BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

METALOPROTEİNLERİN BİYOENFORMATİK ANALİZİ

SERKAN REMZİ KÜÇÜKBAY

YÜKSEK LİSANS TEZİ
2015

METALOPROTEİNLERİN BİYOENFORMATİK ANALİZİ

BIOINFORMATIC ANALYSIS OF METALLOPROTEINS

SERKAN REMZİ KÜÇÜKBAY

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.
2015

“Metaloproteinlerin Biyoenformatik Analizi” başlıklı bu çalışma, jürimiz tarafından 04/08/2015 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI 'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan : Yrd. Doç. Dr. Yakup ÖZKAZANÇ

Üye (Danışman) : Doç. Dr. Hasan OĞUL

Üye : Yrd. Doç. Dr. Emre SÜMER

ONAY

..../08/2015

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Öncelikle tez çalışmam süresince bilgi ve tecrübesi ile kendimi geliştirmemi sağlayan Danışman Hocam Doç.Dr Hasan Oğul'a,
Yardıma ihtiyacım olduğu her anda desteğini benden esirgemeyen meslektaşım ve Sevgili Eşim Selver Ezgi Küçükbay'a,
Motivasyona ihtiyaç duyduğum her an yanımda olan müstakbel meslektaşım Sevgili Kardeşim Furkan Küçükbay'a, Sevgili Annem Doç.Dr. F. Zehra Küçükbay'a ve Sevgili Babam Prof. Dr. Hasan Küçükbay'a

TEŞEKKÜR EDERİM.

ÖZ

METALOPROTEİNLERİN BİYOENFORMATİK ANALİZİ
Serkan Remzi KÜÇÜKBAY
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Protein içerisinde bulunan metal iyonları proteinlerin fonksiyonel görevlerini yerine getirebilmesi, yapısı ve kararlılığı için önem arz etmektedir. Bu sebeple, proteinler üzerinde metal ile bağlanma noktalarının yüksek performans ile tespiti çok önemlidir. Bu çalışma ile Sistein ve Histidin aminoasitlerinin protein dizilimlerini üzerinden metal ile bağlanma durumunu tahminleyen bir çalışma sunulmaktadır. Dört ayrı yöntem belirtilen amaç doğrultusunda kullanılmıştır. Bunlar; Destek Vektör Makinaları, Naive Bayes, Değişken uzunluklu Markov Zincirleri ve Smith Waterman algoritmasının bir sınıflandırıcı gibi kullanılmasıdır. Yukarda belirtilen bütün metotlar bu sınıflandırmayı sadece protein dizilim bilgisi üzerinden gerçekleştirilmiştir. Farklı birçok öznitelik vektörü oluşturulmuş ve bunların sonuçlara olan etkisi gözlemlenmiştir. Bu çalışma ile metal bağlanma noktaları %35 duyarlılık ve %75 anma değerleriyle Naive Bayes kullanarak, %25 duyarlılık ve %23 anma değerleriyle Destek Vektör Makinaları kullanarak, %0.05 duyarlılık ve %60 anma değerleriyle Değişken uzunluklu Markov zincirleri kullanarak ve çok düşük seçicilik performansı ile Smith Waterman algoritması kullanarak tahminlenebilmiştir. Bu çalışmalar sonrasında, seçilen öznitelik vektörlerinin sonuçlara önemli etkiler yaptığı gözlemlenmiştir. Aynı zamanda elde edilen sonuçlar Naive Bayes yönteminin bu alanda rekabetçi sonuçlar ürettiğini göstermiştir.

ANAHTAR SÖZCÜKLER: Metal bağlanma noktası tespiti, protein katmansal yapısı, Destek Vektör Makinaları, Naive Bayes, Değişken uzunluklu Markov zincirleri, Smith Waterman algoritması

Danışman: Doç.Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

BIOINFORMATIC ANALYSIS OF METALLOPROTEINS

Serkan Remzi KÜÇÜKBAY

Başkent University Institute of Science

Department of Computer Engineering

Metal ions in protein are critical to the function, structure and stability of protein. For this reason, accurate prediction of metal binding sites in protein is very important. Here, we present our study which is performed for predicting metal binding sites for histidines (HIS) and cysteines from protein sequence. Four different methods are applied for this task: Support Vector Machine (SVM), Naive Bayes, Variable-length Markov chain and Smith Waterman Algorithm. All these methods use only sequence information to classify a residue as metal binding or not. Several feature sets are employed to evaluate impact on prediction results. We predict metal binding sites for mentioned amino acids at 35% precision and 75% recall with Naive Bayes, at 25% precision and 23% recall with Support Vector Machine and at 0.05% precision and 60% recall with Variable-length Markov chain, at very low performance with Smith Waterman Algorithm. We observe significant differences in performance depending on the selected feature set. The results show that Naive Bayes is competitive for metal binding site detection.

KEYWORDS: Metal binding site detection, protein conformation, predicting metal binding, SVM, Naive Bayes, Variable-length Markov chain.

Supervisor: Assoc. Prof. Dr. Hasan OĞUL, Başkent University, Department of Computer Engineering.

İÇİNDEKİLER LİSTESİ

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

KISALTMALAR

PST	Olasılıksal Suffix Ağaç yapısı (Probabilistic Suffix Tree)
SVM	Destek Vektör Makineleri (Support Vector Machine)
NB	Naïve Bayes
R	Bağıl Pozisyon Değeri
H	Histidin
C	Sistein

1. GİRİŞ

1.1 Problem Tanımı

1960'larda başlayan bilgisayar uygulamalarının biyolojide kullanılması girişimi, her iki alandaki teknolojik gelişime paralel olarak hızla ilerlemiş ve böylelikle ortaya çıkan biyoenformatik dalı bugün en popüler akademik ve endüstriyel sektörlerinin birisi durumuna gelmiştir.

Bilgisayarların moleküler biyolojide kullanımı üç boyutlu moleküler yapıların grafik temsili, moleküler dizilimler ve üç boyutlu moleküler yapı veri tabanları oluşturulması ile başlamıştır. Kısa sürede çok yüksek miktarlarda veri üreten, endüstri düzeyinde gen ekspresyonu, protein-protein ilişkisi, biyolojik olarak aktif molekül araştırmaları, bakteri, maya, hayvan ve insan genom projeleri gibi biyolojik deneylerin doğurduğu talep sonucunda, bu alandaki bilişim uygulamaları neredeyse takip edilemez bir hızda gelişmiştir. Biyoenformatik alanında yapılan çalışmaların önemli amaçlarından biri canlılığın yapı taşı olan proteinlerin incelenmesi ve bunlar hakkında değerli bilgiler üretmektir. Biyolojik aktivitelerin devamlılığında proteinin rolü çok büyüktür. Hücrelerin görevlerini eksiksiz bir şekilde yerine getirmesi proteinler sayesinde olmaktadır. Proteinlerin fonksiyonel görevlerinin belirlenmesi bu sebeple çok önemlidir. Hücrelerde görevlerini yerine getirmeme durumlarında hangi proteinin buna sebep olduğunun saptanması, çözüme yönelik en önemli adımdır.

Proteinlerin fonksiyonel görevlerinin yerine getirmesi onların 3 boyutlu katmanlı yapısının bozulmadan kalması sayesinde olmaktadır. Bu katmanlı yapı ise proteinlerin çevresindeki metallerle yapmış olduğu kuvvetli bağlar sayesinde olmaktadır. Eğer bu bağ herhangi bir çevresel faktör ile bozulmaya uğrarsa, ilgili protein üstlenmiş olduğu yaşamsal fonksiyonları yerine getiremeyebilir. Çevresindeki metaller ile bağ yapmış şekilde bulunan proteinlere metaloprotein denir. Metal atomları, proteinlerin 3 katmanlı yapısının belirlenmesinde ve düzenleyici ya da katalizör olarak katıldığı görevlerde önemli rol oynamaktadır. Metaller aynı zamanda bir çok biyolojik sürece de dahil olurlar. Bunlara örnek olarak, solunum ve fotosentezde kataliz görevini yerine getiren enzimler verilebilir. Bununla birlikte, hali hazırda tam olarak bir tedavi yöntemi bulunamayan Parkinson, Alzheimer ve AIDS gibi hastalıkların metal birikimlerinden

kaynaklandığını gösteren çalışmalar yapılmaktadır. Aynı zamanda programlanmamış hücre bölünmelerinde de bu metal birikimlerinin etkileri gözlenmektedir. Bu sebeple görevleri tanımlanamamış proteinlerin, yaşamsal fonksiyonel sürece olan katılımlarını tanımlamak onların görevlerini belirlemek birçok hastalığın sebebini anlamamızı kolaylaştırmış olur ve çözümlerin üretilmesini hızlandırır. Bu sebeplerden dolayı proteinlerin metal ile bağlantı noktalarının tespiti büyük önem arz etmektedir.

