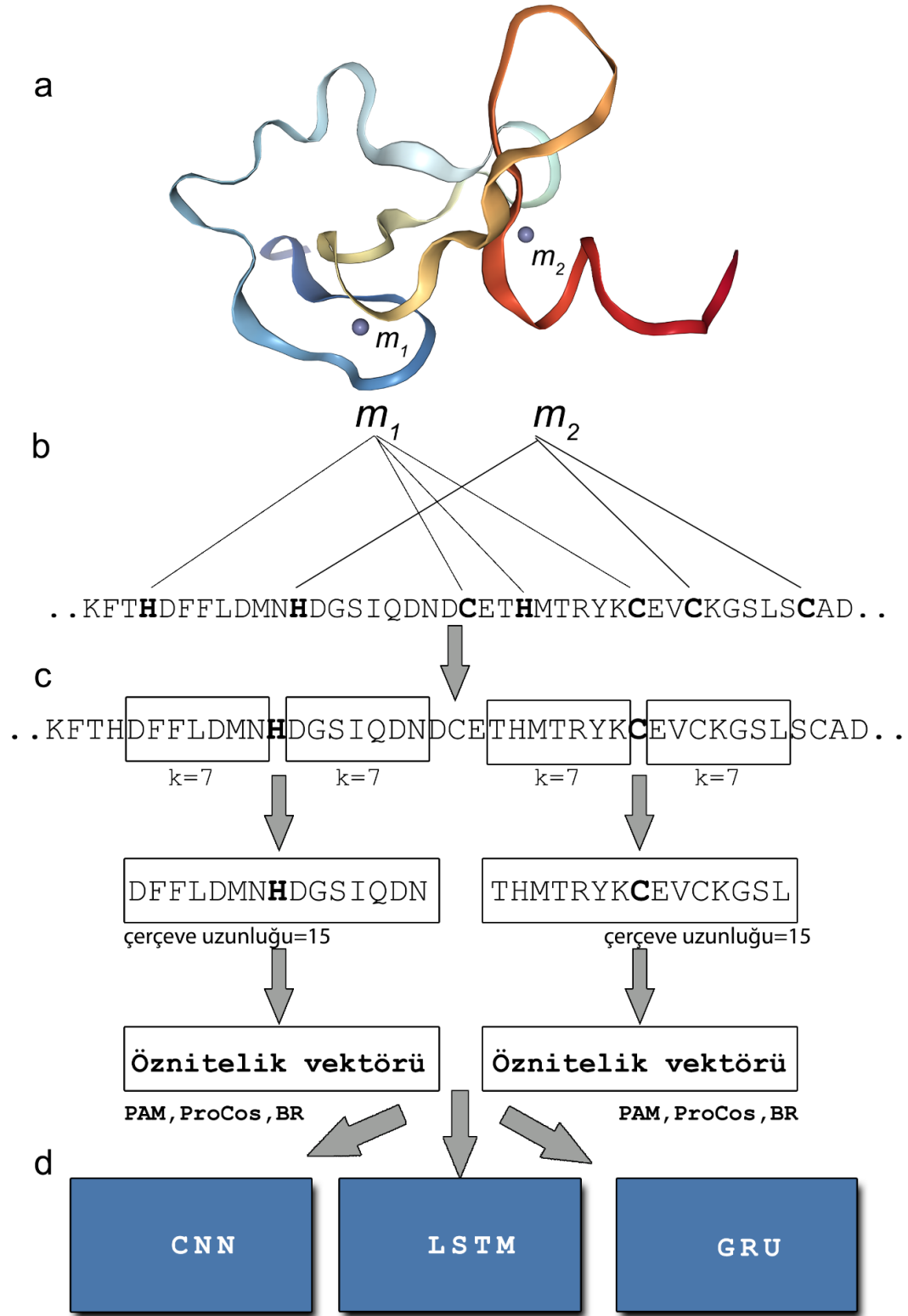
Bir protein yapısını etkili bir şekilde tanımlamak ve sınıflandırmak için protein dizilimi temsil yöntemleri kullanılmaktadır. Literatürde birçok temsil yöntemi önerilmiştir. Bunlar sayısal, grafiksel, geometrik ve hibrit temsil yöntemleri olarak sınıflandırılabilirler. Bu çalışmada 3 farklı sayısal temsil yöntemi kullanılmıştır. Nokta Kabul Edilen Mutasyon (PAM) skarlama matrisi, Protein Kompozisyon Sunucusu (ProCos) ve İkili temsil (BR).

n tane amino asit uzunluğunda bir protein zinciri $S = a_1 a_2 a_3 a_4 \dots a_n$ şeklinde ifade edilir. Burada her a_i kategorik C değerini ifade eder ve i . pozisyonda bir metal bağlanma olup olmadığına göre 0 veya 1 değerini alır.

Bu problemi çözmek için, k uzunluğunda bir çerçeve belirlenir ve a_i merkezi kalıntı (Histidin veya Sistein) olmak üzere, dizilimden a_{i+k} ile a_{i-k} arasında temsili bir alt dizilim çıkarılır. Bu çalışmada çerçeve uzunluğu deneysel olarak $k=7$ optimal değer olarak bulunmuştur. Deneysel çalışmamızın akış diyagramı Şekil 2.15’de gösterilmektedir.

2.4.1. Nokta Kabul Edilen Mutasyon (PAM) Skarlama Matrisi

Amino asit skarlama matrisi, amino asit pozisyonlarına göre indekslenmiş 20x20 boyutunda bir tablodur ve tablodaki X, Y pozisyonu ise amino asit X’in amino asit Y ile hizalama skorunu vermektedir. Nokta Kabul Edilen Mutasyon (PAM), Dayhoff vd. [66] tarafından proteinlerdeki amino asit yer değiştirmelerini sunan Markov modelidir. PAM, bir amino asit kalıntısını bir diğeriyle değiştiren nokta mutasyonlarının, evrim sırasında bir gen soyuna dahil edildiği hızın bir ölçüsüdür. PAM, iki genin veya proteinin homolojisini belirlemek için kullanılan bir puanlama matrisi geliştirmek için kullanılır. Nokta mutasyon sayısına göre çeşitleri vardır. Bu çalışmada en yaygın kullanılanlardan biri olan ve Şekil 2.16’da gösterilen PAM120 matrisi kullanılmıştır. PAM120, her 100 amino asit için 120 nokta mutasyonu için olasılığını verir.



Şekil 2. 15. Metal bağlanma yeri tahmini akış diyagramı. a) Çinko metali ile bağlı (m_1, m_2) üç boyutlu 5X6T proteinini. b) protein diziliminden örnek bir kesit. c) metal bağlı Sistein ve Histidin aminoasitleri için çerçeve çıkarma ve öznitelik vektörü. d) derin öğrenme mimarileri.

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

Şekil 2. 16. PAM120 Skorlama Matrisi.

Bir öznitelikler vektörü $k \times 2$ amino asit içerir ve $F=f_1f_2f_3f_4....f_{14}$ şeklinde ifade edilir, burada her f_i değeri öznitelik vektöründeki i . amino asiti, k ise çerçeve uzunluğunu temsil etmektedir. Her PAM öznitelik vektörü P_i , $(k \times 2) \times 20 = 280$ boyuttan oluşmaktadır ve aşağıdaki şekilde ifade edilmektedir:

$$P_i = [p_1, p_2, p_3, p_4, ..., p_j, ..., p_{280}] \quad (12)$$

Burada her p_i , F öznitelikleri vektöründeki m . amino asit değerini göstermektedir. PAM öznitelik kümesine ait örnek bir kesit Şekil 2.17’de verilmiştir.

2.4.2 Protein Kompozisyon Sunucusu (ProCos)

Protein dizilimlerini sayısal olarak temsil etme yöntemlerinden biri de farklı peptit kompozisyonlarının kombinasyonlarını hesaplamak için geliştirilen Protein Kompozisyon Sunucusu (ProCos)'dur. ProCos, birden fazla protein dizilim dosya formatındaki (FASTA) dizilimi işleme kapasitesine sahip bir sunucu uygulaması olarak geliştirilmiştir.

```

0,3,-3,-1,0,-3,-1,0,1,-3,-1,-3,-2,-2,-4,1,1,1,-7,-4,0,3,-3,-1,0,-3,-1,0,1,-3,-1,-3,-2,-2,-4,1,1,1,-7,-4,0,1,-1,-2,-
3,-4,0,-2,-2,-1,-3,-3,-2,-3,-5,6,1,-1,-7,-6,-2,1,-1,-2,-3,-4,0,-2,-2,-1,-3,-3,-2,-3,-5,6,1,-1,-7,-6,-2,-3,6,-1,-3,-
4,1,-3,-4,1,-2,-4,2,-1,-5,-1,-1,-2,1,-5,-3,-3,-4,-4,-5,-7,-2,-4,-5,-3,1,5,-4,3,0,-3,-4,-3,-3,-2,1,-1,-2,-2,-3,-3,-3,-
3,-4,-4,6,1,-3,1,0,-3,-2,0,-6,-2,3,0,-3,2,5,-7,1,3,0,0,-3,-5,-1,-4,-7,-3,0,-1,-8,-5,-3,1,-1,1,0,0,-2,-1,1,-2,-2,-4,-
1,-2,-3,1,3,2,-2,-3,-2,-3,6,-1,-3,-4,1,-3,-4,1,-2,-4,2,-1,-5,-1,-1,-2,1,-5,-3,0,-3,-3,-3,-3,-3,-2,-3,3,1,-4,1,-3,-
2,-2,0,-8,-3,5,-3,-4,-4,-5,-7,-2,-4,-5,-3,1,5,-4,3,0,-3,-4,-3,-3,-2,1,0,-3,1,3,-7,2,5,-1,-1,-3,-4,-1,-3,-7,-2,-1,-2,-
8,-5,-3,-3,6,-1,-3,-4,1,-3,-4,1,-2,-4,2,-1,-5,-1,-1,-2,1,-5,-3

0,3,-3,-1,0,-3,-1,0,1,-3,-1,-3,-2,-2,-4,1,1,1,-7,-4,0,0,-3,1,3,-7,2,5,-1,-1,-3,-4,-1,-3,-7,-2,-1,-2,-8,-5,-3,-
2,2,1,-1,-7,0,-1,-3,-2,-3,-4,5,0,-7,-2,-1,-1,-5,-5,-4,-1,-2,-2,-3,-3,-3,-4,-4,6,1,-3,1,0,-3,-2,0,-6,-2,3,1,-2,0,-
1,-3,-2,-2,-1,-3,0,-3,-1,-1,-4,-1,2,4,-6,-3,0,1,-2,0,-1,-3,-2,-2,-1,-3,0,-3,-1,-1,-4,-1,2,4,-6,-3,0,1,-4,0,0,-4,-3,-
1,5,-4,-4,-5,-3,-4,-5,-2,1,-1,-8,-6,-2,3,-3,-1,0,-3,-1,0,1,-3,-1,-3,-2,-2,-4,1,1,1,-7,-4,0,0,-3,1,3,-7,2,5,-1,-1,-
3,-4,-1,-3,-7,-2,-1,-2,-8,-5,-3,-3,1,2,0,-4,3,-1,-4,7,-4,-3,-2,-4,-3,-1,-2,-3,-3,-1,-3,-3,-4,-5,-7,9,-7,-7,-4,-4,-3,-
7,-7,-6,-6,-4,0,-3,-8,-1,-3,1,-1,1,0,0,-2,-1,1,-2,-2,-4,-1,-2,-3,1,3,2,-2,-3,-2,-3,-4,-4,-5,-7,-2,-4,-5,-3,1,5,-
4,3,0,-3,-4,-3,-3,-2,1,-1,-1,4,2,-5,0,1,0,2,-2,-4,1,-3,-4,-2,1,0,-4,-2,-3

```

Şekil 2. 17. PAM ile elde edilen öznelitik kümesinden 280 özellik içeren iki örnek.

ProCos ile belirli bir protein sekansından dört farklı karakteristik veri elde edilmektedir. Bu karakteristik bilgiler, Sözde amino asit kompozisyonu (PAAC), Amfifilik sözde amino asit kompozisyonu (APAAC), Beş-faktörlü çözüm skoru (5FSS), Fizyokimyasal özellikler (PC)'dir. Bu parametreler, proteinlerin hücre altı lokalizasyonunun tahmininde, proteinlerin fonksiyonel olarak karakterize edilmesinde ve 2-boyutlu jellerden protein tanımlanmasında kullanılmaktadırlar [67].

PAAC, sekans sırası bilgilerini tamamen kaybetmeden ayrı bir modele sahip bir protein sekansını temsil etmek için kullanılır. Bir protein örneğinin PAAC bileşimi 20'den fazla faktör ile temsil edilir, burada ilk 20 faktör amino asit bileşenlerinin bir kombinasyonunu, ek faktörler ise sekans sıra bilgilerinin bir kısmını çeşitli modlar aracılığıyla birleştirir. Şekil 2.18'de gösterildiği gibi PAAC ile 6 parametre hesaplanmaktadır: hidrofobiklik, hidrofiliklik, yan zincir kütlesi, pK1 (alpha-COOH), pK2 (NH3), pI (at 25oC) [68]. PAAC ile elde edilen öznelitik kümeinden örnek bir kesit Şekil 2.19'da verilmiştir.

APAAC, amino asit kompozisyonu ile aynı forma sahiptir ancak bir peroteinin sekans sırası ve hidrofobik ve hidrofilik amino asitlerin zinciri boyunca dağılımı ile ilgili daha fazla bilgi içerir. İlk 20 sayı amino asitlerin kombinasyonunu, sonraki 2 lambda ile hidrofobiklik ve hidrofiliklik dağılım modellerini yansıtan bir dizi korelasyon faktörüdür [69].

5FSS ile verilen protein zincirinin beş faktöriyel skoru hesaplanmaktadır [70]. Faktör 1, gömülü kalıntılara karşı açıkta kalan kalıntıları, polariteye karşı polarite, hidrofilikliğe karşı

hidrofobiklik, serbest enerjiye karşı bağlanmamış enerjinin bir bölümünde eş zamanlı değişkenliği yansıtır. Faktör 2, çeşitli amino asitlerin sarmal, dönüş veya bobin gibi ikincil yapı konfigürasyonları ile ilişkisini temsil eden ikincil bir yapı faktörüdür. Faktör 3, moleküler boyut veya hacim ile ilgilidir. Faktör 4, çeşitli proteinlerdeki bağıl amino asit kompozisyonunu yansıtır. Faktör 5, izoelektrik nokta ve net yük üzerinde yüksek katsayılı elektrostatik yükü ifade eder [71]. PC ile verilen bir protein zincirinin fiziko kimyasal değerleri hesaplanmaktadır [72].

Amino asit	Hidrofobiklik	Hidrofiliklik	Kütle	pK1	pK2	pI
A	0.62	-0.5	15.0	2.35	9.87	6.11
C	0.29	-1.0	47.0	1.71	10.78	5.02
D	-0.90	3.0	59.0	1.88	9.60	2.98
E	-0.74	3.0	73.0	2.19	9.67	3.08
F	1.19	-2.5	91.0	2.58	9.24	5.91
G	0.48	0.0	1.0	2.34	9.60	6.06
H	-0.40	-0.5	82.0	1.78	8.97	7.64
I	1.38	-1.8	57.0	2.32	9.76	6.04
K	-1.50	3.0	73.0	2.20	8.90	9.47
L	1.06	-1.8	57.0	2.36	9.60	6.04
M	0.64	-1.3	75.0	2.28	9.21	5.74
N	-0.78	0.2	58.0	2.18	9.09	10.76
P	0.12	0.0	42.0	1.99	10.60	6.30
Q	-0.85	0.2	72.0	2.17	9.13	5.65
R	-2.53	3.0	101.0	2.18	9.09	10.76
S	-0.18	0.3	31.0	2.21	9.15	5.68
T	-0.05	-0.4	45.0	2.15	9.12	5.60
V	1.08	-1.5	43.0	2.29	9.74	6.02
W	0.81	-3.4	130.0	2.38	9.39	5.88
Y	0.26	-2.3	107.0	2.20	9.11	5.63

Şekil 2. 18. PAAC tarafından kullanılan parametre değerleri.

```

1,14.2857142857143,0,0,0,0,14.2857142857143,7.14285714285714,7.14285714285714,0,0,14.28571
42857143,7.14285714285714,0,0,0,0,14.2857142857143,0,21.4285714285714,0.204322622154421
,0,0,0,0.204322622154421,0.10216131107721,0.10216131107721,0,0,0.204322622154421,0.10216
131107721,0,0,0,0,0.204322622154421,0,0.306483933231631,1.51263626946173,2.497654519920
66,1.38681568452268,0.838673035355557,0.0524216162810743,0.0524216162810743,0.052421616
2810743,1.20946526404669,2.14383039988778,2.29429292064025,2.37520647372935,1.516687895
93433,1.46418531929241,2.61832504717747,0.614774697771686,0.818147405291631,0.903605756
179951,0,0,0.818147405291631,0.182192630565991,3.09025202679418,0.182192630565991,1.6220
9101859244,0.182192630565991,1.08145358397932,1.57486835790536,1.20946526404669,0.83867
3035355557,0.0524216162810743,1.20946526404669,0.349445638609167,2.37520647372935,1.209
46526404669,1.46418531929241,2.29429292064025,0.565029752907741,1.51668789593433,0.8181
47405291631,0.903605756179951,0.818147405291631,0.182192630565991,2.31193778825034,0,2.
49765451992066,0,0.36986840095088,1.57486835790536,1.20946526404669,1.38681568452268,0.
349445638609167,2.37520647372935,0.0524216162810743,1.51668789593433,2.14383039988778,
0.565029752907741,1.51668789593433,0,1.08204603875518,2.31193778825034

```

Şekil 2. 19. PAAC ile elde edilen veri kümesinden 100 özellik içeren bir örnek.