Bu tez çalışması süresince, üzerinde durmuş olduğumuz problem, protein dizilimleri içerisinde bulunan Histidin ve Sistein aminoasitlerinin metal ile bağlanıp bağlanmadığı konusunda tahminleme yapan bir sistem geliştirmek olmuştur. Bu amaçla, sadece dizilim bilgisinden çıkarılan farklı özniteliklerin bu tahminde ne kadar etkili olduğu test edilmiş ve farklı sınıflandırma alt yapıları üzerinde karşılaştırılmıştır.

1.2 Önceki Çalışmalar

Passerini v.d. [9] sistein ve histidin ile bağlantılı metal bağlanmalar arasındaki geçişleri tanımlayabilmek için yapısal çıktı öğrenimine dayalı yeni bir algoritma geliştirmişlerdir. Geliştirilen sistem, açgözlü (greedy) algoritmasını kullanarak daha etkili bir hesaplama sağlamaktadır. Eğitim kümesi doğruluk tahmininde kullanılanlardan farklı SCOP [10] kümelerine ait protein zincirlerinden oluşmaktadır. Bu ayarlara göre, sistemin 56% duyarlılık ve %60 anma oranlarındadır.

Lippi v.d. [11] çalışmalarında protein dizilimlerini girdi olarak alan ve çıktı olarak da sistein ve histedin kalıntıları için metal bağlanma noktalarının tahminlerini yapan MetalDetector isimli bir internet sunucusu geliştirmişlerdir. Sisteinler için sistem aynı zamanda disülfit bağlanma köprülerini de tahmin etmektedir. MetalDetector temelinde, proteinlerdeki histidin aminoasidini 2 şekilde sınıflandırmaktadır: metal bağlanmış veya bağlanmamış. Sisteinleri ise 3 şekilde sınıflandırmaktadır: metal bağlanmış, bağlanmamış veya disülfit köprüsü oluşturmuş. Çapraz doğrulama performanslarının sonuçlarına göre sistem, disülfit bağlanma durumlarını %88.6 duyarlılık ve %85.1 anma değeri ile tahminlemiştir. Sistem metal bağlanan sisteinleri %79.9 duyarlılık ve %76.8 anma, histidinleri ise %60.8 duyarlılık ve %40.7 anma oranlarıyla tespit etmiştir.

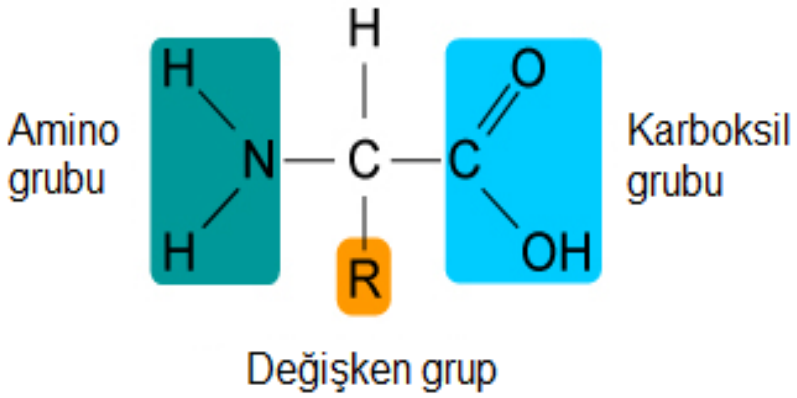
Passerini v.d. [12], [11]'de yapılan uygulamanın geliştirilmiş yeni bir sürümü bulunmaktadır. Yeni sürümde diğerinden farklı olarak, bu uygulamada aynı metal iyonlarında birlikte içeren bağlantı noktalarını tahmin etme özelliği eklenmiştir.

Shu v.d. [13] yaptıkları çalışmada, SVM ile aminoasit dizilimlerinden çinko bağlanma noktalarını tahmin eden bir yöntem önermişlerdir. Önerilen yöntem 2727 adet tekil protein zinciri içeren tekrarsız protein veri bankası kümesi üzerinden elde edilen örneklem kümeleri ile test edildiğinde sistein, histidin, aspartik asit ve glutamik asitleri %75 duyarlılık ve %50 anma değeriyle tespit etmiştir. (Sadece sistin ve histidin için duyarlılık oranı %86'dır.) Bu yöntem için başarım oranı homolog tespit yapıldığında daha fazla çıkmaktadır. sistein, histidin, aspartik asit ve glutamik asitleri için %76 duyarlılık ve %70 anma değeri hesaplanmıştır. (Sadece sistin ve histidin için duyarlılık oranı %90 olmaktadır.)

Passerini v.d. [6], çalışmalarında metal ve demir yapılarındaki çeşitli geçiş metalleri ile bağlanan histidin ve sisteinleri tanıyabilen bir yöntem önermişlerdir. Sistem histidinleri metal bağlanmış veya bağlanmamış şeklinde 2 şekilde tahmin edebilmektedir. Sisteinler ise, metal bağlanmış, bağlanmamış veya disülfit köprüsü oluşturmuş olacak şekilde tahmin edilmiştir. Önerilen yöntem öznitelik vektörleri oluşturulma aşamasında sadece ilgili aminoasit dizilim bilgisini kullanmaktadır. Bunlara ilaveten ilgili aminoasidin protein dizimlerine oranlanmış bağıl pozisyon bilgisi ve daha genel tanımlayıcı bilgiler kullanılmıştır. Çalışmalarında 2 aşamalı makine öğrenme yaklaşımı kullanmışlardır. İlk aşamada, tekil histidinlerin bağlanma durumlarını sınıflandıran destek vektör makineleri eğitilmiştir. İkinci aşamada ise, çift yönlü yinelenen sinir ağları kullanılmıştır. Çalışmanın sonucunda yöntem, histidin ve sisteinlerin metal bağlanmalarını %73 duyarlılık ve %61 anma ile tahmin etmiştir. Sisteinlerin disülfit köprülerinde bulunması ise %86 duyarlılık ve %87 anma ile tespit edilmiştir.

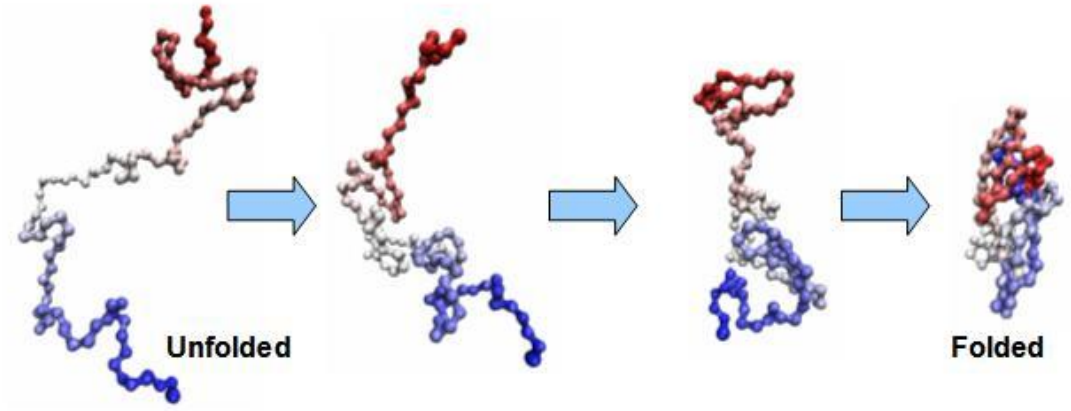
1.3 Alan Bilgisi

Aminoasitler, en basit tanım ile proteinleri oluşturan temel yapı taşlarıdır. Aslında bu yapı taşları bünyesinde amin, karboksil ve fonksiyonel grupları içeren bir moleküldür. 20 çeşit aminoasit bulunmaktadır. Aminoasitler Şekil 1.1'deki gibi fonksiyonel gruplarının özelliklerine göre birbirinden ayrılırlar.



Şekil 1.1 Aminoasit Yapısı

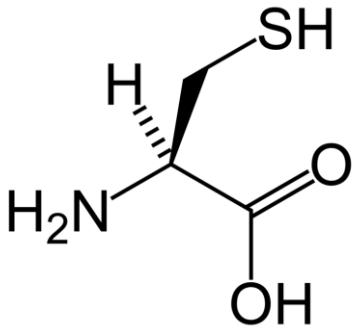
Protein, aminoasitlerin zincir halinde birbirlerine bağlanması sonucu oluşan büyük organik bileşiklerdir. Biyolojik yaşamsal aktivitelerde önemli rolleri bulunmaktadır. Fotosentezden solunuma yaşamın devamlılığı için gereken önemli süreçlerin devamlılığını sağlarlar.



Şekil 1.2 Proteinlerin Katlamalı Yapısı

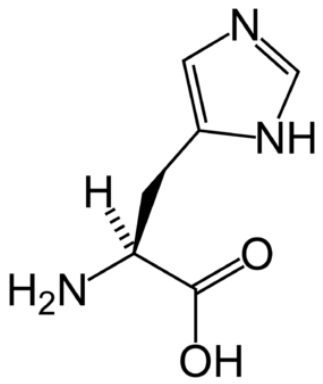
Metaloprotein, yapısında metal içeren proteinlerdir. Proteinlerin büyük bir kısmı bu gruba dâhildir. Proteinlerin yarısından fazlasının metal içerdiği tahmin edilmektedir. Diğer tahmin ise proteinlerin %30 a yakını görevlerini yerine getirmeleri için metale ihtiyaç duymaktadır. Nitekim metaloproteinler hücreler içerisinde birçok görevi üstlenmektedir. Bunlara örnek vermek gerekirse, proteinlerin depolanması ve taşınması işlemleri, enzim olarak hayati fonksiyonlarda görev almaları ve genetik bilgileri aktarma işlemlerini söyleyebiliriz.

Sistein (C/CYS), proteinleri oluşturan 20 aminoasitten biridir. Yan zincirinde kükürt grubu içerir. Polar özelliktedir, ancak fizyolojik pH'da yüksüzdür.



Şekil 1.3 Sistein Kimyasal Yapısı

Histidin (HIS/H), proteinleri oluşturan doğada yaygın olarak bulunan 20 aminoasitten biridir. L-Histidin ve D-Histidin olmak üzere iki farklı formu vardır.



Şekil 1.4 Histidin Kimyasal Yapısı

Ligand, bir biyomoleküle bağlanarak bir karmaşık oluşturan bir bileşiktir. Genelde, iyonik bağlar, hidrojen bağları veya Vander Waals güçleri ile hedef bir proteindeki bağlanma yerine bağlanır.

1.4 Tezin Katkıları

Proteinlerin metal ile bağlanma noktalarının tahminlenmesi konusunda birçok çalışma olmasına rağmen, makine öğrenim tekniklerinin bu araştırma konusunda kullanılması fikri yeni sayılır. Metal ile bağlanma noktalarının tahminlenmesinde, hesaplamaya dayalı olmayan tekniklerin kullanılması birçok dezavantajı beraberinde getirmektedir. Örneğin, X-ray emilim spektroskopi ile bağlanma noktalarının tespitinde kullanılması uygun olduğu görülmüş olmasına rağmen, özel ligandların bağ noktalarını kaçırdığı gözlemlenmiştir. Başka çalışmalarda ise, motif

tabanlı karşılaştırma yöntemi kullanılmıştır. Karşılaştırmalarda düzenli ifadeler ile eşleşen kısımlardan sonuç üreterek bağlanma noktalarının tespiti için tahminler üretmiştir. Bu çalışma da bağlanma noktasının tespitinde kullanılabilir olmasına rağmen düzenli ifadelerin anma belirtiyor olmasından dolayı yüksek oranda yanlış-ret sonucu üretilmiştir. Hesaplamaya dayalı olmayan yöntemlerin sahip olmuş olduğu dezavantajlardan yola çıkarak, bu çalışmada hesaplama tabanlı yöntemler kullanılmıştır. Yapılan çalışma ile makine öğrenim teknikleri kullanılarak metal ile bağlanma noktalarının tespiti yapılabileceği gösterilmiştir.

Tez çalışma süresince, birbirinden farklı hesaplama tabanlı yöntemler kullanılmış, elde edilen sonuçlar karşılaştırılmıştır ve belirleyici sonuçlar elde edilmiştir. Örneğin, bu alanda yapılan çalışmalarda, sıkça kullanılan destek vektör makineleri ile daha az kullanılan Naive Bayes yönteminin sonuçları karşılaştırılmıştır. Elde edilen sonuçlar sonrasında Naive Bayes yönteminin de bu alanda kullanılmaya uygun ve rekabetçi olduğu gözlemlenmiştir.

Bununla beraber çalışma süresince bir birinden farklı öznitelikler çıkarılmış ve elde edilen sonuçlar bir biri ile karşılaştırılarak sonucu pozitif yönde etkileyen öznitelikler belirlenmiştir. Çıkartılan bu özniteliklerin bu alanda yapılan çalışmalar için değerli bir etkisi olacağına inanılmaktadır.

Ayrıca yenilikçi olarak tanımlanabilecek yöntemler kullanılarak sonuçlar üretilmiş ve bu sonuçlar karşılaştırılarak sunulmuştur. Kullanılan bu yenilikçi yöntemlerinde ilgili araştırma alanında pozitif bir etkisi olacağına inanılmaktadır.

1.5 Tezin Organizasyonu

Bölüm 2'de kullanılan yöntemler hakkında bilgiler verilmiştir. Bölüm 3'te elde edilen sonuçlar, bu sonuçları yorumlamamıza yardımcı olacak değerlendirme yöntemleri hakkında bilgiler ve kullanılan veri kümesi hakkında detaylı bilgiler verilmiştir. Son bölümde ise değerlendirmeler ve gelecek çalışma planları aktarılmıştır.

2. KULLANILAN YÖNTEMLER

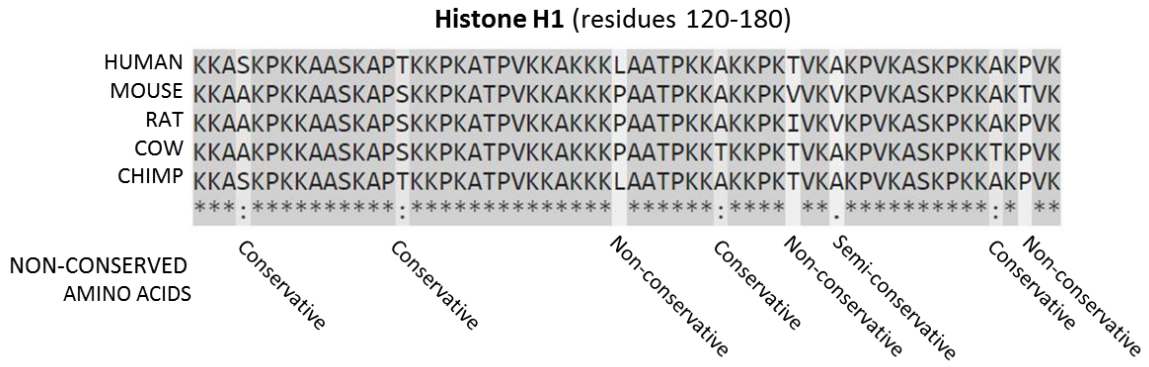
2.1 Metal Bağlanma Tahmini

Metal bağlanma noktalarının tahmini problemi bir sınıflandırma problemi olarak ele alınmıştır. Bu amaçla sistem girdisi tam bir protein dizilimi, çıktısı ise bu dizilim içerisindeki her bir H ve C aminoasitlerinin “bağlanan” veya “bağlanmayan” şeklinde ayrımıdır. Bu ayrımı yapabilmek için üç farklı yaklaşım kullanılmıştır: ikili (pairwise), üretici (generative) ve ayırt edici (discriminative) yaklaşımlar.

2.2 İkili (Pairwise) Yaklaşım

İkili yaklaşım yöntemi iki protein dizilimi arasındaki elde edilen uzaklık/yakınlık skoru üzerinden çalışmaktadır. Karşılaştırılan iki dizilim arasındaki uzaklık ne kadar çok ise, iki örneğin bir biri ile aynı özellikleri göstermeyeceği varsayılmıştır. Bu durum aynı atadan gelen iki protein dizilimi içinde aynı varsayılmıştır. Bu doğrultuda metal ile bağ yaptığını bildiğimiz bir dizilim ile durumunu bilmediğimiz bir dizilimi bu yöntem ile kıyasladığımızda, bu uzaklık değerine göre bir yargı üretilmektedir. Eğer bu yakınlık değerleri aşağıda detaylı bir şekilde anlatılacak olan kurallar içerisindeyse, yeni örnek için metal ile bağ yapar sonucu üretilecektir. Tam tersi durumda doğru olarak kabul edilecektir.