2.4.3. İkili Temsil (BR)

İkili temsil ile protein dizilimlerinin bir çeşit ikili kodlaması yapılmaktadır. Bu yöntemle, her bir çerçeve vektörü ikili vektörlere dönüştürülebilir. İkili temsil, çerçeve vektöründeki her bir amino asidin bir 20'lik vektör ile temsil edilmesini sağlar. Burada amino asit sekans listesine göre çerçeve vektöründeki her köşegen amino asidin değeri 1'dir ve geri kalan 19 amino asidin değeri 0'dır. Tablo 2.2'de amino asitlerin ikili temsilleri gösterilmiştir.

Örneğin;

Histidin için ikili temsil = [0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0]

Sistein için ikili temsil = [0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]

Buna göre çerçevedeki her amino asit için 20 boyutlu bir vektör elde edilir. $(k \times 2)$ boyutlu çerçeve için Histidin ve Sistein amino asitleri hariç tutulur.

Her ikili temsil B_i öznitelik vektörü $(k \times 2) \times 20 = 280$ boyutludur ve aşağıdaki şekilde formüle edilir:

$$B_i = [b_1, b_2, b_3, b_4, \dots, b_j, \dots, b_{280}] \quad (13)$$

Burada, her b_j elemanı, F öznitelik vektörü içindeki m . amino asidin ikili temsil değerini göstermektedir.

Tablo 2. 2. Amino asitlerin ikili temsil olarak gösterilmesi

Amino asit	İkili Temsil																			
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

3. DENEYSEL BULGULAR

Çalışmada yer alan bütün deneysel uygulamalar derin öğrenme ağlarının uygulanması için GPU hızlandırmalı hesaplamaya izin veren bir derin öğrenme kütüphanesi olan Keras platformu üzerinde gerçekleştirilmiştir. Keras, Python dili ile yazılmıştır ve hem CPU hem de GPU destekli TensorFlow üzerinde çalışmaktadır. Uygulama Python programlama dili ile geliştirilmiştir ve veri kümelerinin saklanması için MySQL veritabanı kullanılmıştır.

Burada sunulan bütün modellerin eğitim ve testleri Intel Core i7 2.5 GHz CPU, 16 GB RAM ve NVIDIA GeForce GT 750M grafik kartına sahip bir makinede hesaplanmıştır. Giriş protein sekans öznitelik kümesinin büyüklüğüne göre tahmin süresi değişebilmektedir.

5-kat çapraz doğrulama ile değerlendirilerek, doğruluk, kesinlik, duyarlılık ve f-ölçütü'nden oluşan dört değerlendirme ölçütü kullanılmıştır. Bu ölçütler aşağıdaki denklemlerle ifade edilmektedir:

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

$$\text{Kesinlik} = \frac{TP}{TP+FP} \quad (15)$$

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \quad (16)$$

$$\text{F – skoru} = \frac{2TP}{2TP+FP+FN} \quad (17)$$

Burada TP, TN, FN ve FP değerleri sırası ile doğru pozitif, doğru negatif, yanlış negatif ve yanlış pozitif sayılarını ifade etmektedir. TP, doğru tahmin edilen pozitif örnek sayısını, TN doğru kabul edilen negatif örnek sayısını, FP yanlış pozitif olarak tahmin edilen negatif örnek sayısını, FN ise yanlış negatif olarak tahmin edilen pozitif örneklerin sayısını işaret etmektedir.

3.1. Derin Öğrenme Modelleri

Çalışmada üç farklı derin öğrenme ağ modeli kullanılarak öznitelik kümeleri üzerinde deneyler yapılmıştır. Bu modeller üzerinde en iyi performansı elde edecek bir çok parametrik test yapılmıştır.

3.1.1. CNN Modeli

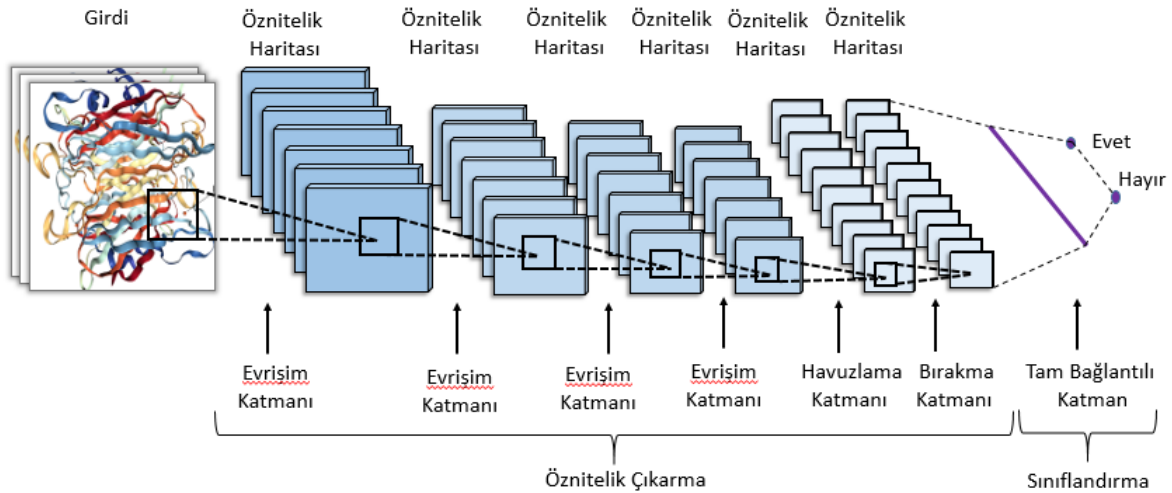
Evrişimsel sinir ağlarının imaj tanıma konusundaki yüksek performansı literatür çalışmalarından bilinmektedir. Bu çalışmada öncelikle her protein dizilimi 2-boyutlu gri tonlamalı imajlarına dönüştürülerek giriş verileri elde edilmiştir.

PAM ve ikili temsil yöntemleriyle çıkarım yapılan 280 boyutlu öznitelikler için 28x10'luk giriş matrisleri, ProCos özelliklerinden 5FSS ile çıkarılan 120 boyutlu öznitelikler için 12x10, PAAC ile çıkarılan 100 öznitelik için 10x10 ve PC ile çıkarılan 30 boyutlu öznitelik için 6x5 boyutlarında giriş matrisleri elde edilmiştir. Elde edilen bu veri kümeleri için 2-boyutlu CNN mimarisi kullanılmıştır.

CNN mimarisi Şekil 3.1’de gösterildiği gibi 4 evrişim katmanı, 2 havuzlama katmanı, 2 bırakma ve 2 tam bağlı katmandan oluşmaktadır. Her katman öğrenilmiş ağırlıklarla bir x girişini alıp bir sapma ekleyerek ve bazı f doğrusal olmayan fonksiyonları uygulayarak bir çıktı üretir. Bir nöronün çıktısı (Φ) şu şekilde hesaplanabilir:

$$\Phi = f(\sum_i W_i + b) \quad (18)$$

burada, önceki katmanların x_i çıktılarını topluyoruz.



Şekil 3. 1. CNN Modeli.

CNN mimarisinde tüm katmanlarda ReLU aktivasyon fonksiyonu, ağımızın çıkışında ise sigmoid fonksiyonu kullanılmıştır. Bu modelde, her evrişim katmanı 3x3 piksellik filtrelerden oluşan alt örneklerden oluşmaktadır. CNN mimarisinde optimum katman sayısı ve optimum

girdi-çıkıtı değerlerine ulaşmak için çok sayıda deneyler yapılmıştır. 280 boyutlu bir girdi verisi için 4 evrişim katmanı ve 2 tam bağlı katmandan oluşan mimari sonucunda 2 olası durum için bir çıkıtı üretmektedir. Optimum katman sayıları ve bu katmanların girdi-çıkıtı sayıları Tablo 3.1'de gösterilmiştir.

Tablo 3. 1. CNN modeli girdi kanalları sayısı ve çıkıtı sayıları

katman	girdi	çıkıtı
evrişim-1	280	128
evrişim-2	128	96
evrişim-3	96	64
evrişim-4	64	32
tam bağlı-1	32	64
tam bağlı-2	64	2

CNN mimarisi, filtre boyutu, filtre sayısı, adım aralıkları, dolgu, bırakma oranı veya aktivasyon fonksiyonu seçimi gibi birçok tasarım seçeneğine sahiptir. Bunların yanında devir sayısı, parti büyüklüğü ve optimize edici oranı gibi parametreleri vardır.

CNN mimarimizde optimize edici olarak Adam kullanıldı ve birçok denemeden sonra parti büyüklüğü optimal olarak 100, devir sayısı olarak 2 ile 100 arasında yapılan deneyler sonucunda Tablo 3.2'de gösterildiği gibi 20 optimal değeri elde edilmiştir. Bu modele ait parametreler Tablo 3.3'de gösterilmiştir.

Tablo 3. 2. CNN mimarisinin devir sayısına bağlı çalışma süreleri ve sonuçları

Devir sayısı	Kesinlik (%)	Duyarlılık (%)	Çalışma süresi (saniye)
2	75	81	44,5
10	78	81	49,7
20	79	82	52,1
30	77	81	52,8
50	76	80	53,4
100	75	79	53,26

Tablo 3. 3. CNN model parametreleri

#çerçeve boyutu	15
devir aralığı	{2,100}
parti büyüklüğü	100
#evrişim katmanı	4
maks. havuzlama boyutu	2
#tam bağlı katman	2
bırakma	{0.15, 0.25}
#toplam parametre	202370

3.1.2. GRU Modeli

Bu çalışmada tekrarlayan sinir ağ algoritması olarak GRU kullanılmıştır. GRU mimarisinin CNN mimarisinde olduğu gibi birçok parametre ve tasarım seçeneği bulunmaktadır. Bu parametreler Tablo 3.4’de verilmiştir. Tekrarlayan sinir ağı yapısı aşağıdaki formüllerle tanımlanabilir:

$$h_t = f(W_x x + W_h h_{t-1} + b_n) \quad (19)$$

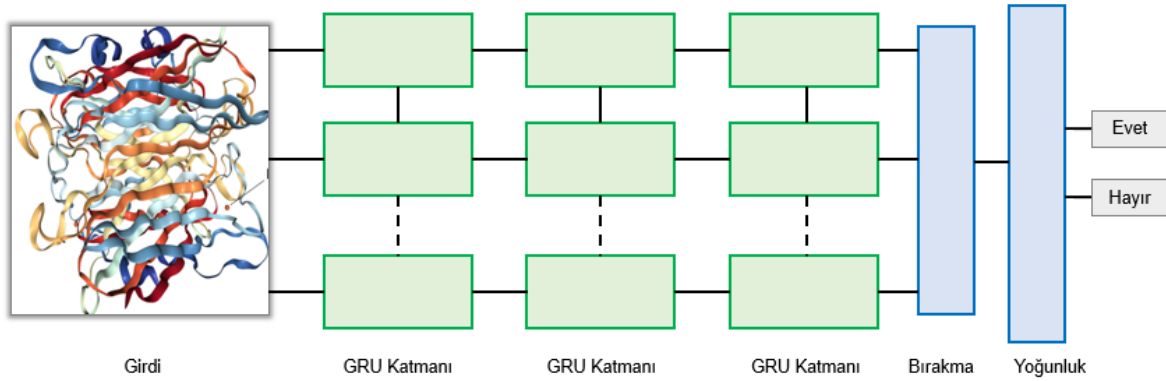
$$y_t = \sigma(W_y h_t + b_y) \quad (20)$$

$T=t$ zamanında, tekrarlayan ağ, önceki $x_1, x_2, x_3...x_{t-1}$ ve mevcut x_t girişinden gelen bilgileri hatırlayabilir.

Tablo 3. 4. GRU model parametreleri

çerçeve boyutu	15
giriş büyüklüğü	256
GRU katman sayısı	3
devir aralığı	{2,50}
parti büyüklüğü	50
tam bağlı katman sayısı	1
bırakma	{0.2}
aktivasyon, hata,optimize fonksiyonu	Softmax, Sparse categorical crossentropy, Adam
Toplam parametre	993538

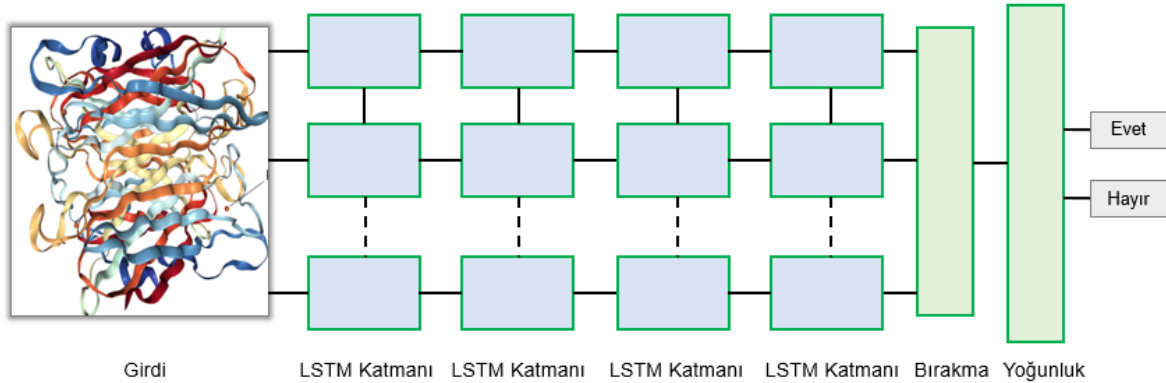
Bu çalışmada Şekil 3.2'de gösterildiği gibi 3 GRU katmanı, 1 bırakma ve 1 tam bağlı katmandan oluşan GRU modeli kullanılmıştır. Aktivasyon fonksiyonu olarak softmax, hata fonksiyonu olarak sparse categorical crossentropy ve optimizasyon fonksiyonu olarak Adam kullanılmıştır. Optimum sonuçlara ulaşınca kadar 2 ile 5 arasında GRU katmanı ve 2 ile 50 arasında devir aralığı deneyi yapılmıştır.



Şekil 3. 2. GRU Model.

3.1.3. LSTM Modeli

Uzun-kısa süreli bellek (LSTM) sinir ağları, RNN'nin hücreğine bir belleğin eşlik ettiği, uzun süreli bağımlılık problemlerinden kaçınmak için tasarlanmış özel bir RNN mimarisidir. Bu mimaride kullanılan parametre değerleri Tablo 3.5'de verilmiştir. LSTM mimarisi Şekil 3.3'te gösterildiği gibi 4 LSTM katmanı, 1 bırakma ve 1 tam bağlı katmandan oluşmaktadır. Aktivasyon fonksiyonu olarak softmax, hata fonksiyonu için Sparse categorical crossentropy ve optimizasyon fonksiyonu için Adam kullanılmıştır. Optimum sonuçlara ulaşmak için 2 ile 50 arasında devir aralığı, 1 ile 6 arasında LSTM katmanı kullanılarak deneyler yapılmıştır.



Şekil 3. 3. LSTM Modeli.

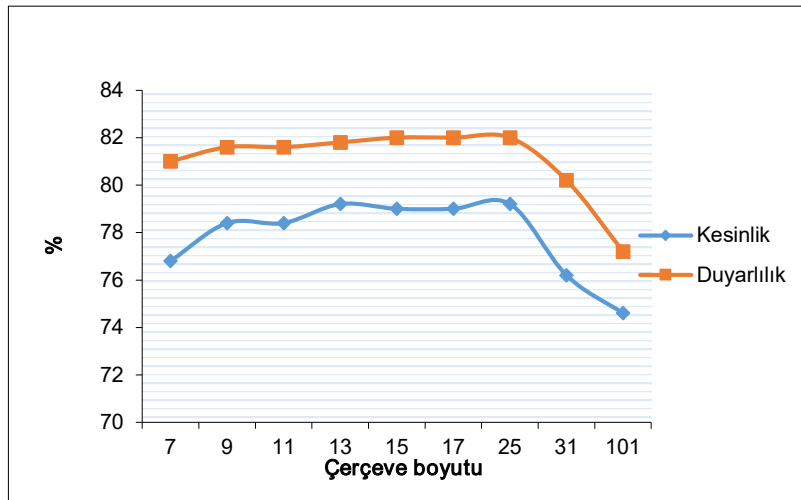
Tablo 3. 5. LSTM model parametreleri

çerçeve boyutu	15
girdi boyutu	50
LSTM katman sayısı	4
devir aralığı	{2,50}
parti büyüklüğü	50
Tam bağlı katman sayısı	1
bırakma	{0.2}
aktivasyon, hata, optimizasyon fonksiyonu	Softmax, Sparse cat. crossentropy, Adam
Toplam parametre	71902

3.2. Metal Bağlanma Yeri Bulguları

Çalışmada her bir protein dizilimi için 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 31, 101 boyutlu çerçeveler çıkarılmıştır. Bu çerçeve boyutlarına karşılık PAM öznitelikleri ve CNN ile elde edilen kesinlik ve duyarlılık sonuçlarını içeren grafiksel gösterim Şekil 3.4'de gösterilmiştir.

Bu sonuçlara göre 5, 17 ve 25 çerçeve ile elde edilen duyarlılık değerleri %82, kesinlik değerleri ise 15 ve 17 için %79, 25 için %79,2 değerleri ile birbirine yakın sonuçlar çıkmıştır. Çerçeve boyutu 15'ten önce ve 25'i geçtikten sonra sonuçların kalitesinin azaldığı görülmektedir. En düşük sonuç 5 ve 101'de elde edilmiştir. Hesaplama süreleri de dikkate alınarak optimal çerçeve değeri W=15 seçilerek, sonraki çalışmalarda bu değer dikkate alınmıştır.



Şekil 3. 4. Bütün çerçeve boyutları için CNN ve PAM kullanılarak elde edilen sonuçlar.

Protein metal bağlanma yeri tahmini için önerilen yöntemin kabiliyetini göstermek için ilk olarak CNN mimarisi ve PAM öznitelik kümesi kullanılarak 3 ve 5 kat çapraz doğrulama ile bağımsız test veri kümesi üzerinde değerlendirme yapılmıştır. 3-kat çapraz doğrulama için kesinlik değeri %77,92, duyarlılık değeri %80,52 elde edilirken, 5-kat çapraz doğrulama için %79 kesinlik, %82 duyarlılık değerleri elde edilmiştir. En iyi sonuçların 5-kat çapraz doğrulama ile elde edildiği dikkate alınarak, çalışmada veri kümesini test etmek için SVM-BRNN [12]'de olduğu gibi 5-kat çapraz doğrulama kullanılmıştır.