Dizilim hizalaması biyoenformatikte, Protein, RNA veya DNA dizilimlerinde benzer bölgelerin tespiti için kullanılmaktadır. Dizilimleri belirli kurallar ile düzenleyerek benzer bölgeleri tespit eder. Benzer bölgeler, dizilimler arasında fonksiyonel, yapısal veya evrimsel bir ilişki olduğu anlamına gelebilir. Hizalanmış nükleoit veya aminoasit dizilimleri genellikle bir matrisin satırı olacak şekilde gösterilir. Aminoasitler arasında benzer bölgeleri yakalamak için boşluklar eklenir. Böylece alt alta gelen iki satır arasında benzer bölgeler alt alta olacak şekilde hizalanmış olur. İki çeşit hizalama yöntemi bulunmaktadır. Bunlar ilgilendikleri bölgelere göre farklılık gösterir. Bunlardan biri genel dizilim hizalama yöntemidir. Bu yöntemde bütün bir aminoasit dizilimi üzerinden benzer bölgeler tespiti yapılmaya çalışılır. Diğer yöntem ise yerel dizilim hizalama yöntemidir. Bu yöntemde, dizilimler arasında Şekil 2.1’deki gibi bölge bölge benzerlikler hizalanmaya çalışılır.



Şekil 2.1 Dizi Hizalama

Bu çalışmada yerel dizilim hizalama yöntemi kullanılmıştır. Hizalama algoritması olarakta Smith-Waterman algoritması kullanılmıştır. Bu algoritma ilk olarak 1981 yılında Temple F. Smith ve Michael S. Waterman tarafından önerilmiştir [1]. Dinamik bir algoritmadır ve yerel hizalamayi en uygun şekilde yapacağını garanti eder. Bu çalışmada Smith-Waterman algoritmasının açık kaynak kodlu bir versiyonu kullanılmıştır. Geliştirilen sistemde, değinilen algoritmanın hizalama yapmak için üretmiş olduğu yakınlık skorlarından yararlanılmıştır. Bu skorların kullanılma mantığı aşağıda açıklanmıştır.

Öncelikle metal ile bağlanma durumunu tahminlemek istemediğimiz hedef aminoasitten ve komşularından oluşan (bu çalışma için bu aminoasitler C ve H dir) belirli bir uzunlukta bir parçayı aminoasit dizilimi üzerinden alıyoruz. Bundan sonraki anlatımlarda bu parça için çerçeve ifadesi kullanılacaktır. Bu çerçeve çıkarım işlemi hem test hem de eğitim verileri üzerinde uygulanıp, her bir hedef aminoasit için çerçeve çıkarılacaktır.

Bütün veriler için istenen uzunlukta çerçeve çıkarım işleminden sonra, test verilerinden bir çerçeve alınır. Bu çerçeve ile eğitim verileri üzerinden çıkartılmış bütün çerçeveler üzerinden bölgesel hizalama skorları alınır. Böylece elimizde, 1 test çerçevesi için, eğitim verilerinden çıkartılmış çerçeve sayısı kadar hizalama skoru bulunmuş olur. Hizalama skorları alındıktan sonra, bu skorlar büyükten küçüğe sıralanır. Bu sıralama sonrasında ilk N tane skor alınır. Burada seçilen N sayısı tek sayı olması gerekmektedir. Çünkü alınan N adet sonuç ile bir seçim algoritması uygulanacaktır. Tek sayı seçilmez ise seçim sonucu eşitlik içerebilir. Bu N tane skorun hangi eğitim verisi üzerinden elde edildiği tespit edilir. Son olarak bu eğitim verilerinin metal ile bağlanan bir aminoasit olup olmadığına

bakılır. Bu N adet sonuç üzerinden hangi tip durum en çok sayıda ise, test edilen aminoasit içinde aynı durum atanır ve ilgili aminoasit için tahminleme gerçekleşmiş olur.

Protein veri bankasında 1bh8_B olarak etiketlenmiş aminoasit dizilimi üzerinden, hedef aminoasitten 5 uzaklıklı test çerçevesi çıkarma örneği şu şekilde gerçekleştirilecektir:

1bh8_B

FSEEQLNRYEMYRRSAFPKAAIKRLIQAAAVKSITGTSVSQNVVIAMSGISKVFVG
EVVE**E*ALDVCEKWGE***MPPLQPKHMREAVRRLKSKGQIP

TÇ1(Test Çerçevesi 1) = {E,A,L,D,V,C,E,K,W,G,E}

Aynı işlem aşağıda gösterilen eğitim verileri için de gerçekleştirilecektir.

1bcp_F

GLPTHLYKNFTVQELALKLKG***KNQEFCLTAFM***SGR***SLVRACLSDAG***DEKDTWF
DTMLGFAISAYALKSRIALTVEDSPYPGTPGDL***LELQICPLNGY***PE

TRÇ1(Eğitim Çerçeve 1) = {“ ”,G,L,P,T,H,L,Y,K,N,F} VE nmbs

TRÇ2 = {K,N,Q,E,F,C,L,T,A,F,M} VE mbs

TRÇ3 = {S,L,V,R,A,C,L,S,D,A,G} VE mbs

TRÇ4 = {L,E,L,Q,I,C,P,L,N,G,Y} VE mbs

TRÇ1 örneğinde görüldüğü üzere, hedef aminoasit, dizilimin başlarında veya sonlarında olması durumunda, belirlenen komşuluk sayısı içerisinde komşuluğu bulunmayabilir. Örneğin bu TRÇ1 de, dizilimin başında bulunan ***H*** hedef aminoasidinin soldan 5. komşusu bulunmamaktadır. Bu tür durumlarda komşuluk olarak boşluk karakteri atanmıştır.

Metot içerisinde, eğitim verilerinden oluşan çerçeveleri bilgisayar üzerinde modellerken, çerçeve karakter dizisi ile beraber ilgili aminoasittin metal ile bağlanma durumunu belirten bir değer de kullanılmıştır. Bu değer ilgili hedef aminoasittin metal ile bağlanıp bağlanmadığını göstermektedir. Bahsedilen bayrak işaretçisi, **mbs** veya **nmbs** olarak tezin anlatımında kullanılacaktır. Bu ifadelerden

mbs, gerçek hayatta metal ile bağlandığını gösterirken, nmbs metal ile bağlanmadığını gösterecektir.

Çerçeve çıkarım işleminden sonra, test çerçevesi ile bütün eğitim çerçeveleri üzerinden yerel hizalama skorları alınacaktır. Yukarda bahsedilen Smith-Waterman algoritması kullanılacaktır. Aşağıda belirtilen skorlar varsayımsaldır.

SW-Skor1(TÇ1, TRÇ1) = 2.25

SW-Skor2(TÇ1, TRÇ2) = 3.53

SW-Skor3(TÇ1, TRÇ3) = 4

SW-Skor4(TÇ1, TRÇ4) = 1.02

Alınan hizalama skorları büyükten küçüğe sıralanacaktır. Bu sıralama sonrasında,

SW-Skor3, SW-Skor2, SW-skor1, SW-skor4 şeklinde sıralama oluşacaktır. Sıralama işleminden sonra final adımı olan, ilk N tane skorun seçilmesi ve bu skorları yaratan eğitim verilerinin metal ile bağlanma durumları üzerinden bir oylama gerçekleştirilecektir. Bu örnek için N sayısını 3 kabul edersek, elimizde SW-Skor3, SW-Skor2, SW-skor1 sonuçları kalacaktır. Bunlar üzerinden oylama yapıldığında 1 adet metal ile bağlanmayan sonuç kalırken, 2 adet metal ile bağlanan sonuç kalacaktır. Bu oylama sonucunda test edilen aminoasidi metal ile bağlanır olarak tahminlemiş olacağız.

2.3 Üretici (Generative) Yaklaşım

Bu çalışmamızda uygulanan üretici yaklaşım için sınıflandırılması istediğimiz her tür için ayrı modeller oluşturulmuştur. Bu yöntemi, ayırt edici yöntemlerden ayıran özelliği ise budur. Ayırt edici yöntemlerde model eğitimi sürecinde bütün örnekler sınıf bilgisinden bağımsız aynı model üretimi için girdi olarak kullanılırken, bu yöntemde her sınıf kendi modelini oluşturmaktadır.

Bu çalışma içerisinde, sınıflandırma işlemi için dört ayrı model oluşturulmuştur. Bu modeller aynı tür aminoasit dizilimlerini girdi olarak alınıp yaratılan PST yapılarıdır. Oluşturulan bu modeller aşağıda detaylı bir şekilde anlatılmaya çalışılmıştır. Yeni bir aminoasit diziliminin sınıflandırılması aşamasında, üretilen dört model içinde bir sonuç elde edilir ve bu sonuçlar üzerinden sınıflandırma işlemi gerçekleştirilir.

Bu çalışma içerisinde, model eğitimi için Markov Zincirleri yaklaşımı kullanılmıştır. Markov zincirleri, bir olayın gerçekleşmesini, o an ki duruma göre olasılıklandırır. Geçmiş bilgilerden yola çıkarak bir olasılıklandırma yapmaz. Markov zincirlerinin bu yaklaşımını, bir canlının beslenme öğünlerini belirleyen bir model oluşturduğunu düşünerek örneklendirirsek aşağıda belirtilen olasılıksal geçişler oluşabilir.

- Günde bir öğün beslenir.
- Eğer bugün peynir yerse, yarın üzüm veya marul yeme olasılığı eşittir.
- Eğer bugün üzüm yerse, yarın üzüm yeme olasılığı 1/10,peynir yeme olasılığı 4/10 ve marul yeme olasılığı 5/10 dur.
- Eğer bugün marul yerse, yarın üzüm yeme olasılığı 4/10 ya da peynir yeme olasılığı 6/10.Ve Marul yeme olasılığı sıfırdır.

Yukardaki olasılıksal geçişlerden de anlaşılacağı üzere, Markov zincirleri bir sonraki olayın gerçekleşmesini sadece bulunduğu ana bağlar. Geçmişe yönelik bir bilgi kullanmaz.

PST yapıları Markov zincirleri üzerinden üretilmiştir. Markov zincirleri, sıralı verileri her bir karakterin sırasını dikkate alarak model oluşturmak için kullanılır. Sıralı veriler üzerinden benzerlik değeri oluşturur. Sıfır-sıralı(zero-order) Markov zinciri, S_1^N sıralısı için benzerlik değerini olasılıksal olarak verir. Bu değer sıralı içerisinde bulunan her bir sembolün olasılıksal değerinin çarpımına eşittir. Denklem 1 de görüldüğü üzere, P(.) olasılıksal değere tekabül ederken, S_j rastgele seçilmiş olan j inci pozisyondaki karakteri sembolize etmektedir.

$$P(S_1^N) = \prod_{j=1}^{N} P(Sj = sj) \quad (1)$$

Bu çalışmada, yüksek sıralamalı (higher order) Markov chain kullanılmıştır. Bu modelleme yönteminde sıralı veri bütün olarak değil, belli alt sıralılar kullanılarak model oluşturulmaktadır. 2 numaralı denklemde matematiksel ifadesi gösterilmiştir.

$$P(S_1^N) = \prod_{j=1}^{N} P\left(S_j = s_j \mid S_{j-L_j}^{j-1} = s_{j-L_j}^{j-1}\right) \quad (2)$$

Değişken uzunluklu Markov zincirinin verimli bir uygulaması olasılıksal sonekleri gösteren ağaç yapısı üzerinden elde edilmiştir (Probabilistic Suffix Tree). Bu

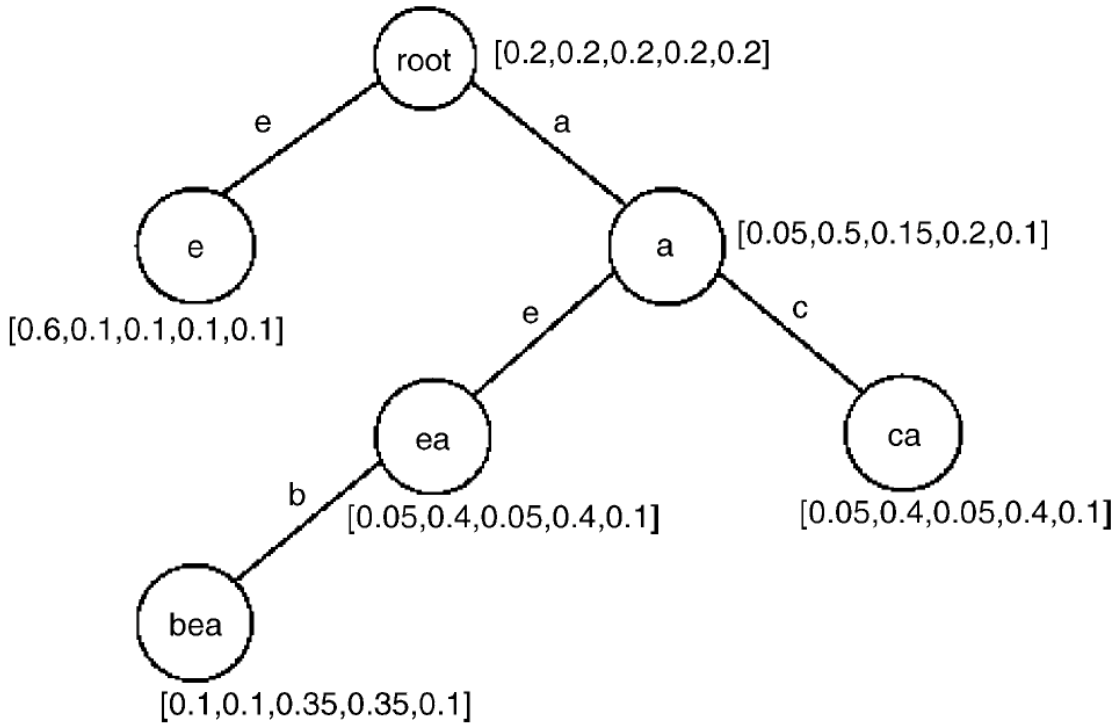
uygulanan yöntem için anlatımı boyunca PST kısaltması kullanılacaktır. PST metodu ilk olarak Bejerano ve Yona tarafından proteinleri modellemek için kullanılmıştır [2].

Orijinal PST mantığı birçok girdi dizilimi üzerinden, önemli olarak nitelendirilebilen parçaların, protein üzerinde bulunduğu yer ile bir bağlantı aramaksızın belirlenmesine dayanır [14].

PST sonlu sayıda karakter üzerinden tanımlanan kenar ve düğümlerden oluşan bir yapıdır. PST içerisinde oluşturulan her bir kenar bir karakter ile ifade edilir. Bir düğüm üzerinden çıkan bir kenara atanmış karakterin aynı düğüm için tekrarına izin verilmez. PST içerisinde bulunan düğümler karakter dizileri ile ifade edilir. Bir PST içerisinde tanımlanmış olan karakter sayısı kadar düğüm bulunabilir. Bu düğümlere ek olarak, deneysel olarak elde edilmiş, tanımlanmış olan karakter sayısı boyutunda olasılık vektörleri eklenir. Bu vektör içerisinde bulunan her bir olasılık değeri, ilgili düğüm sonrasında gelebilecek olan karakterlerin olasılık değerleridir. Şekil 6 da bir PST örneği verilmiştir. Şekil 2.2'de verilen örnekteki oluşuma göre, "bea" ile etiketlenmiş olan düğümün olasılık vektörü (0.1, 0.1, 0.35, 0.35 ve 0.1) dir. Yani, "bea" düğümden sonra "a" karakterinin gelme olasılığı 0.1 iken "c" karakterinin gelme olasılığı 0.35 dir. Bu örnek PST üzerinden, S=abeacad karkater dizisinin bulunma olasılığı 0.2 X 0.5 X 0.2 X 0.6 X 0.35 X 0.2 X 0.4 olacaktır. Buradaki her bir olasılık değeri PST üzerinde dolaşarak elde edilmektedir.