CNN, RNN ve LSTM'den oluşan üç farklı derin öğrenme ağ modeli, PAM, ikili temsil ve ProCos öznitelik kümeleri için uygulanmış ve aşağıdaki sonuçlar elde edilmiştir. Burada en iyi performans PAM öznitelik kümesi kullanarak CNN modeli üzerinde elde edilmiştir. En yüksek sonuçlar PAM kullanarak CNN ve GRU modelinde, PAAC kullanarak LSTM modelinde elde edildiği görülmektedir.

Derin öğrenme mimarilerimizin ortalama çalışma süreleri Tablo 3.6'da gösterilmiştir. Burada, mimarilerde kullanılan katman sayısı ve devir sayıları dikkate alınarak bir devir işlemi için ortalama çalışma süresi saniye olarak tespit edilmiştir.

Tablo 3. 6. Derin öğrenme mimarilerinin 1 devir işlemi için ortalama çalışma süreleri

mimari	ort. çal. süresi (saniye)	devir sayısı	katman sayısı
CNN	52	20	4
GRU	135,6	13	3
LSTM	45,6	20	4

3.2.1. PAM Matirisi ile Elde Edilen Veri Kümesi Bulguları

Protein dizilimlerinde Histidin ve Sistein merkezli çerçevelerden oluşan 22861 PAM öznitelik verisi için sonuçlar elde edilmiştir. Bu veri kümesi için derin öğrenme modellerimiz CNN, GRU ve LSTM ile ayrı ayrı deneyler yapılarak sonuçlar Tablo 3.7'de gösterilmiştir. Bu sonuçlara göre en yüksek değerler olan %79 kesinlik, %82 duyarlılık ve %80,5 f-skoru CNN ile elde edilmiştir. CNN ile elde edilen doğruluk değeri %86 ile diğerlerine göre %5 daha yüksektir. LSTM ve GRU prrformans değerleri birbirine yakın çıkmıştır.

Tablo 3. 7. PAM öznitelikleri ve derin öğrenme modelleri ile elde edilen sonuçlar

	Kesinlik (%)	Duyarlılık (%)	F-skoru (%)	Doğruluk (%)
CNN	79	82	80,5	86
GRU	74,6	81	73,6	80,6
LSTM	73,2	80,8	73,6	80,8

Yine aynı örnek veri kümesine geleneksel makine öğrenme metodlarından SVM, Naïve Bayes, AdaBoost ve Bagging uygulanmış ve Tablo 3.8’de verilen sonuçlar elde edilmiştir. Topluluk öğrenme yöntemlerinden AdaBoost ve Bagging sonuçları birbirine yakın değerler olarak görülmektedir. SVM ile elde edilen %65 kesinlik değerinin diğer bütün yöntemlerden daha düşük olduğu, Naïve Bayes ile elde edilen %73 duyarlılık değeri de diğer yöntemlerden daha düşük olduğu görülmektedir.

Tablo 3. 8. PAM öznitelikleri ve makine öğrenme yöntemleri ile elde edilen sonuçlar

	Kesinlik (%)	Duyarlılık (%)	F-skoru (%)	Doğruluk (%)
SVM	65	79	73	81
Naïve Bayes	72	73	71	73
AdaBoost	75	78	75	81
Bagging	76	78	75	81

3.2.2. İkili Temsil ile Elde Edilen Veri Kümesi Bulguları

İkili temsil ile elde edilen sonuçlar Tablo 3.9’da verilmiştir. En yüksek sonuçlar PAM’da olduğu gibi burada da CNN ile elde edilmiştir. %81,4 doğruluk, %77,8 kesinlik, %81,2 duyarlılık ve %75,6 f-skoru olarak GRU ve LSTM değerlerinden daha iyi performans göstermiştir. Burada kesinlik değerleri üç modelde de birbirine yakın çıkmıştır.

Tablo 3. 9. İkili temsil öznitelik kümesi ile elde edilen sonuçlar

	Kesinlik (%)	Duyarlılık (%)	F-skoru (%)	Doğruluk (%)
CNN	77,8	81,2	75,6	81,4
GRU	68,8	81	72,2	80,8
LSTM	66	81	72	81

Kesinlik sonuçlarına göre CNN %77,8 ile en yüksek değer elde edilirken, LSTM ile %66 en düşük sonuç elde edilmiştir. Bu sonuçlar Tablo 3.7'de ki PAM sonuçları ile karşılaştırıldığında ikili temsil ile elde edilen sonuçların daha düşük olduğu görülmektedir.

3.2.3. ProCos ile Elde Edilen Veri Kümesi Bulguları

ProCos karakteristikleri PC, 5FSS ve PAAC için sonuçlar sırası ile Tablo 3.10, Tablo 3.11 ve Tablo 3.12'de gösterilmiştir. PC için en düşük sonuç CNN ile elde edilmiştir. Diğer değerlendirme kriterlerine göre bu üç model için birbirine yakın sonuçlar ortaya çıkmıştır. PC ile elde edilen sonuçlar, PAM ve ikili temsil ile elde edilen sonuçlardan düşüktür. Bunun nedeni, PC özniteliklerinin sayısı PAM ve ikili temsil öznitelik sayısından daha az olması olabilir.

Tablo 3. 10. PC öznitelik kümesi ile elde edilen sonuçlar

	Kesinlik (%)	Duyarlılık (%)	F-skoru (%)	Doğruluk (%)
CNN	67	82	73	81
GRU	71	78	72	80
LSTM	72,4	81	72,6	81

Tablo 3.11, PAAC öznitelikleri için CNN'nin bütün değerlendirme kriterleri için en iyi sonuçların elde edildiğini göstermektedir. GRU modeli ile burada elde edilen sonuçlar, PC ve 5FSS öznitelikleri ile elde edilen sonuçlardan daha düşük olduğu görülmektedir. Buradaki LSTM modeli ile elde edilen sonuçlar, PC ve 5FSS öznitelik kümeleri sonuçlarından daha yüksektir.

Tablo 3. 11. PAAC öznitelik kümesi ile elde edilen sonuçlar

	Kesinlik (%)	Duyarlılık (%)	F-skoru (%)	Doğruluk (%)
CNN	75	81	75	81
GRU	70,6	72,8	71,4	72,8
LSTM	73,8	80,6	73,2	80,6

5FSS öznitelik kümesinden elde edilen sonuçlara göre en düşük sonuçlar Tablo 3.12'de de görüldüğü gibi GRU modeli ile, en yüksek sonuçlar ise CNN ile elde edilmiştir. Burada LSTM modeli ile elde edilen sonuçlar PAAC öznitelik kümesi ile elde edilen sonuçlardan daha düşük olduğu görülmektedir.

Tablo 3. 12. 5FSS öznitelik kümesi ile elde edilen sonuçlar

	Kesinlik (%)	Duyarlılık (%)	F-skoru (%)	Doğruluk (%)
CNN	75,4	81,4	72,4	80,8
GRU	71	74,4	72,4	74,4
LSTM	73,2	80,4	73,2	80,4

3.2.4. Performan Karşılaştırması

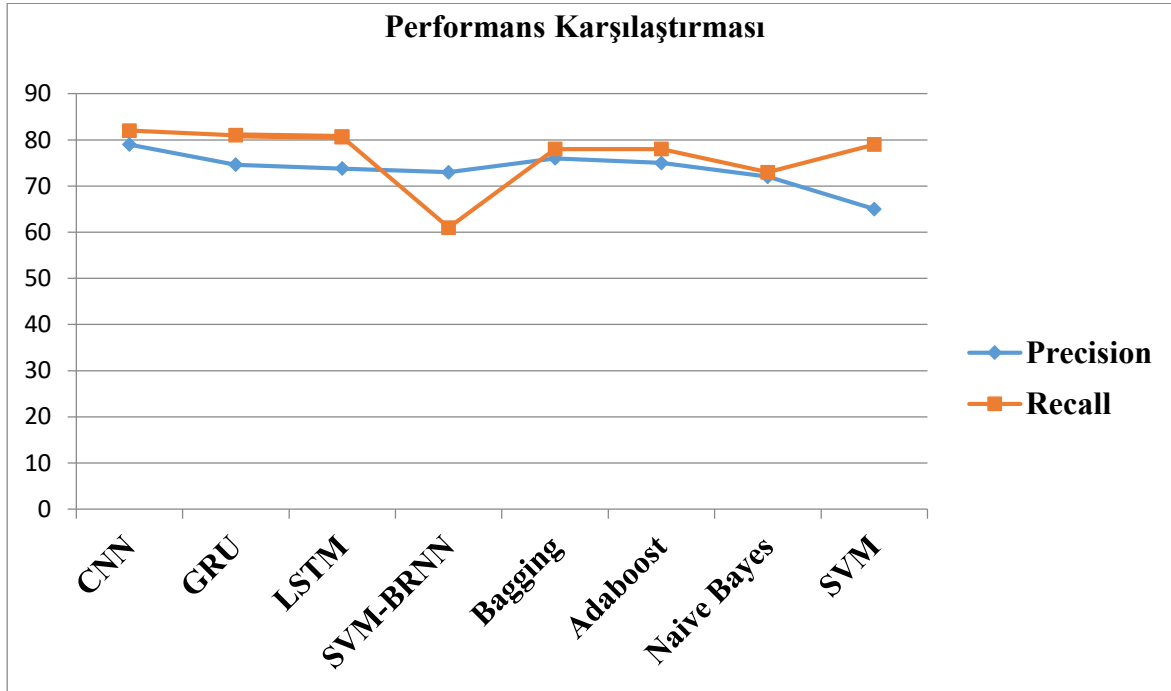
Çalışmada her bir model ile elde edilen en yüksek sonuçlar ile kıyaslama veri kümesi ile en son teknoloji çalışması olan SVM-BRNN [12], kesinlik ve duyarlılık değerlendirme ölçütlerine göre karşılaştırılmıştır. SVM-BRNN modeli, SVM tahminleri ve BRNN kombinasyonuna dayanan hibrit bir çözüm sunmak üzere tasarlanmıştır. Burada birkaç geçiş metali ile bağlanan Histidin ve Sistein'i tanımlamak için iki aşamalı bir makine öğrenmesi yaklaşımı tanıtılmaktadır. İlk aşamada Histidin ve Sistein'in durumunu tahmin eden bir SVM, ikinci aşamada ise metal bağlı veya metal bağlı olmayan Histidin ve Sistein'i ayırmak için bir BRNN modeli kullanılmaktadır. Çalışmada protein dizilimi temsili için PSSM (Konuma özel skorlama matrisi) öznitelik çıkarma metodu kullanılmıştır. Bu çalışma da geçiş metalleri ile bağlı Histidin ve Sisteinleri %73 kesinlik ve %61 duyarlılık değerleri ile tahmin etmektedir. Bu tez çalışmasında önerilen üç farklı derin öğrenme modeli ile elde edilen sonuçlar Tablo 3.13'te gösterildiği gibi SVM-BRNN'den daha yüksek bir başarımlı performansını sağlamıştır.

Tablo 3.13'de görüldüğü gibi en yüksek sonuçlar PAM öznitelik kümesi kullanılarak CNN modeli ile elde edilmiştir. Bu sonuçlar diğer modellerimizden ve SVM-BRNN'den daha yüksek bir performans göstermektedir. SVM-BRNN sonuçlarına göre, CNN modeli ile elde edilen duyarlılık değeri %21, kesinlik değeri de %6 artmıştır. GRU ve LSTM modeli ile elde edilen sonuçlarda duyarlılık değeri sırasıyla %20 ve %19,6; kesinlik değeri ise %1,6 ve %0,8 artmıştır. Derin öğrenme modelleri yanında geleneksel makine öğrenme teknikleri SVM, AdaBoost, Naive Bayes ve Bagging ile PAM öznitelik kümeleri test edilmiş ve sonuçlar Tablo 3.13'e eklenmiştir. Boosting ve Bagging topluluk öğrenme teknikleri çok fazla makine öğrenme algoritmalarının tahminlerini birleştiren bir tekniktir. Bu tekniklerde karar ağaçları temel öğrenici algoritması olarak seçilmiştir. Sonuçlara göre Bagging ve AdaBoost teknikleri SVM-BRNN'den daha iyi performans göstermiştir.

Tablo 3. 13. Performans karşılaştırması

Model	Öznitelik	Kesinlik (%)	Duyarlılık (%)
CNN	PAM	79	82
GRU	PAM	74,6	81
LSTM	PAAC	73,8	80,6
SVM-BRNN ^[10]	PSSM	73	61
Bagging	PAM	76	78
AdaBoost	PAM	75	78
Naïve Bayes	PAM	72	73
SVM	PAM	65	79

Öznitelik kümelerine bağlı sonuçlar Şekil 3.5'te grafiksel olarak gösterilmiştir. Sonuçlarda üç derin öğrenme modeli için beş farklı öznitelik kümesi uygulanarak elde edilen kesinlik ve duyarlılık sonuçları verilmiştir.



Şekil 3. 5. Performans karşılaştırması.

3.3. Metal Bağlanma Yerlerinin Geometrik Tahimini

Burada, metale bağlı olduğu bilinen bir protein zinciri içerisinde ki Sistein ve Histidin kalıntılarının orjinal bağlanma duruma göre (bağlı veya bağısız) elde ettiğimiz tahminlerimizin (bağlı veya bağısız) sayısını ve hangi metale bağlandıkları tespit edilerek metal bağlanma yerlerinin geometrisi yeniden değerlendirilmiştir. Bölüm 3.2’de metal bağlanma yeri tahmininde kullanılan aynı veri kümesi içerisinde ki 233 metal bağlı protein seçilerek geometrik tahmin için kullanılmıştır.

Veri kümesindeki Sistein ve Histidin kalıntılarının hangi metallere bağlandığına ait bilgiler Tablo 3.14’de gösterilmiştir. Bu çalışmada, protein dizilimi içerisindeki tahmin edilen Sistein ve Histidin kalıntılarının kaç tanesinin bir metale bağlandığını ve bu metallerin neler olduğu tespit etmeye çalışılmıştır. Sonuçlar Tablo 3.15’de gösterilmektedir. Bölüm 3.1.1’de oluşturduğumuz CNN modeli ve bölüm 2.4.1’de ki PAM öznitelik çıkarma yöntemi ile elde edilen sonuçlar, bölüm 3.2.4’te gösterildiği gibi en yüksek başarımlı değerleridir. Buna dayanarak metal bağlanma yerlerinin geometrik tespiti için de CNN + PAM modeli kullanılmıştır.

Tablo 3. 14. Veri kümesindeki Histidin ve Sistein bağlı metal sayıları

Metal	Toplam Sistein ve Histidin Sayısı	%	Sistein	Histidin
ZN	468	49,79	310	158
FE	312	33,19	182	130
CU	58	6,17	13	45
CD	57	6,06	47	10
AG	12	1,28	12	0
GA	8	0,85	5	3
HG	8	0,85	7	1
NI	7	0,74	0	7
MG	6	0,64	0	6
CA	4	0,43	1	3
Toplam	940	-	577	363

Veri kümesinde Zn, Fe, Cu, Cd, Ag, Ga, Hg, Ni, Mg ve Ca'dan oluşan 10 metal, toplam 940 kalıntı ve bu kalıntıların 577'si Sistein ve 363'ü Histidin kalıntılarını içermektedir. Bu veri kümesinde Histidin ve Sistein kalıntılarına bağlı metallerin yaklaşık %82'si geçiş metalleri grubundan Zn ve Fe oluşturmaktadır. Bu metallerin %61'i Sistein'e, %39'u Histidin kalıntısı ile bağ yapmaktadır. Ni ve Mg metalleri hiçbir Sistein kalıntısı ile bağ yapmazken, Ag metali ise Histidin kalıntısı ile hiçbir bağ yapmamaktadır.

Tablo 3. 15. Tahmin sonucu bağlı metal sayıları

Metal	Toplam Sistein ve Histidin Sayısı	%	Sistein	Histidin
ZN	123	56,42	114	40
FE	77	35,32	37	9
CD	8	3,67	8	0
CU	6	2,75	5	1
GA	2	0,92	2	0
CA	1	0,46	0	1
AG	1	0,46	1	0
Toplam	218		167	51

CNN+PAM modeli uygulaması sonucu oluşan sonuçlara göre Histidin ve Sistein ile bağ yapan metallerin %92'sini geçiş metallerinden Zn ve Fe oluşturduğu görülmektedir. Bu metallerin %76'sı Sistein'e, %24'ü ise Histidin kalıntısı ile bağ yapmışlardır. Ca metali hiçbir Sistein kalıntısı ile bağ yapmaz iken, Cd, Ga ve Ag metalleri de hiçbir Histidin kalıntısı ile bağ yapmadığı görülmüştür.