Bu çalışmada metal bağlanma sınıflandırması yapmak için hedef aminoasitler üzerinden çerçeveler alınmıştır. Bu çerçeve 2.1'deki yöntem gibi, merkezinde hedef aminoasidin olduğu yanlarında ise sol ve sağ komşuluklarından oluşan bir alt aminoasitler dizilimidir. Bu yöntemde, çerçeveler eğim verisi üzerinden çıkartılırken ayrı sınıflar olduğu varsayılmıştır. Eğitim verisi üzerinden çıkartılan çerçevelerde, merkez aminoasit olan hedef aminoasidin metal ile bağlanıp bağlanmama durumuna göre sınıf ataması yapılmıştır. Daha sonra bu çerçeveler ortadan ikiye ayrılmıştır. Ve 2 ayrı sınıf daha oluşturulmuştur. Bu sınıflara ayırma işleminden sonra 4 farklı sınıf yaratılmıştır. Bunlar, hedef aminoasidi metal ile bağlanıp, bu hedef aminoasidin solunda kalan örnekler(sınıf1); hedef aminoasidi metal ile bağlanıp, bu hedef aminoasidin sağında kalan örnekler(sınıf2); hedef

aminoasidi metal ile bağ yapmayan, bu hedef aminoasidin sağında kalan örnekler(sınıf3) ve son olarak hedef aminoasidi metal ile bağ yapmayan, bu hedef aminoasidin solunda kalan örnekler(sınıf4)'dür.



Şekil 2.2 [a-e] Karakterleri Üzerinde Tanımlı PST

Sınıf1 ve sınıf2 ayrı sınıflar olmasına rağmen ata sınıfı metal ile bağlananlar olarak belirlenmiştir. Sınıf3 ve sınıf4 ise aynı mantık üzerinden metal ile bağlanmayanlar olarak belirlenmiştir.

1bcp_F

GLPTHLYKNFTVQELALKLKG***KNQEFCLTAFM***SGRSLVRACLSDAGDEKDTWFD
TMLGFAISAYALKSRIALTVEDSPYPGTPGDLLELQICPLNGYPE

Aşağıda 1bcp_F dizilimi üzerinden bahsedilen dört ayrı sınıfın oluşturulma işlemi adım adım anlatılmıştır.

- İlk olarak hedef aminoasit üzerinden bir çerçeve çıkarılacaktır. Ç1 = {***K, N, Q, E, F, C, L, T, A, F, M***}.
- Bir sonraki aşama olarak çıkarılan bu çerçeve hedef aminoasitin (C) bulunduğu noktada ikiye bölünecektir. Ç1L = { **K, N, Q, E, F**} Ç1R = {**M, F, A, T,L**}
- Son aşama olarak, elde edilen bu iki farklı çerçeve daha sonra hedef aminoasitin metal ile bağlanma durumuna göre ayrı sınıflara atanacaktır. Bu örnek için hedef aminoasit metal ile bağ yaptığını varsayarsak, Ç1L sınıf 1 olarak işaretlenirken, Ç1R sınıf 3 olarak işaretlenecektir.

Bu adımlar eğitim verisinde bulunan bütün aminoasitler için gerçekleştirilecektir. Böylece eğitim verisi dörde bölünmüş olup, dört ayrı sınıftan oluşmuş olacaktır. Bu dört ayrı veri kümeleri için ayrı ayrı PST'ler oluşturulacaktır. Bu işlemin sonucunda elimizde dört ayrı PST modeli olmuş olacaktır.

PST modelleri oluşturulduktan sonra, test dataları içerisinde hedef aminoasit üzerinden çerçeve çıkarılma işlemi uygulanacaktır. Çıkartılan bu çerçeve bütün olarak bir önceki aşamada elde edilmiş olan PST modelleri üzerinden test edilip, her bir PST için ayrı skorlar elde edilecektir. Elde edilen bu PST skorları, yukarıda bahsedilen sınıf1 ve sınıf2 kümesine dâhil olanlar kendi içinde çarpılacak, sınıf3 ve sınıf4 kümesine dâhil olan kendi içinde çarpılacaktır. Daha sonra elde edilen bu iki değerden hangisi daha büyük ise test edilen veriye büyük olan skorun sınıfınki ile aynı sınıf ile işaretlendirilecektir.

2.4 Ayırt Edici (Discriminative) Yaklaşım

Ayırt edici yaklaşım, iki veya daha fazla farklı türden oluşan verileri, gözlemler sonucu elde edilmiş bilgileri kullanarak bir birinden ayırt etme mantığına dayanmaktadır. Bu çalışma içerisinde bir türe ait olma veya olmama durumunu ayırt edilmeye çalışılmıştır. Bahsedilen bu tür, metal ile bağ yapan aminoasit ve metal ile bağ oluşturmayan aminoasit olarak kullanılmıştır.

2.4.1 Sınıflandırma yöntemi

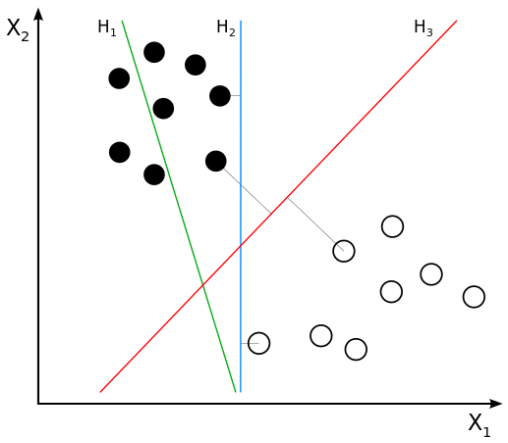
Sınıflandırma, temel anlamda karar verme amacıyla kullanılan bir işlemdir. Bu işlemin amacı, sınıflandırılması için verilen örneklerin arasında ayırıcı modeller

oluşturmak ve yeni gelen örnekleri üretilen ayırt edici modellere göre hangi sınıfa ait olduklarını tayin etmektir. Bu çalışmada sınıflandırma yöntemlerinden destek vektörleri (SVM) ve Naïve Bayes yöntemi (NB) kullanılmıştır.

2.4.1.1 Destek vektör makineleri (SVM)

SVM (Support Vector Machine) kendi kendine ayırt edici bir fonksiyon çıkartabilen öğrenme modelidir. Verilen veriyi analiz ederek bir örüntü çıkartır ve bu örüntü üzerinden ayırt edici olarak kullanılacak olan bir fonksiyon üretir. Üretilen bu model sınıflandırma ve regresyon analizinde kullanılır. SVM' nin üretmiş olduğu model aslında, kendisine verilen her bir örneğin uzayda bir noktaya atanmış bir ifadesi olarak açıklayabiliriz. SVM'nin asıl amacı ise, bu uzaydaki noktaları bir birinden ayırt eden belirgin boşluklar oluşturmaktır. Yani iki farklı sınıfın uzaydaki görüntüsü bir birinden ayırt edilecek kadar uzakta olmalıdır. Böylece SVM bir model oluşturduktan sonra, yeni gelen bir örneği, oluşturmuş olduğu örnekler uzayında nereye konumlandığına bakar ve hangi sınıfın bölgesine düştüyse, o sınıftandır kararını verir. SVM birçok alanda kullanılan ve başarılı sonuçlar elde edilen bir sınıflandırma yöntemidir. Orijinal SVM Vladimir N. Vapnik ve Alexey Ya. Chervonenkis tarafından 1963 yılında keşfedilmiştir. Şu anda kullanılan standart SVM algoritması 1995 yılında dünyaya duyurulmuştur [3].

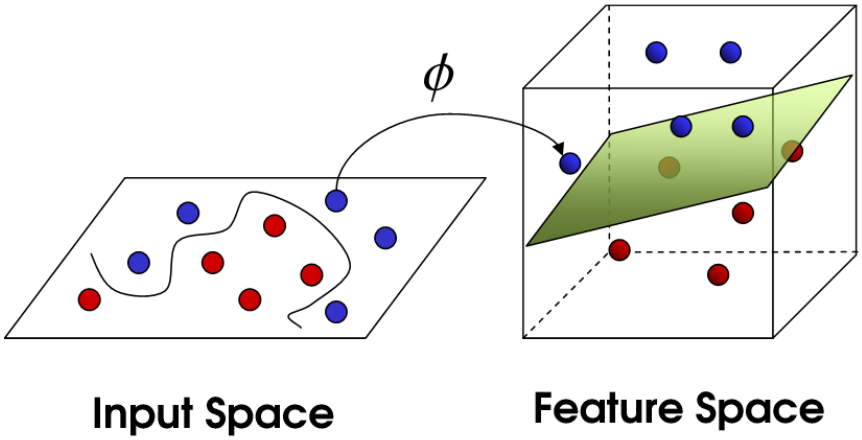
Şekil 2.3'de SVM ye verilen örnekleri tek boyutlu bir uzayda nasıl noktalandırdığı gösterilmektedir. H fonksiyonlarını ise ayırt edici fonksiyonlar olarak söyleyebilir. Bu örnekte en iyi ayırt ediciliği H3 fonksiyonu sağlamıştır.



Şekil 2.3 SVM Tarafından Üretilmiş Ayırt Edici Model[1]

[1] https://en.wikipedia.org/wiki/Support_vector_machine'den alınmıştır.