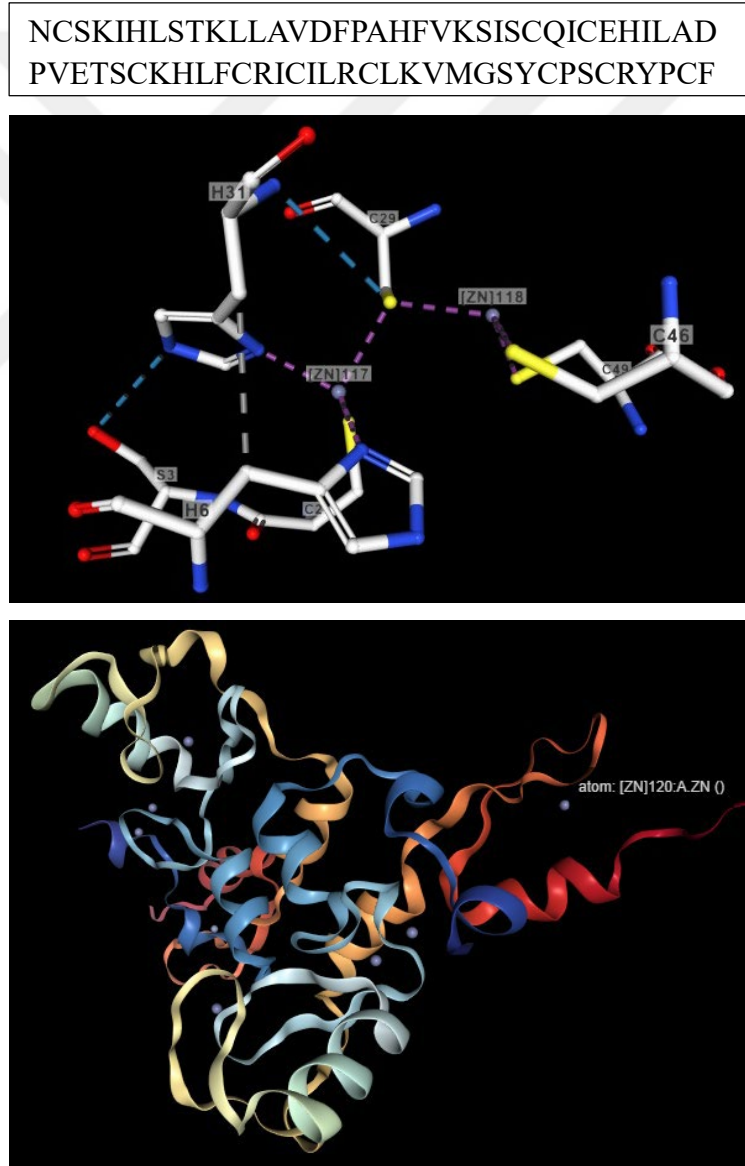
Histidin ve Sistein kalıntıları sadece bir metale bağlanabileceği gibi aynı anda birden fazla metale de bağlanabilmektedir. Buna ilişkin sonuçlar Tablo 3.16'te gösterilmiştir.

Tablo 3. 16. Histidin ve Sistein 'in aynı anda bağlandığı metal sayıları

	1 Metal	2 Metal	3 Metal	4 Metal
Sistein	138	29	3	0
Histidin	38	13	4	3
Toplam	176	42	7	3

Sistein kalıntısı aynı anda 3 farklı yerde aynı 3 metale birden bağlanırken, 29 yerde 2 metale birden, 138 yerde ise 1 metale bağlanmıştır. Histidin kalıntısı ise, aynı anda 3 farklı yerde aynı 4 metale, 4 farklı yerde aynı 3 metale, 13 farklı yerde aynı 2 metale, 38 farklı yerde aynı 1 metal ile bağ yapmışlardır. Toplam olarak Histidin ve Sistein kalıntıları aynı anda 176 farklı yerde 1 metal ile bağ yaparken, 42 farklı yerde 2 metal ile 7 farklı yerde ise 3 metal ile birden bağ yaptıkları tespit edilmiştir.

Veri kümesinde yer alan 1RMD proteinin dizilimi, 2 ve 3 boyutlu yapısı Şekil 3.6'da gösterilmiştir. Bu proteine ilişkin bilinen metal bağlanma yerleri, bu çalışmada tahmin edilen bağlanma yerleri ve hangi metale bağlandığına ilişkin bilgiler Şekil 3.7'de verilmiştir.



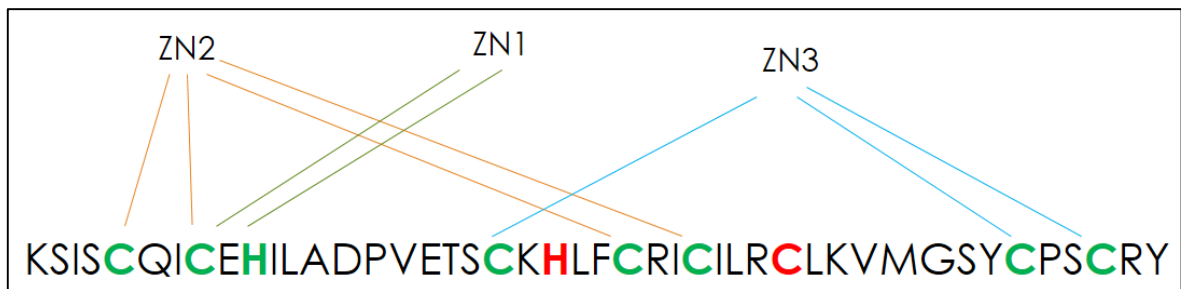
Şekil 3. 6. 1RMD proteini dizilimi (üstte), 2 boyutlu (ortada) ve 3 boyutlu yapısı (altta).

Şekil 3.7’de ki 1RMD proteininin metal bağlanma yerleri kişisine göre yeşil renkle gösterilen Sistein ve histidin kalıntıları Zn metali ile bağ yapana, kırmızı renkliler ise bağ yapmayan yerleri göstermektedir. Bu kesitte, protein 15 farklı yerde Sistein ve Histidin kalıntıları ile Zn metali arasında bağlantı yaptığı bilinmektedir. CNN+PAM modeli ile yaptığımız çalışmada 8 farklı yerde Zn metali ile bağ yaptığı tespit edilmiştir. Yine bu proteinin 4 farklı yerde Histidin ve Sistein kalıntıları ile bağ yapmadığı bilinmektedir. CNN+PAM modeli ile bu 4 farklı yerdeki bağ yapılmayan yerler doğru tespit edilmiştir. Bu protein sonuçlarına göre. %66,6 oranında bir doğru tespit elde edildiği görülmektedir.

Protein:	1RMD														
	NCSKIHLSTKLLAVDFPAHFVKSISQCICEHILADPVETSCKHL														
Bağ:	1	1				0		1	1	1				1	1
Tahmin:	0	0				0		1	1	1				0	1
Metal:	Zn1	Zn1						Zn2	Zn1	Zn1				Zn3	Zn3
	FCRICILRCLKVMGSYCPSCRYPCFPTDLESPVKSFLNILNSLM														
Bağ:	1	1	0			1	1	0							
Tahmin:	1	1	0			1	1	0							
Metal:	Zn2	Zn2				Zn3	Zn3								
	VKCPAQDCNEEVSLKYNHVVSSHKESK														
Bağ:	1	1				0	1	1							
Tahmin:	0	0				0	0	0							
Metal:	Zn4	Zn4					Zn4	Zn4							

Şekil 3. 7. 1RMD proteini için metal bağlanma yerlerinden bir kesit.

Şekil 3.8’de 1RMD protein diziliminden alınan kesit üzerinde metal bağlanma yeri doğru tahmin edilen kalıntılar (yeşil renkli) ve bu kalıntıların bağlandığı ZN metali gösterilmiştir.



Şekil 3. 8. 1RMD proteini için metal bağlanma yerlerinden grafiksel bir kesit.

Şekil 3.9’da ise 1RMD protein dizliminde yer alan Histidin ve Sistein amino asitlerinin pozisyonları (Yer), bu amino asitlerin metale bağlanma durumları (MBS), çalışmada elde edilen tahmin (Tahmin) ve bağlandığı metal (Metal) gösterilmektedir.

<u>Yer</u>	<u>AA</u>	<u>MBS</u>	<u>Tahmin</u>	<u>Metal</u>
2	C	1	0	ZN1
6	H	1	0	ZN1
19	H	0	-	-
26	C	1	1	ZN2
29	C	1	1	ZN1,ZN2
31	H	1	1	ZN1
41	C	1	0	ZN3
43	H	1	1	ZN3
46	C	1	1	ZN2
49	C	1	1	ZN2
53	C	0	-	-
61	C	1	1	ZN3
64	C	1	1	ZN3
68	C	0	-	-
91	C	1	0	ZN4
96	C	1	0	ZN4
107	H	0	-	-
108	H	1	0	ZN4
112	H	1	0	ZN4

Şekil 3. 9. 1RMD proteininin ZN metali ile 4 farklı bağ yeri.

Burada daha açık bir şekilde metal bağlanma yerlerinin geometrik tespitleri verilmiştir. Yeşil renkle gösterilen kalıntılar birinci Zn metali ve geometrik yerini, sarı renkli kalıntılar ikinci Zn metali ve geometrik yerini, mavi renkli kalıntılar üçüncü Zn metali ve geometrik bağlantı yerinin doğru tahmin edildiğini, siyah renkli kalıntı ve Zn metalleri ise yanlış tahmin edildiğini göstermektedir. Burada protein diziliminin 29 numaralı pozisyonunda yer olan Sistein kalıntısı iki Zn metali ile (Zn1, Zn2) aynı anda bağ yaptığı görülmektedir.

4. TARTIŞMA VE SONUÇ

Bir proteinin metal ile bağlanma yerinin tespit edilmesi, o proteinin biyolojik olarak hangi görevleri olduğu, hangi öneme sahip olduğu gibi çok çeşitli özellikleri hakkında bilgiler edinilmesini sağlayabilir. Bunun için literatürde çok farklı teknikler kullanılarak çok sayıda çalışma yürütülmüştür. Son yıllarda gelişen teknoloji ile derin öğrenme ağlarının kullanımını artırmış ve farklı alanlarda yüksek başarımlar göstermesini sağlamıştır. Bu kapsamda protein metal bağlanma yeri tespiti için derin öğrenme ağlarının kullanılması da önem kazanmıştır.

Bu çalışmada protein metal bağlanma yerlerini tahmin etmek için üç farklı derin öğrenme mimarisi sunulmuştur. Protein dizilimlerini sayısal olarak temsil etmek için üç farklı öznitelik çıkarma tekniği kullanılmıştır. Derin öğrenme modelleri farklı kriterlere göre kapsamlı bir şekilde test edilmiş ve performansları karşılaştırılmıştır. Yine aynı veri kümesi içerisinde seçilen veriler kullanılarak metal bağlanma yerlerinin geometrik tespiti de yapılmıştır.

Protein dizilimlerinde Histidin ve Sistein kalıntılarını merkeze alan optimal çerçeve uzunluğu elde edilmiştir. Bu çerçeveler üzerinde hesaplama işlemi yapılabilmesi için PAM skorlama matrisi, protein kompozisyonları ve ikili temsilden oluşan öznitelik çıkarım metodları kullanılarak numerik olarak temsil edilmesi sağlanmıştır.

CNN ve GRU derin öğrenme modellerimizde PAM ile çıkarılan öznitelik kümelerinde daha iyi performans göstermiştir. Bu nedenle PAM öznitelik kümelerinin protein dizilimlerini daha iyi temsil ettiği söylenebilir. Burada PAM'ın bir amino asit kalıntısını bir başkasıyla değiştiren mutasyonların evrim sırasında bir gen hattında dahil edilme hızının bir ölçüsü olması etkin olabilir. Ayrıca, PAM ile elde edilen çok sayıda öznitelik de derin öğrenme teknikleri performansı için önemli bir kriterdir. Proteinlerin metallerle bağ yapması fizyokimyasal bir işlem olmasına bağlı olarak PAAC ile elde edilen özniteliklerden de daha iyi sonuç elde edilmesi beklenirdi ama evrimsel bir işlem olan PAM öznitelikleriyle daha yüksek sonuçlar elde edilmesi tartışılabilir. CNN'nin yüksek sayıda öznitelik sayısına sahip veri kümelerinde daha iyi performans gösterdiği bilinmektedir.

İkili temsil öznitelik kümesi ile elde edilen sonuçlar diğer öznitelik çıkarma yöntemlerine göre daha düşük çıkmıştır. Bu, öznitelik kümesinin yalnızca 1 ve 0 değerlerinden ibaret olmasından kaynaklanabilir. LSTM ve GRU yöntemleri sıralı veri tahmini problemlerinde daha iyi performans göstermektedir ancak bu çalışmada dizilimler 2 boyutlu bir imaja dönüştürüldükleri için CNN ile karşılaştırıldıklarında daha düşük performans göstermişlerdir.

Geleneksel makine öğrenme yöntemlerinden Naïve Bayes ve SVM ile yapılan testlerde PAM skorumla matrisi öznitelikleri ile duyarlılık sonuçları SVM-BRNN çalışmasından daha yüksek çıkmıştır. Aynı şekilde kolektif sınıflandırma algoritmaları AdaBoost ve Bagging algoritmaları PAM özniteliklerine uygulanarak hem kesinlik hem de duyarlılık değerleri daha yüksek sonuç elde edilmiştir.

Derin öğrenme ağları, yüksek öznitelik sayısı ile daha iyi performans göstermektedir. Burada ProCos protein kompozisyon öznitelikleri ile PAM öznitelikleri birleştirilerek daha büyük bir öznitelik kümesi elde edilip yeni çalışmalar yapılabilir. Çalışmadaki derin öğrenme ağ modelleri SVM-BRNN'de kullanılan PSSM ile elde edilen öznitelik kümelerine uygulanarak sonuçlar karşılaştırılabilir. Bu çalışma aynı zamanda, Histidin ve Sistein dışındaki diğer amino asitlerin protein metal bağlama yeri, tüm metaloproteinlerle veya Fe, Cu, Ni, Zn gibi belirli bir metal iyonu ile tahmin edilmesinde kullanılabilir. Bu çalışmada sunulan derin öğrenme modelleri başka öznitelik kümelerine kolayca uygulanabilir. Yine bu çalışmada protein temsili için kullanılan öznitelik çıkarma yöntemleri başka veri kümelerine uygulanabilir.

Derin öğrenme ağları, diğer biyoinformatik problemlerde olduğu gibi bu çalışmada da diğer hesaplama yöntemlerinden daha iyi performans göstermiştir. Bu problem için en iyi performansa sahip SVM-BRNN çalışması ile karşılaştırıldığında, kıyaslama veri kümesi üzerinde önerilen bu yöntemler daha yüksek performans göstermiştir. Derin öğrenme modellerimizin iyi performans göstermesi, protein metal bağlama yeri tahmin problemi için potansiyel uygulama olduğunu göstermektedir. Derin öğrenme modellerinin protein metal bağlanma yerlerinin tahmin edilmesindeki bu kapsamlı araştırması, gelecekteki metal bağlanma çalışmaları için rekabetçi bir araç sağlayabilir.

Protein dizilimlerinden metal bağlama yerlerinin geometrisini belirlemek, yani bir metal iyonunun koordinasyonunda hangi kalıntıların birlikte yer aldığını tespit etmek de önem arz

etmektedir. Bu çalışmada, metal bağlanma yerlerinin geometrik olarak tespit edilmesi yeniden değerlendirilmiştir. Çalışmada derin öğrenme ağ mimarileri kullanılarak elde edilen sonuçlara göre protein dizilimlerinin metal bağlanma yeri geometrileri de tespit edilmiştir. Metal ile bağlandığı bilinen dizilimler üzerinde, derin öğrenme mimarileri ile elde edilen sonuçlar incelenerek, bir dizilimin hangi bölgesinde kaç metal ile bağ olduğu bilgilerine ulaşılabilmektedir. Bir protein dizilimi kalıntısı herhangi bir metal ile bağ yapmayacağı gibi birden fazla metal ile de bağ yapabileceği sonucu da elde edilebilmektedir. Bazı proteinler aynı metal ile birçok yerden bağ yapabilirken bazıları ise farklı metaller ile bu bağları yapabilmektedirler. Çalışmada Histidin ve Sistein kalıntılarının 1 metal, 2 metal, 3 metal ve 4 metal ile yaptığı bağlar tespit edilmiştir.

Metallerle en çok bağ yaptığı bilinen Histidin ve Sistein kalıntılarının, dizilim içerisinde hiç bağ yapmadığı yerler de görülmektedir. Çalışmada, Zn ve Fe metalleri en çok bağ yapan geçiş metalleri olarak tespit edilmiştir. Metal bağlanma yerlerinin geometrik tespiti ile protein yapı ve fonksiyonları hakkında daha fazla bilgilere erişilebilmektedir.

Bu çalışmada tanımlanan modellerin, diğer yapısal biyoloji problemlerinde uygulama potansiyeline sahip olduğuna inanılmaktadır. Gelecekteki çalışmalarda, protein metal bağlama yeri tahminine, protein 3-boyutlu yapısı ve diğer protein yapı bilgileri de eklenerek tahmin edilen bu sonuçlar iyileştirilebilir.

Çalışmada elde edilen başarının iyileştirilmesi için, oluşturulan öznitelik oluşturma yöntemlerinin yeni veri kümeleri ve yeni sınıflandırma yöntemleri ile zenginleştirilmesi gerektiği düşünülmektedir.

Protein metal bağlanma bölgelerinde meydana gelen mutasyonlar bazı hastalıkların ortaya çıkmasına sebep olduğu bilinmektedir. Bu çalışma ile elde edilen yaklaşımlar bu durumun tespit edilmesine katkı sağlayabilecektir.

Önerilen bu derin öğrenme yaklaşımları, protein-metal etkileşimi, protein ikincil yapı tahmini ve protein fonksiyon tahmini gibi birçok potansiyel uygulamaya sahip olacaktır.

KAYNAKLAR

- [1] M. Zarándi, J. Szolomájer, “Amino acids: chemistry diversity and physical properties,” *Amino Acids, Peptides and Proteins*, Volume 42, pages:1-84, 2018.
- [2] C. Zhu, Y. Gao, H. Li, et al, “Characterizing hydrophobicity of amino acid side chains in a protein environment via measuring contact angle of a water nanodroplet on planar peptide network,” *Proc. Natl. Acad. Sci*, 113, 12946-12951, 2016.
- [3] <https://www.rcsb.org/structure/1AXC>, (Accessed: Aug.16, 2020).
- [4] S. M. Marino and V.N. Gladyshev, “Cysteine function governs its conservation and degeneration and restricts its utilization on protein surfaces,” *Journal of molecular biology*, 404(5), 902–916. 2010.
- [5] S. M. Liao, Q. S. Du, J. Z. Meng, Z. W. Pang, R. B. Huang, “The multiple roles of histidine in protein interactions,” *Chemistry Central journal*, 7(1), 44. 2013.
- [6] İ. Haberal and H. Oğul, "DeepMBS: Prediction of Protein Metal Binding-Site Using Deep Learning Networks," *2017 Fourth International Conference on Mathematics and Computers in Sciences and in Industry (MCSI)*, Corfu, pp. 21-25, 2017.
- [7] İ. Haberal and H. Oğul, "Prediction of Protein Metal Binding Sites Using Deep Neural Networks," *Molecular Informatics*, 38,1800169, 2019.
- [8] J.S. Sodhi, K. Bryson, L.J. McGuffin, J.J. Ward, L. Wernisch, D.T. Jones, “Predicting metal-binding site residues in low-resolution structural models,” *Journal of Molecular Biology*, 2004.

- [9] H. Deng, G. Chen, W. Yang, J.J. Yang, "Predicting calcium—binding sites in proteins—A graph theory and geometry approach," *Proteins: Structure, Function, and Bioinformatics*, 64(1): 34–42. 2006.
- [10] C.T. Lin, K.L. Lin, C.H. Yang, I.F. Chung, C.D. Huang, Y.S. Yang, "Protein metal binding residue prediction based on neural networks," *Int J Neural Syst.*, 2005.
- [11] H.H. Lin, L.Y. Han, H.L. Zhang, C.J. Zheng, B. Xie, et al., "Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach," *BMC Bioinformatics*, 2006.
- [12] A. Passerini, M. Punta, A. Ceroni, B. Rost, P. Frasconi, "Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks," *Proteins*, 2006.
- [13] K. Goyal, S.C. Mande, "Exploiting 3D structural templates for detection of metal-binding sites in protein structures," *Proteins*, 2008.
- [14] J.B. Andrew, "Predicting small ligand binding sites in proteins using backbone structure," *Bioinformatics*, 2008.
- [15] M. Babor, S. Gerzon, B. Raveh, V. Sobolev, M. Edelman, "Prediction of transition metal-binding sites from apo protein structures," *Proteins*, 2008.
- [16] M. Lippi, A. Passerini, M. Punta, B. Rost, P. Frasconi, "MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence," *Bioinformatics*, 2008.
- [17] A. Passerini, M. Lippi, P. Frasconi, "MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence," *Nucleic Acids Res.*, 2011.

- [18] M. Brylinski, J. Skolnick, "FINDSITE-metal: Integrating evolutionary information and machine learning for structure-based metal binding site prediction at the proteome level," *Proteins*, 2011.
- [19] W. Zhao, M. Xu, Z. Liang, B. Ding, L. Niu, H. Liu, M. Teng, "Structure-based de novo prediction of zinc-binding sites in proteins of unknown function," *Bioinformatics*, 2011.
- [20] C.H. Lu, Y.F. Lin, J.J. Lin, C.S. Yu, "Prediction of metal ion-binding sites in proteins using the fragment transformation method," *PLoS One*, 7(6), e39252. 2012.
- [21] C. Zheng, M. Wang, K. Takemoto, T. Akutsu, Z. Zhang, J. Song, "An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins", *PLoS One*, vol. 7, no. 11, 2012.
- [22] A. Passerini, M. Lippi, P. Frasconi, "Predicting metal-binding sites from protein sequence," *IEEE/ACM Trans Comput Biol Bioinform*, 2012.
- [23] C. Andreini, G. Cavallaro, S. Lorenzini, A. Rosato, "Metal PDB: A database of metal sites in biological macromolecular structures," *Nucleic Acids Research*, 41, D312–319. 2013.
- [24] Z. Chen, Y. Wang, Y.F. Zhai, J. Song, Z. Zhang, "Zinc Explorer: An accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences," *Molecular BioSystems*, 9, 2213–2222. 2013
- [25] Z. Liu, Y. Wang, C. Zhou, Y. Xue, W. Zhao, H. Liu, "Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, Volume 1844, Issue 1, Part B, Pages 171-180, 2014.
- [26] H. Zheng, M.D. Chordia, D.R. Cooper, et al., "Validation of metal-binding sites in macromolecular structures with the CheckMyMetal web server," *Nature Protocols*, 2014.

- [27] W. Zhou, G.W. Tang, R.B. Altman, "High Resolution Prediction of Calcium-Binding Sites in 3D Protein Structures Using FEATURE," *Journal of Chemical Information and Modeling*, 2015.
- [28] W. He, Z. Liang, M. Teng, L. Niu, "mFASD: a structure-based algorithm for discriminating different types of metal-binding sites," *Bioinformatics*, Volume 31, Issue 12, Pages 1938–1944, 2015.
- [29] X. Hu, Q. Dong, J. Yang, Y. Zhang, "Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals," *Bioinformatics*, 32(21), 3260–3269. 2016.
- [30] Y.F. Lin, C.W. Cheng, C.S. Shih, J.K. Hwang, C.S. Yu, C.H. Lu, "MIB: Metal Ion-Binding Site Prediction and Docking Server," *Journal of Chemical Information and Modeling*. 56 (12), 2287-2291. 2016.
- [31] X. Cao, X. Hu, X. Zhang, S. Gao, C. Ding, Y. Feng, et al, "Identification of metal ion binding sites based on amino acid sequences," *PLoS One*, 12(8): e0183756. 2017.
- [32] H. Li, D. Pi, Y. Wu and C. Chen, "Integrative Method Based on Linear Regression for the Prediction of Zinc-Binding Sites in Proteins," *IEEE Access*, vol. 5, pp. 14647-14656, 2017.
- [33] J. Song, C. Li, C. Zheng, J. Revote, Z. Zhang, G.I. Webb, "MetalExplorer, a bioinformatics tool for the improved prediction of eight types of metal-binding sites using a random forest algorithm with two-step feature selection," *Current Bioinformatics*, 12(6), 480-489. 2017.
- [34] A. Srivastava and M. Kumar, "Prediction of zinc binding sites in proteins using sequence derived information," *Journal of Biomolecular Structure and Dynamics*, 36:16, 4413-4423. 2018.

- [35] H. Li, D. Pi, C. Chen, H. Li, "A Novel Prediction Method for Zinc-Binding Sites in Proteins by an Ensemble of SVM and Sample-Weighted Probabilistic Neural Network," *Access IEEE*, vol. 7, pp. 186147-186157, 2019.
- [36] I. Halperin, D. S. Glazer, S. Wu, R.B. Altman, "The Feature Framework for Protein Function Annotation: Modeling New Functions, Improving Performance, and Extending to Novel Applications," *BMC Genomics*, 9. S2.10. 2008.
- [37] B. Alipanahi, A. DeLong, M. Weirauch, et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat Biotechnol*, 33, 831–838. 2015.
- [38] T. Jo, J. Hou, J. Eickholt, J. Cheng, "Improving Protein Fold Recognition by Deep Learning Networks," *Nature*, 2015.
- [39] J. Zhou, O.G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," *Nat Methods*. 12:931–4, 2015.
- [40] D.R. Kelley, J. Snoek, J.L. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks," *Genome Res*. 26:990–9. 2016.
- [41] M. Hamanaka, K. Taneishi, H. Iwata, J. Ye, J. Pei, J. Hou, Y. Okuno, "CGBVS-DNN: Prediction of Compound-Protein Interactions Based on Deep Learning," *Molecular Informatics*, 2016.
- [42] K. Tian, M. Shao, Y. Wang, J. Guan, S. Zhou, "Boosting compound-protein interaction prediction by deep learning," *Methods*. 110:64–72. 2016.
- [43] S. Wang, J. Peng, J. Ma, J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Sci Rep*. 6:18962. 2016.
- [44] S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model," *PLOS Computational Biology* 13(1): e1005324. 2017.

- [45] J.J.A. Armenteros, C.K. Sønderby, S.K. Sønderby, H. Nielsen, O. Winther, “DeepLoc: prediction of protein subcellular localization using deep learning,” *Bioinformatics*, Volume 33, Issue 21, Pages 3387–3395. 2017.
- [46] X. Pan, H.B. Shen, “Predicting rna-protein binding sites and motifs through combining local and global deep convolutional neural networks,” *Bioinformatics*. 34(20): 3427-36. 2018.
- [47] M. Kulmanov, M.A. Khan, R. Hoehndorf, J. Wren, “DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier,” *Bioinformatics*, 34(4), 660–668. 2018.
- [48] D.T. Jones and S.M. Kandathil, “High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features,” *Bioinformatics*, 34(19), 3308–3315. 2018.
- [49] K. Degtyarenko and S. Contrino, “COMe: the ontology of bioinorganic proteins,” *BMC structural biology*, 4, 3. 2004.
- [50] K.Y. Hsin, Y. Sheng, M.M. Harding, P. Taylor, M.D. Walkinshaw, “MESPEUS: a database of the geometry of metal sites in proteins,” *J. Appl. Cryst.* 41, 963–968. 2008.
- [51] K. Nakamura, A. Hirai, M. Altaf-Ul-Amin, H. Takahashi, “MetalMine: a database of functional metal-binding sites in proteins,” *Plant biotechnology*, 26(5), 517-521. 2009.
- [52] K. Hemavathi, K.M. Kalaivani, A. Udayakumar, G. Sowmiya, J. Jeyakanthan, K. Sekar, “MIPS: metal interactions in protein structures,” *J. Appl. Cryst.* 43, 196–199. 2010.
- [53] H. Choi, H. Kang, H. Park, “MetLigDB: a web-based database for the identification of chemical groups to design metalloprotein inhibitors,” *J. Appl. Cryst.* 44, 878–881. 2011.

- [54] Y. Valasatava, A. Rosato, L. Banci, C. Andreini, “MetalPredator: a web server to predict iron-sulfur cluster binding proteomes,” *Bioinformatics*. 32, 2850–2852. 2016
- [55] R. Levy, M. Edelman, V. Sobolev, “Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates,” *Proteins*, 76, 365–374. 2009.
- [56] C. Andreini, G. Cavallaro, A. Rosato, Y. Valasatava, “MetalS2: A Tool for the Structural Alignment of Minimal Functional Sites in Metal-Binding Proteins and Nucleic Acids,” *Journal of Chemical Information and Modeling* 2013 53 (11), 3064–3075. 2013.
- [57] M.N. Wass, L.A. Kelley, M.J.E. Sternberg, “3DLigandSite: predicting ligand-binding sites using similar structures,” *Nucleic Acids Res.* 38, W469–W473. 2010.
- [58] C. Andreini, G. Cavallaro, S. Lorenzini, “FindGeo: a tool for determining metal coordination geometry,” *Bioinformatics*, 28, 1658–1660. 2012.
- [59] C. Cortes, V. Vapnik, “Support-vector networks,” *Machine Learning*, **20**, 273–297, 1995.
- [60] R. Blagus, L. Lusa, “Boosting for high-dimensional two-class prediction,” *BMC Bioinformatics*, 16, 300, 2015.
- [61] L. Breiman, “Bagging predictors,” *Machine Learning*, 24, 123–140, 1996.
- [62] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning,” *Nature*. 521:436–44. 2015.
- [63] H. A. Song, S.Y. Lee, “Hierarchical Representation Using NMF, 20th International Conference on Neural Information Processing (ICONIP 2013),” Daegu, Korea, 2013.
- [64] M. Seonwoo, L. Byunghan, Y. Sungroh, “Deep learning in bioinformatics,” *Briefings in Bioinformatics*, Volume 18, Issue 5, Pages 851–869, 2017.

- [65] I. Wallach, M. Dzamba, A. Heifets, “AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery,” *arXiv:1510.0285*, Oct. 2015.
- [66] D. T. Jones, W. R. Taylor, J. M. Thornton, “The rapid generation of mutation data matrices from protein sequences,” *Comput. Appl. Biosci.* 8, 275–82. 1992.
- [67] L. Rishishwar, N. Mishra, B. Pant, K. Pant, K.R. Pardasani, “ProCoS - PROtein COmposition Server,” *Bioinformation*, 5(5): 227. PMC: 3040505. 2010.
- [68] K.C. Chou, “Prediction of protein cellular attributes using pseudo-amino-acid-composition,” *Proteins: Structure, Function, and Genetics*, 43: 246-255. 2001.
- [69] K.C. Chou, “Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes,” *Bioinformatics*, 21: 10-19. 2005.
- [70] Chou, K.C. and Cai Y.D. “Prediction of membrane protein types by incorporating amphipathic effects,” *J. Chem Inf Model.* 45(2):407-13. 2005.
- [71] J. Atchley, A.D. Zhao, D. Fernandes, T. Drüke, “Solving the protein sequence metric problem,” *Proc Natl Acad Sci.* 102:6395–6400. 2005.
- [72] L. Yu, Y. Zhang, I. Gutman, Y. Shi, M. Dehmer, “Protein Sequence Comparison Based on Physicochemical Properties and the Position-Feature Energy Matrix,” *Scientific reports*, 7, 46237, 2017.