Şekil 2.4'de ise SVM'nin n boyutlu sistem üzerinde modellemesi gösterilmiştir.



Şekil 2.4 SVM'nin n Boyutlu Sistem Üzerinde Modellenmesi[2]

SVM yöntemi çekirdek fonksiyonunun değişmesi ile farklılaştırılabilir. Doğrusal olarak ayrılabilen iki sınıflı bir sınıflandırma probleminde DVM'nin eğitimi için k sayıda örnekten oluşan eğitim verisinin $\{x_i, y_i\}, i = 1, \dots, k$ olduğu kabul edilirse, optimum hiper düzleme ait eşitsizlikler aşağıdaki şekilde olur.

$$w.xi + b \geq +1 \text{ her } y = +1 \text{ için} \quad (3)$$

$$w.xi + b \leq +1 \text{ her } y = -1 \text{ için} \quad (4)$$

Bu denklemlerde x n boyutlu uzayı, y ise sınıf etiketlerini ve w ise ağırlık vektörünü (hiper düzlemin normali)ifade etmektedir. Optimum hiper-düzlemin belirlenebilmesi

için bu düzleme paralel ve sınırlarını oluşturacak iki hiper-düzlemin belirlenmesi gerekir. Bu hiper-düzlemleri oluşturan noktalar destek vektörleri olarak adlandırılır ve bu düzlemler $w.xi + b = \pm 1$ şeklinde ifade edilirler.

Optimum hiper-düzlemin sınırının maksimuma çıkarılması için w ifadesinin minimum hale getirilmesi gerekir. W ifadesini maksimum değerlerle ifadesi sonucunda SVM için karar denklemi 5'deki gibi ifade edilebilir.

$$f(x) = \text{sign}\left(\sum_{i=1}^{k} \lambda i.\, yi(x.xi)\right) + b \quad (5)$$

Bu çalışmada karar denklemi olarak radial tabanlı olan SVM kullanılmıştır (denklem 6). Çalışma içerisinde her bir hedef aminoasit için öz nitelikler çıkartılmıştır ve bu öz nitelikler metal ile bağlanan için 1 olarak etiketlenirken,

[2] http://boryazilim.com/kernel-vektor-destek-makineleri-svm/ 'den alınmıştır.

metal ile bağlanmayanlar için -1 olarak etiketlenmiştir. SVM için github üzerinde açık kaynak kodlu paylaşılan libSVM paketi kullanılmıştır. (gamma = 0.05, cost = 0.1)

$$\mathrm{k}(\mathrm{x},\mathrm{y}) = \mathrm{e}^{-\gamma||(\mathrm{x}-\mathrm{xi})||^2} \quad (6)$$

2.4.1.2 Naive Bayes sınıflandırıcı

Naïve Bayes teoremi, olasılıksal sınıflandırma yapan sınıflandırma algoritmaları ailesindendir. Uyguladığı yöntem bir durumun meydana gelmesi, diğer durumların meydana gelmesini etkilemez teoreminden oluşmaktadır. Bu yaklaşımı girdi olarak almış olduğu öz niteliklere de uygulayarak sınıflandırma işlemini gerçekleştirir. Naïve Bayes ilk olarak bilgi çıkartımı alanında farklı bir isimde 1960 yılında duyurulmuştur [4]. Ve bu alanda popülerliğini korumaktadır. Bahsi geçen çalışmada kelime tanıma amaçlı kullanılmıştır. Aynı zamanda günümüzde gereksiz zararlı e-posta tanıma sistemlerinde yardımcı fonksiyon olarak kullanılmaktadır.

Naïve Bayes gerçek dünyada birçok kullanımda oldukça başarılı sonuçlar üretmiştir [8].

Bayes teoremi kısaca aşağıdaki formülle ifade edilebilir.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (7)$$

P(A|B); B olayı gerçekleştiği durumda A olayının meydana gelme olasılığıdır. Bu değere koşullu olasılıkta denilmektedir. Bayes teoreminin önemli bir yaklaşımıdır. Koşullu olasılık kavramı, bir olayın gerçekleşme olasılığının hesaplanmasında ek bilginin kullanılmasına olanak tanır. Örneğin bir kişinin iki çocuğu olduğunu düşünürsek, her ikisinin de kız olma olasılığı 1/4 olur. Ancak birinin kız olduğunu önceden bilirsek, bu olasılık 1/3 olarak değişir. Ama herhangi biri değil de birincisi (yaşça büyük olan) kız olduğu biliniyorsa olasılık 1/2 olur. Yani bu iki durumda, her iki çocuğun da kız olma olasılığı, birinin kız olması koşullu olarak hesaplanır.

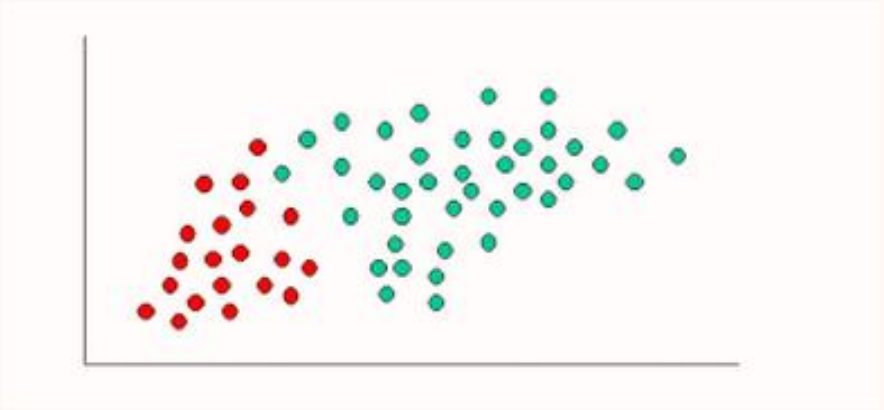
P(B|A); A olayı gerçekleştiğinde B olayının meydana gelme olasılığıdır.

P(A) ve P(B); A ve B olaylarının önsel olasılıklarıdır.

Önsel olasılıkta Naive Bayes teoreminin önemli kavramlarından biridir. Geçmişten gelen, tecrübeye dayalı olasılık oranıdır. Örneğin Şekil 2.5 de görülen dağılım doğrultusunda, yeni gelecek rengi bilinmeyen bir nesnenin yeşil olma önsel olasılığı denklem 8 de, kırmızı olma önsel olasılığı fonksiyon denklem 9 da verilmiştir.

$$P(Y) = \left({yeşil\ örnek\ sayısı}/{toplam\ örnek\ sayısı} \right) \quad (8)$$

$$P(Y) = \left({kırmızı\ örnek\ sayısı}/{toplam\ örnek\ sayısı} \right) \quad (9)$$



Şekil 2.5 Kümelenmiş Veri Örneği

Bayes karar teoreminde istatistik olarak bağımsızlık önermesinden yararlanılırsa bu tip sınıflandırmaya Naive bayes sınıflandırılması denir.

Bayes teoremi sınıflandırma aşamasında, verilen bir sınıf olan y ile [x1-xn] arasında olan örneklem uzayında, denklem 10 da gösterilen ilişkiyi oluşturur.

Bayes teoremi sınıflandırma aşamasında, verilen bir sınıf olan y ile [x1-xn] arasında olan örneklem uzayında, denklem 10 da gösterilen ilişkiyi oluşturur.

$$P\left(y \mid x_{1,\ldots,}x_n\right) = \frac{P(y)P(x_{1,..,}x_n \mid y)}{P\left(x_{1,..,}\, x_n\right)} \quad (10)$$

Denklem 10 ye her bir olayın oluşumunun bir birinden bağımsızlık kuralı uygulanırsa, bu denklem 11 deki gibi ifade edilecektir.

$$P\left(y \mid x_{1,\ldots,}x_n\right) = \frac{P(y)\prod_{i=1}^{n} P(Xi \mid y)}{P\left(x_{1,..,}\, x_n\right)} \quad (11)$$

Naive Bayesin kullanmış olduğu karar denklemi ise denklem 11 den yola çıkarak denklem 12 daki gibi ifade edilmiştir.

$$(y) = argmaxP(y)\prod_{i=1}^{n} P(Xi\,|y) \tag{12}$$

Naive Bayes karar denklemi kullanılan (Çekirdek fonksiyonu) dağılım değeri olan P($X_i \mid y$) değiştirilerek farklılaşabilir. Bu çalışma içerisinde kullanılan öz nitelikler sürekli sayılardan oluştuğu için, Gaussian Naïve Bayes kullanılmaya karar verilmiştir. Uygulanan çekirdek fonksiyonu denklem 4 de gösterilmiştir.