EKLER

EK 1: Çalışmada Kullanılan Protein Listesi

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1	101m_	31	1aop_	61	1bl9A	91	1cxpA
2	1a0b_	32	1aq8A	62	1blu_	92	1cxpC
3	1aoo_	33	1aqe_	63	1bnlA	93	1cxyA
4	1a1tA	34	1ar1A	64	1bor_	94	1cy5A
5	1a3z_	35	1ast_	65	1bs4A	95	1d09B
6	1a42_	36	1aw5_	66	1bt0A	96	1d0qA
7	1a4aA	37	1aw6_	67	1bt1A	97	1d4uA
8	1a4lA	38	1axq_	68	1bvb_	98	1d6uA
9	1a56_	39	1b0jA	69	1bxiB	99	1d7bA
10	1a6bB	40	1b0nA	70	1byfA	100	1d8wA
11	1a6f_	41	1b0nB	71	1c3rA	101	1d9yA
12	1a6yA	42	1b0pA	72	1c52_	102	1dax_
13	1a71A	43	1b13A	73	1c6oA	103	1dcqA
14	1a73A	44	1b25A	74	1c7kA	104	1ddzA
15	1a7d_	45	1b3dA	75	1c7mA	105	1de0A
16	1a7i_	46	1b4tA	76	1c9qA	106	1dfeA
17	1a7tA	47	1b4uB	77	1cc1L	107	1dfsA
18	1a7vA	48	1b55A	78	1cc5_	108	1dftA
19	1a7w_	49	1b57A	79	1cfmA	109	1diiC
20	1a8e_	50	1b66A	80	1cfwA	110	1dj7A
21	1a8vA	51	1b71A	81	1cfzA	111	1djnA
22	1adt_	52	1b7vA	82	1chc_	112	1dk4A
23	1af2A	53	1b80A	83	1cjxA	113	1dl6A
24	1afrA	54	1b9mA	84	1cl6A	114	1dmtA
25	1ah7_	55	1b9rA	85	1cnoA	115	1dmwA
26	1ahjA	56	1bbo_	86	1cnqA	116	1do6A
27	1aigL	57	1bf6A	87	1co4A	117	1dq3A
28	1ajyA	58	1bh5A	88	1cvmA	118	1dqsA
29	1ak0_	59	1biqA	89	1cvrA	119	1drtA
30	1aol_	60	1bkcA	90	1cw0A	120	1dt6A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
121	ldw0A	151	leqwA	181	lfre_	211	lgupA
122	ldwlB	152	leteA	182	lfs7A	212	lgvyA
123	ldxrC	153	levkA	183	lfu9A	213	lgw6A
124	ldypA	154	lexkA	184	lfv5A	214	lgl1A
125	ldzwP	155	leyfA	185	lfvsA	215	lgycA
126	le08A	156	lf0jA	186	lfwqA	216	lgytA
127	le08D	157	lf22A	187	lfy2A	217	lgzhA
128	le0eA	158	lf2iG	188	lfyzA	218	lh0hB
129	le1hA	159	lf35A	189	lg12A	219	lh21A
130	le29A	160	lf37A	190	lg25A	220	lh2kA
131	le3dA	161	lf3z_	191	lg47A	221	lh31B
132	le4uA	162	lf4sP	192	lg5cA	222	lh96A
133	le53A	163	lf56A	193	lg71A	223	lhc1_
134	le5sA	164	lf81A	194	lg8fA	224	lh9A
135	le7dA	165	lf88A	195	lg8jB	225	lhp1A
136	le7pB	166	lfaq_	196	lg9kA	226	lhslA
137	le7pC	167	lfb1A	197	lgaqB	227	lhu9A
138	leb6A	168	lfcdC	198	lgaxA	228	lhuxA
139	leb7A	169	lfauA	199	lgkkA	229	lhwtC
140	lec5A	170	lffuA	200	lgks_	230	lhyiA
141	lee8A	171	lfioA	201	lgl4A	231	li1iP
142	lef4A	172	lfj3A	202	lgm4A	232	li3jA
143	lefzA	173	lfjgD	203	lgmwA	233	li3qA
144	leg9A	174	lfjgN	204	lgnf_	234	li3qC
145	leh6A	175	lfn9A	205	lgof_	235	li3qI
146	lehbB	176	lfnoA	206	lgp5A	236	li3qL
147	lei6A	177	lfp0A	207	lgpc_	237	li6nA
148	leilA	178	lfp4A	208	lgqgA	238	li7hA
149	lejeA	179	lfp4B	209	lgqwA	239	li7wA
150	lep1B	180	lfqtA	210	lgt7A	240	lia6A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
241	lia9A	271	ljnrB	301	lldjB	331	lnn7A
242	liag_	272	ljpuA	302	llfwA	332	lnuiA
243	libyA	273	ljq4A	303	lli5A	333	lnx4A
244	liciA	274	ljr3A	304	llm3B	334	lnztA
245	lid2A	275	ljr7A	305	llml_	335	lo12A
246	lim5A	276	ljreA	306	llpvA	336	lo7dA
247	lio0A	277	ljt1A	307	llr5A	337	loccB
248	liqzA	278	ljwhC	308	lltlA	338	loccF
249	lirxA	279	ljwqA	309	llv3A	339	locyA
250	lisuA	280	lk0vA	310	lm1pA	340	lodhA
251	litqA	281	lk24A	311	lm2oA	341	loi0A
252	liuz_	282	lk2fA	312	lm4oA	342	loj6A
253	liwlA	283	lk2yX	313	lm55A	343	lopmA
254	liymA	284	lk6wA	314	lm65A	344	loqjA
255	lizoA	285	lk7hA	315	lm7jA	345	lox7A
256	lj2oA	286	lk81A	316	lmg2D	346	lp1mA
257	lj3gA	287	lkbeA	317	lmj4A	347	lpft_
258	lj5lA	288	lkbpA	318	lmjgM	348	lpguA
259	lj6wA	289	lkcqA	319	lmkmA	349	lpi1A
260	lj79A	290	lkcw_	320	lmm2A	350	lpih_
261	lj7nA	291	lkeaA	321	lmr1C	351	lpmi_
262	ljb0A	292	lkfiA	322	lmwqA	352	lpyiA
263	lji3A	293	lk19A	323	lmwzA	353	lpzwA
264	lji4A	294	klrA	324	lmzbA	354	lq16A
265	ljjdA	295	lkqfC	325	ln0zA	355	lq16C
266	ljyA	296	lkqkA	326	ln25A	356	lq68A
267	ljk4A	297	lkwgA	327	ln2zA	357	lqh3A
268	ljkeA	298	ll1oC	328	lncs_	358	lqjA
269	ljm7A	299	ll5pA	329	lnj3A	359	lqtW
270	ljm7B	300	llbu_	330	lnkuA	360	lqx0A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
361	1qyp_	391	16vpA	421	1a4l_	451	1a8y_
362	1r3nA	392	1a05A	422	1a44_	452	1a93A
363	1rmd_	393	1a0aA	423	1a48_	453	1a93B
364	1tafA	394	1a0dA	424	1a4rA	454	1a95A
365	1tafB	395	1a0i_	425	1a50A	455	1a99A
366	1tbn_	396	1a0oB	426	1a50B	456	1a9nA
367	1tf3A	397	1a0sP	427	1a5e_	457	1a9nB
368	1udvA	398	1a14N	428	1a5nA	458	1a9o_
369	1ui0A	399	1a15A	429	1a5nB	459	1a9v_
370	1v3wA	400	1a17_	430	1a5v_	460	1a9xA
371	1vf5A	401	1a1mB	431	1a67_	461	1a9xB
372	1vhh_	402	1a1vA	432	1a6cA	462	1aa1C
373	1xer_	403	1a1w_	433	1a6jA	463	1aa2_
374	1yuiA	404	1a22A	434	1a6q_	464	1aa7A
375	1zbdB	405	1a22B	435	1a6r_	465	1aa8A
376	1zfd_	406	1a26_	436	1a6s_	466	1aab_
377	1zfo_	407	1a2cL	437	1a6x_	467	1aalA
378	2aw0_	408	1a2j_	438	1a76_	468	1aax_
379	2drpD	409	1a2kA	439	1a79A	469	1aazA
380	2hrvA	410	1a2t_	440	1a7aA	470	1ab9C
381	3caoA	411	1a2xB	441	1a7j_	471	1abv_
382	3hipA	412	1a2yB	442	1a7kA	472	1ac0_
383	3paz_	413	1a2zA	443	1a7m_	473	1acbI
384	117eA	414	1a34A	444	1a7s_	474	1acc_
385	11asA	415	1a39_	445	1a82_	475	1acj_
386	132l_	416	1a3c_	446	1a87_	476	1acmA
387	13pkA	417	1a3d_	447	1a8a_	477	1acx_
388	152l_	418	1a3k_	448	1a8d_	478	1ad1A
389	153l_	419	1a3p_	449	1a8o_	479	1adjA
390	16gsA	420	1a40_	450	1a8p_	480	1adoA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
481	ladr_	511	lajsA	541	larl_	571	laz0A
482	ladx_	512	lak2_	542	larzA	572	laz6_
483	lafal	513	lakhB	543	las4A	573	1b05A
484	lafh_	514	lal04	544	lat0_	574	1b09A
485	lafoA	515	lalu_	545	lata_	575	1b0aA
486	lafp_	516	lam4A	546	lauk_	576	1b0o_
487	lafvA	517	lam9A	547	laun_	577	1b0zA
488	lag1O	518	lamf_	548	lauq_	578	1b12A
489	lag2_	519	lamx_	549	lauuA	579	1b1cA
490	lag7_	520	lan1I	550	lauxA	580	1b1uA
491	lagg_	521	lan4A	551	lauyB	581	1b1yA
492	lagm_	522	la06A	552	lavaC	582	1b22A
493	lagqA	523	laocA	553	lavdA	583	1b2iA
494	lagt_	524	laod_	554	lavgI	584	1b2lA
495	lagy_	525	laohB	555	lavoA	585	1b2mA
496	lah1_	526	laoiA	556	lavpA	586	1b2pA
497	lah5_	527	laoiB	557	lavqA	587	1b33N
498	lahjB	528	laoiC	558	lawj_	588	1b34B
499	ahl_	529	laoiD	559	lax4A	589	1b35A
500	lahzA	530	laonO	560	lax8_	590	1b35B
501	lai4A	531	laoxA	561	laxb_	591	1b35C
502	lai4B	532	laoy_	562	laxh_	592	1b35D
503	laigH	533	lap4_	563	laxj_	593	1b3cA
504	laihA	534	lapj_	564	laxr_	594	1b3tA
505	laipA	535	lapq_	565	lay2_	595	1b41B
506	laipC	536	lapyA	566	lay7A	596	1b47A
507	lair_	537	lapyB	567	lay7B	597	1b49A
508	laisB	538	laq5A	568	layj_	598	1b4fB
509	laiw_	539	laqzA	569	laym4	599	1b4g_
510	lajkA	540	larb_	570	layoA	600	1b4rA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
601	1b4sA	631	1bal_	661	1bfd_	691	1bm9A
602	1b4uA	632	1bb8_	662	1bff_	692	1bmfA
603	1b54_	633	1bbg_	663	1bg0_	693	1bmlC
604	1b5fB	634	1bbn_	664	1bg8A	694	1bmp_
605	1b5oA	635	1bby_	665	1bgc_	695	1bmV2
606	1b62A	636	1bbzA	666	1bgk_	696	1bnb_
607	1b65A	637	1bc4_	667	1bgw_	697	1bo8A
608	1b6cA	638	1bcg_	668	1bh4_	698	1bob_
609	1b6cB	639	1bci_	669	1bh7_	699	1boeA
610	1b6g_	640	1bcmA	670	1bh8A	700	1boqA
611	1b6rA	641	1bcpA	671	1bh8B	701	1bosA
612	1b6tA	642	1bcpB	672	1bhe_	702	1bp7A
613	1b70A	643	1bcpD	673	1bhi_	703	1bpoB
614	1b77A	644	1bcpF	674	1bhp_	704	1bq0_
615	1b7bA	645	1bcrA	675	1bhu_	705	1bq6A
616	1b7fA	646	1bcrB	676	1bi6H	706	1bqhG
617	1b7tZ	647	1bd3A	677	1bix_	707	1bqpB
618	1b8aA	648	1bd7A	678	1bjaA	708	1bqqT
619	1b8cA	649	1bdg_	679	1bjoA	709	1br2A
620	1b8fA	650	1bds_	680	1bjrE	710	1brv_
621	1b8kA	651	1bdyA	681	1bjx_	711	1brwA
622	1b8uA	652	1be3H	682	1bjyA	712	1brz_
623	1b8wA	653	1be3I	683	1bk1_	713	1bs0A
624	1b9gA	654	1beg_	684	1bk7A	714	1bs9_
625	1b9lA	655	1bei_	685	1bkb_	715	1bshA
626	1b9pA	656	1bev1	686	1bkjA	716	1bteA
627	1b9uA	657	1bf0_	687	1bku_	717	1btn_
628	1b9wA	658	1bf2_	688	1bl1_	718	1bu6O
629	1b9xB	659	1bf5A	689	1bl8A	719	1bu8A
630	1ba0_	660	1bf9_	690	1bm4A	720	1buiC

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
721	1bunB	751	1c1yB	781	1c9uA	811	1ci4A
722	1bus_	752	1c2aA	782	1cb2A	812	1ci6A
723	1buyA	753	1c39A	783	1cb7B	813	1ci6B
724	1bvnP	754	1c3aA	784	1cb9A	814	1cid_
725	1bvnT	755	1c3d_	785	1cbg_	815	1cixA
726	1bw3_	756	1c3gA	786	1cbkA	816	1ckeA
727	1bw5_	757	1c3qA	787	1cc7A	817	1cklA
728	1bw9A	758	1c3yA	788	1ccm_	818	1ckv_
729	1bwvA	759	1c44A	789	1ccvA	819	1cl1A
730	1bwzA	760	1c49A	790	1cd1A	820	1cl7I
731	1bx4A	761	1c4eA	791	1cd3B	821	1cliA
732	1bx7_	762	1c4oA	792	1cd3F	822	1clqA
733	1bxεA	763	1c4rG	793	1cd5A	823	1clvI
734	1bxiA	764	1c4tA	794	1cd9B	824	1cm9A
735	1bxyA	765	1c55A	795	1cdh_	825	1cmr_
736	1by2_	766	1c5a_	796	1cdq_	826	1cmxA
737	1by3A	767	1c5eA	797	1cdzA	827	1cnsA
738	1bykA	768	1c5kA	798	1ce3A	828	1cnt1
739	1byqA	769	1c6wA	799	1ce4A	829	1cnv_
740	1bywA	770	1c7sA	800	1ceeB	830	1cof_
741	1bzg_	771	1c7yA	801	1cf2O	831	1coi_
742	1c01A	772	1c82A	802	1cf3A	832	1cokA
743	1c02A	773	1c8bA	803	1cf4B	833	1coo_
744	1c07A	774	1c8dA	804	1cf7A	834	1cozA
745	1c0eA	775	1c8nA	805	1cfb_	835	1cq3A
746	1c0fA	776	1c8zA	806	1cfe_	836	1cqiA
747	1c0fS	777	1c9fA	807	1cfh_	837	1cqiB
748	1c17M	778	1c9kA	808	1cgiI	838	1cqkA
749	1c1hA	779	1c9oA	809	1chl_	839	1cr0A
750	1c1kA	780	1c9sV	810	1chpD	840	1cs3A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
841	1csbA	871	1d2pA	901	1dcfA	931	1dioB
842	1csgA	872	1d2sA	902	1dcjA	932	1dioG
843	1csmA	873	1d2tA	903	1ddlA	933	1dipA
844	1cv8_	874	1d2zC	904	1dd9A	934	1div_
845	1cviA	875	1d2zD	905	1ddf_	935	1dizA
846	1cvwL	876	1d3bE	906	1ddgA	936	1dj0A
847	1cw3A	877	1d3bL	907	1ddt_	937	1dj7B
848	1cw5A	878	1d3yA	908	1debA	938	1dj1A
849	1cx1A	879	1d4aA	909	1dec_	939	1dk1A
850	1cx8A	880	1d4bA	910	1dekA	940	1dkcA
851	1cxwA	881	1d5tA	911	1deoA	941	1dkiA
852	1cxzB	882	1d5wA	912	1devB	942	1dklA
853	1cy1A	883	1d5yA	913	1df3A	943	1dkrA
854	1cz1A	884	1d6bA	914	1df9C	944	1dl0A
855	1cz3A	885	1d6gA	915	1dfcA	945	1dl2A
856	1cz4A	886	1d71A	916	1dfnA	946	1dlgA
857	1d02B	887	1d7qA	917	1dfuP	947	1dliA
858	1d0dA	888	1d8cA	918	1dfvA	948	1dlvA
859	1d0gR	889	1d8dA	919	1dgnA	949	1dm0A
860	1d0iA	890	1d8hA	920	1dgpA	950	1dm9A
861	1d0kA	891	1d8jA	921	1dgrA	951	1dmlB
862	1d0rA	892	1d9kC	922	1dgrM	952	1dmlE
863	1d0sA	893	1d9uA	923	1dgrN	953	1dmuA
864	1d1hA	894	1dabA	924	1dh3A	954	1dn1A
865	1d11A	895	1danT	925	1dhn_	955	1dnlA
866	1d1nA	896	1danU	926	1dhs_	956	1dnv_
867	1d2iA	897	1dapA	927	1di2A	957	1dnyA
868	1d2jA	898	1dbfA	928	1di6A	958	1do8A
869	1d2lA	899	1dbrA	929	1dil_	959	1doaB
870	1d2nA	900	1dc1B	930	1dioA	960	1dofA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
961	ldowA	991	ldxmA	1021	le4qA	1051	leciA
962	ldowB	992	ldylA	1022	le4sA	1052	leciB
963	ldp5B	993	ldynA	1023	le4tA	1053	lecmB
964	ldp7P	994	ldyoA	1024	le52A	1054	lecrA
965	ldpuA	995	ldytA	1025	le54A	1055	lecsA
966	ldqcA	996	ldz1A	1026	le5bA	1056	led7A
967	ldqeA	997	ldz7A	1027	le5dA	1057	ledg_
968	ldqgA	998	ldzlA	1028	le5jA	1058	lee1A
969	ldqwA	999	ldzoA	1029	le5pA	1059	lee6A
970	ldr9A	1000	ldzrA	1030	le5tA	1060	leejA
971	ldrs_	1001	le0aB	1031	le6cA	1061	leemA
972	ldsyA	1002	le0bA	1032	le6fA	1062	lef1A
973	ldt3A	1003	le0fI	1033	le77A	1063	lef1C
974	ldtdB	1004	le0mA	1034	le79H	1064	lefdN
975	ldtwA	1005	le0vA	1035	le79I	1065	leg3A
976	ldtwB	1006	le12A	1036	le7pA	1066	leg7A
977	ldtyA	1007	le1aA	1037	le8cA	1067	leg9B
978	ldu2A	1008	le20A	1038	le8oA	1068	legaA
979	lduc_	1009	le2aA	1039	le8oB	1069	legH
980	ldviA	1010	le2b_	1040	le8pA	1070	legjA
981	ldvkA	1011	le2hA	1041	le8rA	1071	lego_
982	ldvoA	1012	le2xA	1042	le8tA	1072	leh4A
983	ldw9A	1013	le3gA	1043	le96B	1073	leh5A
984	ldwmA	1014	le3mA	1044	le9kA	1074	lehdA
985	ldwnA	1015	le3uA	1045	le9tA	1075	lehs_
986	ldwpA	1016	le3yA	1046	leanA	1076	lei0A
987	ldx7A	1017	le42A	1047	lebaA	1077	lei1A
988	ldx9A	1018	le4eA	1048	lebfA	1078	lei7A
989	ldxeA	1019	le4gT	1049	lebmA	1079	leigA
990	ldxhA	1020	le4jA	1050	lecbA	1080	leit_

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1081	leiwA	1111	leptA	1141	lewxA	1171	1f2nA
1082	lej1B	1112	leptB	1142	lex1A	1172	1f2uA
1083	lej2A	1113	leq1A	1143	lex2A	1173	1f2uB
1084	lej6D	1114	leq2D	1144	lex7A	1174	1f2vA
1085	lej8A	1115	leq6A	1145	lex9A	1175	1f32A
1086	lejbA	1116	leqfA	1146	lexbA	1176	1f39A
1087	lejfA	1117	leqkA	1147	lexdA	1177	1f3lA
1088	lejiA	1118	lerc_	1148	lexg_	1178	1f3uA
1089	lek8A	1119	lerd_	1149	lexzA	1179	1f3uF
1090	lek9A	1120	lerp_	1150	leyrA	1180	1f42A
1091	lekfA	1121	lery_	1151	leyvA	1181	1f43A
1092	lekrA	1122	lerzA	1152	leyxA	1182	1f47B
1093	lekuA	1123	les3A	1153	lezfB	1183	1f4mA
1094	lekzA	1124	les7B	1154	lezgA	1184	1f51A
1095	lel0A	1125	lesc_	1155	lezjA	1185	1f53A
1096	lel5A	1126	lesnA	1156	lezvB	1186	1f5aA
1097	lel6A	1127	lesxA	1157	lezvG	1187	1f5mA
1098	leljA	1128	let1A	1158	lezvH	1188	1f5qB
1099	lem2A	1129	lethB	1159	lezwA	1189	1f61A
1100	lem8A	1130	leuaA	1160	1f02I	1190	1f6dA
1101	lem8D	1131	leulA	1161	1f08A	1191	1f6vA
1102	lem9A	1132	lev0A	1162	1f09A	1192	1f6yA
1103	lemn_	1133	levsA	1163	1f0kA	1193	1f75A
1104	lemxA	1134	levuA	1164	1f0nA	1194	1f7lA
1105	leni_	1135	lew4A	1165	1f0oA	1195	1f7uA
1106	leo2A	1136	lewfA	1166	1f0zA	1196	1f8rA
1107	leokA	1137	lewiA	1167	1fleA	1197	1f8vA
1108	lep1A	1138	lewnA	1168	1flzA	1198	1f8vD
1109	lep7A	1139	lewsA	1169	1f2hA	1199	1f8xA
1110	lepaA	1140	lewwA	1170	1f2lA	1200	1f9lA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1201	1f93A	1231	1fgp_	1261	1fjsL	1291	1fsb_
1202	1f94A	1232	1fgs_	1262	1fjxA	1292	1fseB
1203	1f98A	1233	1fgxA	1263	1fkmA	1293	1fshA
1204	1f9dA	1234	1fh6A	1264	1fl7B	1294	1fsiA
1205	1f9fA	1235	1fhoA	1265	1flcB	1295	1fskA
1206	1f9nB	1236	1fhvA	1266	1fleI	1296	1fswA
1207	1f9tA	1237	1fi5A	1267	1flgA	1297	1fu5A
1208	1fafA	1238	1fi8C	1268	1flkA	1298	1fuxA
1209	1faoA	1239	1fib_	1269	1fltX	1299	1fvl_
1210	1fapB	1240	1fik_	1270	1fm0D	1300	1fvzA
1211	1fazA	1241	1fit_	1271	1fm0E	1301	1fw2A
1212	1fbmA	1242	1fiuA	1272	1fmhA	1302	1fw9A
1213	1fbr_	1243	1fj1E	1273	1fmhB	1303	1fwkA
1214	1fc3A	1244	1fj2A	1274	1fnnA	1304	1fwnA
1215	1fc5A	1245	1fjgB	1275	1fntF	1305	1fwoA
1216	1fc7A	1246	1fjgC	1276	1fntL	1306	1fwyA
1217	1fcdA	1247	1fjgE	1277	1fntM	1307	1fx2A
1218	1fcjA	1248	1fjgG	1278	1fntN	1308	1fxkB
1219	1fcqA	1249	1fjgH	1279	1fo0A	1309	1fxwA
1220	1fct_	1250	1fjgI	1280	1fo0B	1310	1fxxA
1221	1feaA	1251	1fjgJ	1281	1fo8A	1311	1fy7A
1222	1fexA	1252	1fjgK	1282	1foeA	1312	1fyc_
1223	1fezA	1253	1fjgL	1283	1fokA	1313	1fyvA
1224	1ff3A	1254	1fjgM	1284	1fouA	1314	1fyzC
1225	1ff9A	1255	1fjgP	1285	1fp3A	1315	1fyzE
1226	1ffh_	1256	1fjgQ	1286	1fp8A	1316	1fzpB
1227	1fftC	1257	1fjgS	1287	1fpoA	1317	1fzqA
1228	1ffuB	1258	1fjgT	1288	1fps_	1318	1g0xA
1229	1fg9C	1259	1fjnA	1289	1fpzA	1319	1g0yR
1230	1fggA	1260	1fjrA	1290	1fqjC	1320	1g10A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1321	lg13A	1351	lg6eA	1381	lggpB	1411	lgps_
1322	lg15A	1352	lg6wD	1382	lgh6A	1412	lgpwA
1323	lg1cA	1353	lg6zA	1383	lgh8A	1413	lgpwB
1324	lg1eB	1354	lg72B	1384	lgh9A	1414	lgq8A
1325	lg1pA	1355	lg73A	1385	lghhA	1415	lgq9A
1326	lg26A	1356	lg7dA	1386	lghj_	1416	lgqeA
1327	lg2cA	1357	lg7eA	1387	lghqB	1417	lgqiA
1328	lg2cB	1358	lg7oA	1388	lghsA	1418	lgqpA
1329	lg2hA	1359	lg84A	1389	lgjjA	1419	lgr3A
1330	lg2iA	1360	lg86A	1390	lgjsA	1420	lgrj_
1331	lg2pA	1361	lg8eA	1391	lgk4A	1421	lgrqA
1332	lg2rA	1362	lg8pA	1392	lgkaA	1422	lgs5A
1333	lg2wA	1363	lg8qA	1393	lgl0I	1423	lgsa_
1334	lg31A	1364	lg90A	1394	lgl1I	1424	lgsmA
1335	lg38A	1365	lg99A	1395	lgl2A	1425	lgsoA
1336	lg3gA	1366	lg9mG	1396	lgl2C	1426	lgtdB
1337	lg3jB	1367	lg9pA	1397	lgl2D	1427	lgttA
1338	lg3kA	1368	lga1A	1398	lgl4B	1428	lgu7A
1339	lg3uA	1369	lga3A	1399	lgmeA	1429	lgubA
1340	lg41A	1370	lga7A	1400	lgmiA	1430	lguiA
1341	lg4fA	1371	lgakA	1401	lgmlA	1431	lguuA
1342	lg4pA	1372	lgccA	1402	lgmmA	1432	lgv9A
1343	lg4yB	1373	lgczA	1403	lgngX	1433	lgvmA
1344	lg57A	1374	lgd2E	1404	lgnkA	1434	lgvnA
1345	lg5gA	1375	lgd7A	1405	lgnyA	1435	lgvp_
1346	lg5hA	1376	lgdtA	1406	lgo3F	1436	lgw5A
1347	lg5rA	1377	lgefA	1407	lgo4G	1437	lgw5M
1348	lg5uA	1378	lgff2	1408	lgo5A	1438	lgw5S
1349	lg60A	1379	lgglA	1409	lgo9A	1439	lgwvA
1350	lg61A	1380	lggoA	1410	lgpqA	1440	lgwyA

Sira No	Protein	Sira No	Protein	Sira No	Protein	Sira No	Protein
1441	1gx5A	1471	1h3qA	1501	1hbmB	1531	1hkyA
1442	1gxcA	1472	1h4lD	1502	1hbmC	1532	1hl2A
1443	1gxlA	1473	1h4yA	1503	1hbwA	1533	1hlgA
1444	1gx0A	1474	1h54A	1504	1hbxE	1534	1hllA
1445	1gxqA	1475	1h5oA	1505	1hcc_	1535	1hlvA
1446	1gxrA	1476	1h5pA	1506	1hcd_	1536	1hn3A
1447	1gxtA	1477	1h65A	1507	1hcfX	1537	1hn6A
1448	1gxyA	1478	1h67A	1508	1hcrA	1538	1hnf_
1449	1gy3B	1479	1h6dA	1509	1hd5A	1539	1ho4A
1450	1gyfA	1480	1h6fA	1510	1hd6A	1540	1ho8A
1451	1gyjA	1481	1h6gA	1511	1hdfA	1541	1hq0A
1452	1gywA	1482	1h6qA	1512	1hdlA	1542	1hqmA
1453	1gz5A	1483	1h70A	1513	1hdoA	1543	1hqmE
1454	1gzjA	1484	1h75A	1514	1hekA	1544	1hqz1
1455	1gzsB	1485	1h7cA	1515	1hezE	1545	1hr0W
1456	1h0aA	1486	1h7dA	1516	1hf2B	1546	1hruA
1457	1h0mD	1487	1h7lA	1517	1hfjA	1547	1hs7A
1458	1h0zA	1488	1h7zA	1518	1hgdA	1548	1hssA
1459	1h1dA	1489	1h80A	1519	1hhsA	1549	1htlA
1460	1h1yA	1490	1h8bA	1520	1hic_	1550	1htlC
1461	1h20A	1491	1h8mA	1521	1hiwA	1551	1htrP
1462	1h2dA	1492	1h8pA	1522	1hjrA	1552	1hu3A
1463	1h2iA	1493	1h8vA	1523	1hjsA	1553	1hulA
1464	1h2kS	1494	1h99A	1524	1hjzA	1554	1huvA
1465	1h3fA	1495	1ha6A	1525	1hk9E	1555	1hv6A
1466	1h3iA	1496	1ha8A	1526	1hkeA	1556	1hx1B
1467	1h3lB	1497	1ha9A	1527	1hkfA	1557	1hx2A
1468	1h3mA	1498	1hae_	1528	1hks_	1558	1hx6A
1469	1h3oA	1499	1havA	1529	1hkvA	1559	1hx8A
1470	1h3oB	1500	1hbmA	1530	1hkxA	1560	1hxdA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1561	lhxiA	1591	li3qF	1621	liapA	1651	lijfB
1562	lhy9A	1592	li3qH	1622	liawA	1652	lijqA
1563	lhyhA	1593	li3qK	1623	lib8A	1653	lijxA
1564	lhykA	1594	li3zA	1624	librB	1654	lik9A
1565	lhynQ	1595	li40A	1625	libvF	1655	likoP
1566	lhyp_	1596	li4dA	1626	lic1A	1656	liloA
1567	lhyrA	1597	li4nA	1627	lic8B	1657	lilr1
1568	lhysB	1598	li4wA	1628	lica_	1658	lim3D
1569	lhywA	1599	li5hW	1629	licfB	1659	lim8A
1570	lhz4A	1600	li5kC	1630	licfI	1660	limjA
1571	lhzmA	1601	li5nA	1631	lichA	1661	limt_
1572	li0rB	1602	li5pA	1632	licwA	1662	lin4A
1573	li16_	1603	li5xA	1633	lid0A	1663	linp_
1574	li17A	1604	li6aA	1634	lid1A	1664	lio2A
1575	li19A	1605	li6kA	1635	lidqA	1665	liokA
1576	li1dD	1606	li6wA	1636	lig0B	1666	liomA
1577	li1gA	1607	li78A	1637	lig6A	1667	lip0A
1578	li1jA	1608	li7gA	1638	lignA	1668	lip9A
1579	li24A	1609	li7qA	1639	ligrA	1669	liqoA
1580	li25A	1610	li7wB	1640	liguB	1670	liqrA
1581	li26A	1611	li82A	1641	lihmB	1671	liqtA
1582	li2hA	1612	li85A	1642	lihnA	1672	lirjA
1583	li2mB	1613	li8fF	1643	lihqA	1673	lirl_
1584	li2tA	1614	li8iA	1644	lii2A	1674	lirqB
1585	li2uA	1615	li8nA	1645	liieA	1675	lirzA
1586	li36A	1616	li9bA	1646	liioA	1676	lis7K
1587	li3cA	1617	li9dA	1647	liirA	1677	lisfA
1588	li3oA	1618	li9sA	1648	liiyA	1678	litxA
1589	li3oB	1619	li9zA	1649	lijaA	1679	lityA
1590	li3qE	1620	lia0B	1650	lijdB	1680	liu9A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1681	liuhA	1711	lj1nA	1741	ljb0D	1771	ljfmA
1682	liukA	1712	lj20A	1742	ljb0E	1772	ljfrA
1683	liurA	1713	lj24A	1743	ljb0F	1773	ljfuA
1684	liv8A	1714	lj2jB	1744	ljb0I	1774	ljfxA
1685	livhA	1715	lj2mA	1745	ljb0J	1775	ljfzA
1686	livnA	1716	lj46A	1746	ljb0K	1776	ljgcA
1687	livtA	1717	lj49A	1747	ljb0L	1777	ljh3A
1688	livzA	1718	lj4wA	1748	ljb7A	1778	ljhfA
1689	liw4A	1719	lj53A	1749	ljb7B	1779	ljhgA
1690	liweA	1720	lj5jA	1750	ljbA	1780	ljhnA
1691	liwgA	1721	lj5pA	1751	ljbI	1781	lji7A
1692	liwmA	1722	lj5sA	1752	ljbjA	1782	lji8A
1693	lixdA	1723	lj5uA	1753	ljbkA	1783	ljifA
1694	lixkA	1724	lj5wB	1754	ljc4A	1784	ljiwI
1695	lixlA	1725	lj6oA	1755	ljc7A	1785	ljj2L
1696	lixyA	1726	lj6rA	1756	ljchA	1786	ljjgA
1697	liycA	1727	lj75A	1757	ljctA	1787	ljjhA
1698	liygA	1728	lj7dA	1758	lje5A	1788	ljjO
1699	liykA	1729	lj7lA	1759	lje6A	1789	ljjOE
1700	liyrA	1730	lj7qA	1760	lje8A	1790	ljjUB
1701	liytA	1731	lj7vR	1761	ljeiA	1791	ljjuC
1702	lizlB	1732	lj7xA	1762	ljekA	1792	ljk0B
1703	liziA	1733	lj83A	1763	ljekB	1793	ljkiA
1704	liznA	1734	lj8cA	1764	ljepA	1794	ljkW_
1705	liznB	1735	lj8iA	1765	ljeqA	1795	ljkzA
1706	lj0aA	1736	lj8rA	1766	ljeYB	1796	ljli_
1707	lj0sA	1737	lj9aA	1767	ljf8A	1797	ljmaA
1708	lj0tA	1738	lj9jA	1768	ljfgA	1798	ljmcA
1709	ljldB	1739	ljadA	1769	ljfiA	1799	ljmkC
1710	ljldC	1740	ljaxA	1770	ljfiB	1800	ljmmA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1801	ljmqA	1831	ljud_	1861	1k19A	1891	1k85A
1802	ljmsA	1832	ljunA	1862	1k1gA	1892	1k8hA
1803	ljmtA	1833	ljv1A	1863	1k1vA	1893	1k8kC
1804	ljmvA	1834	ljv2A	1864	1k1wA	1894	1k8kF
1805	ljo6A	1835	ljvr_	1865	1k1zA	1895	1k8kG
1806	ljofA	1836	ljvsA	1866	1k26A	1896	1k8mA
1807	ljogD	1837	ljw2A	1867	1k28D	1897	1k99A
1808	ljovA	1838	ljweA	1868	1k30A	1898	1k9sA
1809	ljoyA	1839	ljwfA	1869	1k36A	1899	1k9yA
1810	ljpeA	1840	ljwhA	1870	1k39A	1900	1ka5A
1811	ljphA	1841	ljwyB	1871	1k3bA	1901	1ka8A
1812	ljpgA	1842	ljx4A	1872	1k3bB	1902	1kafA
1813	ljq3A	1843	ljx6A	1873	1k3bC	1903	1kbaA
1814	ljqoA	1844	ljx7A	1874	1k3eB	1904	1kbhA
1815	ljr3D	1845	ljxcA	1875	1k3kA	1905	1kbzA
1816	ljr5A	1846	ljxhA	1876	1k3sA	1906	1kc6A
1817	ljr8A	1847	ljxsA	1877	1k42A	1907	1kcfA
1818	ljrfA	1848	lji1A	1878	1k47A	1908	1kcxA
1819	ljriA	1849	lji2O	1879	1k4nA	1909	1kczA
1820	ljrmA	1850	lji2P	1880	1k4rA	1910	1kd8B
1821	ljs3A	1851	lji4A	1881	1k4uP	1911	1kdxA
1822	ljs9A	1852	ljiyA	1882	1k4zA	1912	1kezA
1823	ljsbA	1853	ljiyoA	1883	1k5dB	1913	1kf6C
1824	ljsdB	1854	ljiyoE	1884	1k5dC	1914	1kf6D
1825	ljsg_	1855	ljztA	1885	1k5kA	1915	1kg0C
1826	ljxSA	1856	1k04A	1886	1k62B	1916	1kg1A
1827	ljyA	1857	1k0hA	1887	1k6iA	1917	1kgyA
1828	lji0A	1858	1k0rA	1888	1k75A	1918	1khcA
1829	ljtgB	1859	1k0sA	1889	1k77A	1919	1khiA
1830	lju8A	1860	1k12A	1890	1k7yA	1920	1khvA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1921	1kiaA	1951	1kp0A	1981	1kxgA	2011	1l6xB
1922	1kilB	1952	1kp6A	1982	1kxlA	2012	1l7dB
1923	1kilC	1953	1kpgA	1983	1kxoA	2013	1l7kA
1924	1kilD	1954	1kplA	1984	1kxrA	2014	1l7lA
1925	1kiuB	1955	1kppA	1985	1ky6A	2015	1l7vA
1926	1kj6A	1956	1kpsA	1986	1kyhA	2016	1l7vC
1927	1kjkA	1957	1kpsB	1987	1kytA	2017	1l7yA
1928	1kjjB	1958	1kptA	1988	1kywC	2018	1l8jA
1929	1kkeA	1959	1kq1H	1989	1kzhA	2019	1l8oA
1930	1kl7A	1960	1kqcA	1990	1l0oA	2020	1l8pA
1931	1klo_	1961	1kqhA	1991	1l1dA	2021	1l8qA
1932	1klqA	1962	1kqqA	1992	1l1mA	2022	1l8rA
1933	1klxA	1963	1kqrA	1993	1l1nA	2023	1l8yA
1934	1km5A	1964	1krqA	1994	1l1oA	2024	1l9kA
1935	1kmaA	1965	1ks9A	1995	1l1oB	2025	1l9lA
1936	1kmiZ	1966	1ksgB	1996	1l1pA	2026	1l9wA
1937	1kmxA	1967	1kskA	1997	1l1sA	2027	1lajA
1938	1kn6A	1968	1ktbA	1998	1l2mA	2028	1lbs_
1939	1kn7A	1969	1ktgA	1999	1l2qA	2029	1lc0A
1940	1kncA	1970	1ktzB	2000	1l2wI	2030	1lcyA
1941	1knlA	1971	1ku2A	2001	1l3gA	2031	1lddA
1942	1knqA	1972	1ku9A	2002	1l3pA	2032	1ldkA
1943	1knt_	1973	1kuuA	2003	1l3sA	2033	1ldkB
1944	1knvA	1974	1kv8A	2004	1l3yA	2034	1ldkE
1945	1knyA	1975	1kvdA	2005	1l5aA	2035	1ldl_
1946	1knzA	1976	1kvdB	2006	1l6eA	2036	1lfdA
1947	1ko6A	1977	1kvkA	2007	1l6hA	2037	1lfo_
1948	1ko7A	1978	1kvnA	2008	1l6lC	2038	1lg7A
1949	1komA	1979	1kw4A	2009	1l6oA	2039	1lghA
1950	1kozA	1980	1kwaA	2010	1l6wA	2040	1lghB

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
2041	1li1A	2071	1lqtA	2101	1ly1A	2131	1m6nA
2042	1lir_	2072	1lqvC	2102	1lyaA	2132	1m6uA
2043	1lis_	2073	1lr7A	2103	1lyvA	2133	1m6yA
2044	1liuA	2074	1lrrA	2104	1lzwA	2134	1m8oB
2045	1lizA	2075	1lr_	2105	1lzwB	2135	1m8yB
2046	1lj2A	2076	1ls4A	2106	1m0dA	2136	1m93C
2047	1lj8A	2077	1lshA	2107	1m0tA	2137	1m98A
2048	1ljoA	2078	1lslA	2108	1m0vA	2138	1m9gA
2049	1lk9A	2079	1lssA	2109	1m12A	2139	1maeL
2050	1lknA	2080	1lt9A	2110	1m1fA	2140	1mai_
2051	1lktA	2081	1lu4A	2111	1m1gA	2141	1mb6A
2052	1lkvX	2082	1lu9A	2112	1m1sA	2142	1mbmA
2053	1ll8A	2083	1lucB	2113	1m2eA	2143	1mc0A
2054	1lm5B	2084	1lujB	2114	1m32A	2144	1mdaH
2055	1lmmA	2085	1lupA	2115	1m33A	2145	1mdbA
2056	1lmrA	2086	1luzA	2116	1m3jA	2146	1mec1
2057	1ln0A	2087	1lv9A	2117	1m3uA	2147	1mejA
2058	1ln1A	2088	1lvhA	2118	1m44A	2148	1mek_
2059	1ln4A	2089	1lvmA	2119	1m4jA	2149	1mfpA
2060	1lnwA	2090	1lvoA	2120	1m4rA	2150	1mfqC
2061	1lo7A	2091	1lvwA	2121	1m4uA	2151	1mg7B
2062	1lplA	2092	1lw7A	2122	1m4vB	2152	1mg8A
2063	1lq9A	2093	1lwrA	2123	1m4zA	2153	1mhmB
2064	1lqbA	2094	1lwsA	2124	1m56D	2154	1mhnA
2065	1lqbB	2095	1lwuA	2125	1m5hA	2155	1mi1A
2066	1lqbC	2096	1lx8A	2126	1m5iA	2156	1mijA
2067	1lqjC	2097	1lxa_	2127	1m5zA	2157	1miuB
2068	1lqlA	2098	1lxjA	2128	1m62A	2158	1mivA
2069	1lqrA	2099	1lxl_	2129	1m63C	2159	1mj2A
2070	1lqsL	2100	1lxnA	2130	1m6eX	2160	1mjhA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
2161	1mk4A	2191	1mw5A	2221	1n5hA	2251	1neu_
2162	1mkaA	2192	1mwmA	2222	1n69A	2252	1nexC
2163	1mkcA	2193	1mwpA	2223	1n6jA	2253	1nf2A
2164	1mkeA	2194	1mwwA	2224	1n6zA	2254	1nf3C
2165	1mkfA	2195	1mxiA	2225	1n71A	2255	1nfjA
2166	1mkiA	2196	1myn_	2226	1n7eA	2256	1nfkA
2167	1mknA	2197	1mzgA	2227	1n7tA	2257	1ng5A
2168	1mkp_	2198	1mzkA	2228	1n7uA	2258	1ng7A
2169	1ml8A	2199	1n08A	2229	1n7zA	2259	1ngmB
2170	1mla_	2200	1n0lB	2230	1n81A	2260	1ngnA
2171	1mlvB	2201	1n0wA	2231	1n87A	2261	1ngr_
2172	1mm0A	2202	1n0wB	2232	1n8lA	2262	1ngvA
2173	1mm4A	2203	1n10A	2233	1n91A	2263	1nh2C
2174	1mm8A	2204	1n1cA	2234	1n93X	2264	1nh2D
2175	1mmc_	2205	1n1fA	2235	1n9sC	2265	1nh7A
2176	1mn3A	2206	1n1jB	2236	1nar_	2266	1nhyA
2177	1mn8D	2207	1n2dC	2237	1nas_	2267	1ni3A
2178	1mof_	2208	1n2fA	2238	1nbaA	2268	1ni5A
2179	1mp1A	2209	1n3gA	2239	1nbfA	2269	1nigA
2180	1mpxA	2210	1n3jA	2240	1nbwA	2270	1nijA
2181	1mpzA	2211	1n3kA	2241	1nbwB	2271	1nixA
2182	1mr4A	2212	1n4cA	2242	1nc7A	2272	1nj7A
2183	1mswD	2213	1n4mA	2243	1nct_	2273	1nj8A
2184	1mszA	2214	1n4tA	2244	1nd7A	2274	1njhA
2185	1mt0A	2215	1n52A	2245	1nd9A	2275	1njqA
2186	1mt5A	2216	1n52B	2246	1ndbA	2276	1nkl_
2187	1mtpB	2217	1n57A	2247	1ne2A	2277	1nkoA
2188	1mu5A	2218	1n5bA	2248	1ne5A	2278	1nkvC
2189	1mugA	2219	1n5dA	2249	1nepA	2279	1nkzA
2190	1mv3A	2220	1n5gA	2250	1neq_	2280	1nlfA

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
2281	1nlqA	2311	1nvuS	2341	1o59A	2371	1od2A
2282	1nlwA	2312	1nw1A	2342	1o5hA	2372	1odsA
2283	1nlyA	2313	1nw3A	2343	1o65A	2373	1oe4A
2284	1nmwA	2314	1nxiA	2344	1o6aA	2374	1oeyA
2285	1nn4A	2315	1nxuA	2345	1o6uE	2375	1oeyJ
2286	1nnwA	2316	1nxzA	2346	1o70A	2376	1of3A
2287	1no1A	2317	1ny1A	2347	1o7dB	2377	1of5A
2288	1no4A	2318	1ny4A	2348	1o7dD	2378	1ofcX
2289	1nocB	2319	1ny71	2349	1o7dE	2379	1oftA
2290	1nofA	2320	1nyaA	2350	1o7fA	2380	1ofuA
2291	1nogA	2321	1nycA	2351	1o7iA	2381	1ofzA
2292	1nq4A	2322	1nyoA	2352	1o7kA	2382	1ogcA
2293	1nqdA	2323	1nytA	2353	1o8rA	2383	1ogpA
2294	1nqzA	2324	1nzpA	2354	1o8vA	2384	1ogqA
2295	1nr2A	2325	1o0pA	2355	1o9cA	2385	1oh1A
2296	1nr3A	2326	1o0uA	2356	1o9gA	2386	1ohcA
2297	1nriA	2327	1o0yA	2357	1o9lA	2387	1ohhH
2298	1nrjA	2328	1o13A	2358	1o9vA	2388	1ohmA
2299	1nrzA	2329	1o1yA	2359	1oa8D	2389	1ohuB
2300	1ns1A	2330	1o1zA	2360	1obbA	2390	1ois_
2301	1nslA	2331	1o20A	2361	1oc0B	2391	1oj4A
2302	1nstA	2332	1o22A	2362	1occD	2392	1okcA
2303	1nt2A	2333	1o2bA	2363	1occE	2393	1oksA
2304	1nt2B	2334	1o2fB	2364	1occG	2394	1om2A
2305	1ntcA	2335	1o3uA	2365	1occH	2395	1omiB
2306	1ntvA	2336	1o4vA	2366	1occI	2396	1omxA
2307	1nu3B	2337	1o4wA	2367	1occJ	2397	1omyA
2308	1nu7D	2338	1o50A	2368	1occK	2398	1on0A
2309	1nupA	2339	1o54A	2369	1occL	2399	1on4A
2310	1nvmB	2340	1o57A	2370	1occM	2400	1oo0A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
2401	loo0B	2431	loz9A	2461	lpgjA	2491	lpuoB
2402	looeA	2432	lozpA	2462	lpgs_	2492	lpv6A
2403	loofA	2433	lp0zA	2463	lpjzA	2493	lpw4A
2404	lop1A	2434	lp16A	2464	lpkhA	2494	lpwkA
2405	lopd_	2435	lp1cA	2465	lplfA	2495	lpwxA
2406	loq1A	2436	lp32A	2466	lplr_	2496	lpx5B
2407	loqdK	2437	lp35A	2467	lpls_	2497	lpymA
2408	loqeK	2438	lp4eB	2468	lplzA	2498	lpyoD
2409	lor7B	2439	lp4lD	2469	lpm3A	2499	lpytA
2410	lor7C	2440	lp4tA	2470	lpnbA	2500	lpz4A
2411	loruA	2441	lp5uB	2471	lpnbB	2501	lq02A
2412	loryA	2442	lp65A	2472	lpnh_	2502	lq0tA
2413	loscF	2443	lp68A	2473	lpnj_	2503	lq1lA
2414	losmA	2444	lp7bA	2474	lpoc_	2504	lq1oA
2415	lot4A	2445	lp7dA	2475	lpoiB	2505	lq2kA
2416	lotgA	2446	lp7hL	2476	lpov3	2506	lq2yA
2417	lotrA	2447	lp9kA	2477	lpp8M	2507	lq4mB
2418	louoA	2448	lp9nA	2478	lpprM	2508	lq4tA
2419	louwC	2449	lp9oA	2479	lppt_	2509	lq53A
2420	lovqA	2450	lp9vA	2480	lppxA	2510	lq5oA
2421	low1A	2451	lpA7A	2481	lpqrA	2511	lq5vD
2422	low4A	2452	lpavA	2482	lpqwA	2512	lq5xA
2423	lowtA	2453	lpb5A	2483	lpreA	2513	lq5zA
2424	lowwA	2454	lpbuA	2484	lprzA	2514	lq6aA
2425	lowxA	2455	lpbwB	2485	lpsrA	2515	lq6wC
2426	loxjA	2456	lpdgA	2486	lpsuB	2516	lq77A
2427	loxwC	2457	lpdkB	2487	lpswA	2517	lq88B
2428	loy1A	2458	lpdo_	2488	lpt7A	2518	lq8cA
2429	loygA	2459	lpeh_	2489	lpu6B	2519	lq8hA
2430	loyzA	2460	lpfsA	2490	lpugA	2520	lq98A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
2521	lqa4A	2551	lqlmA	2581	lqysA	2611	ltceA
2522	lqabA	2552	lqlwA	2582	lqzgA	2612	ltfr_
2523	lqamA	2553	lqmeA	2583	lr3eA	2613	lthtA
2524	lqaxA	2554	lqmgA	2584	lr4xA	2614	lthx_
2525	lqb3A	2555	lqmiA	2585	lr5rA	2615	ltif_
2526	lqbeB	2556	lqmpA	2586	lr72A	2616	ltiiC
2527	lqcnA	2557	lqnoA	2587	lregX	2617	ltiiD
2528	lqcsA	2558	lqo0E	2588	lrepC	2618	ltit_
2529	lqd1B	2559	lqo7A	2589	lri6A	2619	ltl2A
2530	lqdeA	2560	lqolA	2590	lrmg_	2620	ltocR
2531	lqfhA	2561	lqp6A	2591	lscjB	2621	ltpg_
2532	lqfjA	2562	lqpoA	2592	lsemA	2622	ltsg_
2533	lqftA	2563	lqpsA	2593	lsfp_	2623	ltsk_
2534	lqfxA	2564	lqqeA	2594	lsftA	2624	ltsp_
2535	lqg3A	2565	lqqp1	2595	lsh1_	2625	ltul_
2536	lqgiA	2566	lqqvA	2596	lshsA	2626	ltyfA
2537	lqgkB	2567	lqr0A	2597	lsknP	2627	lu2fA
2538	lqgtA	2568	lqr4A	2598	lskz_	2628	luadC
2539	lqgvA	2569	lqrjB	2599	lsopA	2629	luanA
2540	lqgwA	2570	lqsaA	2600	lsra_	2630	lub1A
2541	lqgwC	2571	lqsdA	2601	lsro_	2631	lub4C
2542	lqhbA	2572	lqsmD	2602	lsscA	2632	lub7A
2543	lqhkaA	2573	lqu0A	2603	lstfl	2633	lucrB
2544	lqhlA	2574	lqu5A	2604	lstmA	2634	lud2A
2545	lqjgA	2575	lqu7A	2605	lsur_	2635	ludnA
2546	lqjtA	2576	lqv9A	2606	lsvfB	2636	lue9A
2547	lqk6A	2577	lqvcA	2607	lsw6A	2637	luenA
2548	lqk7A	2578	lqwtA	2608	lt2rA	2638	lueoA
2549	lqkrA	2579	lqxfA	2609	ltbd_	2639	luewA
2550	lqkyA	2580	lqydA	2610	ltc3C	2640	luf0A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
2641	1uf2B	2671	1wkt_	2701	2ligA
2642	1uf2E	2672	1wq1G	2702	2ncm_
2643	1uf3A	2673	1xat_	2703	2pjrB
2644	1ufvA	2674	1xnaA	2704	2pldA
2645	1ugiA	2675	1yacA	2705	2pth_
2646	1uglA	2676	1ycqA	2706	2sas_
2647	1uhnA	2677	1ycsB	2707	2sicI
2648	1ujlA	2678	1yghA	2708	2spcA
2649	1unnD	2679	1yua_	2709	2sqcA
2650	1uocA	2680	1zaq_	2710	2std_
2651	1uox_	2681	1zug_	2711	2stv_
2652	1uoyA	2682	2aaiA	2712	2thiA
2653	1uqvA	2683	2acy_	2713	2trcP
2654	1us4A	2684	2arcB	2714	2u1a_
2655	1vcc_	2685	2azoB	2715	2u2fA
2656	1vdc_	2686	2bamA	2716	2utgA
2657	1vdrA	2687	2bidA	2717	2vsgA
2658	1vf5B	2688	2btvB	2718	3crd_
2659	1vh0A	2689	2eboA	2719	3ljrA
2660	1vib_	2690	2ezi_	2720	3mspA
2661	1vie_	2691	2ezl_	2721	3ncmA
2662	1vig_	2692	2fivA	2722	3rhn_
2663	1vmoA	2693	2fmr_	2723	3ygsP
2664	1vokA	2694	2fmtA	2724	4otaA
2665	1vpu_	2695	2hgf_	2725	6paxA
2666	1vtx_	2696	2hntC	2726	7ahlA
2667	1vubA	2697	2hpqP	2727	9antA
2668	1wba_	2698	2irfG		
2669	1wdcA	2699	2lbp_		
2670	1wit_	2700	2lefA		

EK 2: Çalışmada kullanılan metal bağlı protein listesi

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
1	101M	31	1B0N_A	61	1C7K_A	91	1DD9_A
2	1A1T_A	32	1B0P_A	62	1C7M_A	92	1DE0_A
3	1A3Z	33	1B13_A	63	1C9Q_A	93	1DFC_A
4	1A42	34	1B25_A	64	1CC5	94	1DFE_A
5	1A4A_A	35	1B3D_A	65	1CFM_A	95	1DFS_A
6	1A4L_A	36	1B4T_A	66	1CFW_A	96	1DFT_A
7	1A56	37	1B4U_B	67	1CFZ_A	97	1DIO_A
8	1A6B_B	38	1B55_A	68	1CHC	98	1DJ7_A
9	1A6F	39	1B57_A	69	1CJX_A	99	1DJN_A
10	1A71_A	40	1B66_A	70	1CL6_A	100	1DK1_A
11	1A73_A	41	1B71_A	71	1CNO_A	101	1DL6_A
12	1A7D	42	1B9M_A	72	1CO4_A	102	1DMT_A
13	1A7I	43	1B9R_A	73	1CVR_A	103	1DO6_A
14	1A7T_A	44	1BBO	74	1CW0_A	104	1DQ3_A
15	1A7V_A	45	1BF6_A	75	1CXY_A	105	1DQS_A
16	1A8E	46	1BH5_A	76	1CY5_A	106	1DRT_A
17	1AF2_A	47	1BIQ_A	77	1D09_B	107	1DW0_A
18	1AH7	48	1BL9_A	78	1D0Q_A	108	1DWL_B
19	1AHJ_A	49	1BLU	79	1D2P_A	109	1DXR_C
20	1AIG_L	50	1BNL_A	80	1D4U_A	110	1DYP_A
21	1AK0	51	1BOR	81	1D6U_A	111	1DZW_P
22	1A00	52	1BS4_A	82	1D7B_A	112	1E08_A
23	1AQ8_A	53	1BT0_A	83	1D8W_A	113	1E0E_A
24	1AQE	54	1BT1_A	84	1D9Y_A	114	1E29_A
25	1AR1_A	55	1BUY_A	85	1DAX	115	1E3D_A
26	1AST	56	1BVB	86	1DBF_A	116	1E4U_A
27	1AW5	57	1BXI_B	87	1DBR_A	117	1E5S_A
28	1AW6	58	1C3R_A	88	1DC1_B	118	1E7D_A
29	1AXQ	59	1C52	89	1DCF_A	119	1E7P_B
30	1B0J_A	60	1C6O_A	90	1DD1_A	120	1EB6_A

Sıra No	Protein	Sıra No	Protein	Sıra No	Protein	Sıra No	Protein
121	1EB7_A	151	1FP0_A	181	1I3Q_A	211	1MM2_A
122	1EC5_A	152	1FP4_A	182	1I3Q_I	212	1MWZ_A
123	1EE8_A	153	1FP4_B	183	1I3Q_L	213	1N0Z_A
124	1EF4_A	154	1FRE	184	1I6N_A	214	1NJ3_A
125	1EFZ_A	155	1FU9_A	185	1I7H_A	215	1NUI_A
126	1EG9_A	156	1FV5_A	186	1IQZ_A	216	1OCC_F
127	1EH6_A	157	1FVS_A	187	1ITQ_A	217	1ODH_A
128	1EHK_B	158	1FWQ_A	188	1IUZ	218	1OPM_A
129	1EI6_A	159	1FY2_A	189	1J2O_A	219	1OX7_A
130	1EIL_A	160	1FYZ_A	190	1J5L_A	220	1PFT
131	1EJE_A	161	1G12_A	191	1JJD_A	221	1PYI_A
132	1EP1_B	162	1G25_A	192	1JM7_A	222	1PZW_A
133	1EQW_A	163	1G5C_A	193	1JM7_B	223	1Q16_A
134	1ETE_A	164	1G71_A	194	1JNR_B	224	1QJK_A
135	1EXK_A	165	1G8F_A	195	1JQ4_A	225	1QYP
136	1EYF_A	166	1G8J_B	196	1JR3_A	226	1R3N_A
137	1F0J_A	167	1G9K_A	197	1JWH_C	227	1RMD
138	1F22_A	168	1GAQ_B	198	1K0V_A	228	1UI0_A
139	1F37_A	169	1GKS	199	1K2F_A	229	1XER
140	1F4S_P	170	1GM4_A	200	1KEA_A	230	1ZFO
141	1F56_A	171	1GMW_A	201	1KLR_A	231	2AW0
142	1F81_A	172	1GOF	202	1KQK_A	232	2HRV_A
143	1FCD_C	173	1GP5_A	203	1KWG_A	233	3CAO_A
144	1FEU_A	174	1GQG_A	204	1LPV_A		
145	1FFU_A	175	1H0H_B	205	1LTL_A		
146	1FJ3_A	176	1H21_A	206	1LV3_A		
147	1FJG_D	177	1H31_B	207	1M1P_A		
148	1FJG_N	178	1HC1	208	1M2O_A		
149	1FN9_A	179	1HYI_A	209	1MG2_D		
150	1FNO_A	180	1I1I_P	210	1MJG_M		