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(v-\mu_c)^2}{2\sigma_c^2}} \tag{13}$$

Naïve Bayes sınıflandırıcı olarak çalışmalar süresince açık kaynaklı olarak paylaşılan MATLAB kütüphaneleri kullanılmıştır.

2.4.1.3 Öznitelikler

Öznitelik makine öğrenim alanında, gözlemlenmiş olayların, ölçülebilen ve ifade edilebilen bağımsız özellik değerleridir. Ayırt edici, bilgi verici ve bağımsız öznitelik seçimi sınıflandırma işlemleri için büyük öneme sahiptir. Öznitelikler genellikle sayısal değerlerden oluşsa da, yapısal öznitelikler de kullanılmaktadır. Bunlara en iyi örnek çizgeler ve dizgiler verilebilir.

Bu bölümde sınıflandırma yöntemleri için kullanılan öznitelikler hakkında bilgiler verilecektir. Bu çalışmada aminoasit dizilimlerinden yola çıkarak sayısal ve sürekli olan öznitelikler elde edilmiştir. Çalışmanın en önemli noktalarından biri de sınıflandırma işlemini sadece dizilim bilgisinden yola çıkarak elde edilmiş olan özniteliklerden gerçekleştirmektedir.

Bu çalışmada aminoasitler üzerinden öznitelik çıkartılmasında birden çok yöntem kullanılmıştır. Ve elde edilen her bir öznitelik için ayrı sınıflandırma skorları alınmıştır. Kullanılan bu yöntemlerden biride bir aminoasidin bir başka aminoasidin yerine geçebilme ihtimalini gösteren değişiklik matrisi veya kabul edilebilir mutasyon noktaları olarak da isimlendirilen (substitution matrix veya point accepted mutation) PAM matris değerleri olmuştur.

PAM doğal seleksiyon açısından proteinlerin birincil yapısında (yani aminoasit dizilimleri şeklinde olan yapısında) doğal seleksiyon süreçleri bakımından kabul

edilebilir derece oluşan değişikliklerdir. Kısacası bir aminoasidin başka bir aminoasit ile değişmesidir. PAM matrisi, kolonları ve satırları 20 aminoasit için sayısal ifade gösteriminden oluşan bir matristir. PAM matrisi biyoenformatik alanında sıkça dizilim hizalamada skorlama aşamasında kullanılmaktadır.

Bu çalışmada PAM matrisi, hedef aminoasidin komşularına olan yakınlık derecesini göstermek amacı ile kullanılmıştır. Bu yakınlık bilgisi bir çerçeve üzerinde gösterilmiştir. Yani hedef aminoasidin her bir komşusu çerçeve üzerinde 20 adet sayısal bilgi ile gösterilmiştir. Biyoenformatik alanında kullanılan farklı farklı PAM matrisleri bulunmaktadır. Bu çalışma için yaygın olarak kullanılan PAM 120 matrisi kullanılmıştır. Şekil 2.6' da PAM 120 matrisinin içeriği gösterilmiştir.

Yukarda bahsedilen hedef aminoasidin komşularına olan benzerlik değerinin bir çerçeve üzerinde gösterimi aşaması aşağıda örnek üzerinde açıklanmıştır. Bir önceki bölümlerde kullanılan protein örneği burada kullanılacaktır.

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

Şekil 2.6 PAM 120 Matris Değerleri

Yukarda bahsedilen hedef aminoasidin komşularına olan benzerlik değerinin bir çerçeve üzerinde gösterimi aşaması aşağıda örnek üzerinde açıklanmıştır. Bir önceki bölümlerde kullanılan protein örneği burada kullanılacaktır.

1bh8_B	FSEEQLNRYEMYRRSAFPKAAIKRLIQAAAVKSITGTSVSQNVVIAMS GISKVFVGEVVE***E<u>ALDVCEKWGE</u>***MPPLQPKHMREAVRRLKSKGQIP

Şekil 2.7 Aminoasit Dizilim Örneği

Şekil 2.7'de koyu harflerle gösterilen hedef aminoasit komşularını PAM matris değerlerini kullanarak çerçeve oluşturma aşaması anlatılacaktır.

- İlk olarak aminoasit üzerinden hedef aminoasidin komşularını içeren N komşuluklu komşuluk çerçevesi çıkartılır.
  Çerçeve1= {E,A,L,D,V,E,K,W,G,E}
- Çerçeve bölgesinin belirlenmesinden sonra, çerçeve içerisindeki her bir aminoasit için PAM 120 matrisinde ki değeri alınır ve ilgili aminoasidin yerine yazılır. Aşağıda çerçevede bulunan her bir aminoasidin PAM 120 değeri listelenmiştir.
  E = (0,-3,1,3,-7,2,5,-1,-1,-3,-4,-1,-3,-7,-2,-1,-2,-8,-5,-3);
  A = (3,-3,-1,0,-3,-1,0,1,-3,-1,-3,-2,-2,-4,1,1,1,-7,-4,0);
  L = (-3,-4;-4,-5,-7,-2,-4,-5,-3,1,5,-4,3,0,-3,-4,-3,-3,-2,1);
  D = (0,-3,2,5,-7,1,3,0,0,-3,-5,-1,-4,-7,-3,0,-1,-8,-5,-3,);
  V = (0,-3,-3,-3,-3,-3,-3,-2,-3,3,1,-4,1,-3,-2,-2,0,-8,-3,5);
  K = (-2, 2,1,-1,-7,0,-1,-3,-2,-3,-4,5,0,-7,-2,-1,-1,-5,-5,-4);
  W = (-7,1,-4,-8,-8,-6,-8,-8,-3,-6,-3,-5,-6,-1,-7,-2,-6,12,-2,-8);
  G = (1,-4,0,0,-4,-3,-1,5,-4,-4,-5,-3,-4,-5,-2,1,-1,-8,-6,-2);
- Komşuluk çerçevesinde bulunan her bir aminoasidin yerine PAM120 matris değeri yazılması sonucunda elde edilen öznitelik sayısı C = (N x 2) x 20 dir. Bu örnekte komşuluk sayısı sağdan ve soldan 5 olacak şekilde ayarlanmıştır. Bu incelenen durumda Şekil 2.8'deki gibi komşuluk çerçevesinin oluşturduğu toplam öznitelik sayısı 200 olmuş olur.

0,-3,1,3,-7,2,5,-1,-1,-3,-4,-1,-3,-7,-2,-1,-2,-8,-5,-3,3,-3,-1,0,-3,-1,0,1,-3,-1,-3,-2,-2,-4,1,1,1,-7,-4,0-3,-4;-4,-5,-7,-2,-4,-5,-3,1,5,-4,3,0,-3,-4,-3,-3,-2,10,-3,2,5,-7,1,3,0,0,-3,-5,-1,-4,-7,-3,0,-1,-8,-5,-3,0,-3,-3,-3,-3,-3,-3,-2,-3,3,1,-4,1,-3,-2,-2,0,-8,-3,5-2,2,1,-1,-7,0,-1,-3,-2,-3,-4,5,0,-7,-2,-1,-1,-5,-5,-4-7,1,-4,-8,-8,-6,-8,-8,-3,-6,-3,-5,-6,-1,-7,-2,-6,12,-2,-8,1,-4,0,0,-4,-3,-1,5,-4,-4,-5,-3,-4,-5,-2,1,-1,-8,-6,-2

Şekil 2.8 Komşuluk Çerçevesinin PAM Matrisi ile Oluşturulmuş Örneği

Bu çalışma içerisinde sınıflandırma yöntemleri için kullanılan diğer bir öznitelik ise hedef aminoasidin dizilimi üzerinde bulunduğu pozisyonun orantısal değeridir. Bu orantısal değer, hedef aminoasidin bulunduğu pozisyonun, aminoasit diziliminin uzunluğuna bölünmesi ile elde edilir. Denklem 14'te bu değer RP ile gösterilmiştir ve bundan sonraki anlatımlarda RP olarak bahsedilecektir.

$$RP = Pozisyon\ Indeks/Sekans\ Uzunluğu \tag{14}$$

Böylece PAM120 matris değerleriyle oluşturulmuş olan özniteliklere RP değerinin eklenmesiyle toplam öznitelik sayımız bu devam eden örneğimiz için 201 olmuştur.

Yukarda bahsedilen özniteliklere ek olarak PAM120 matrisi dönüşümü yapılmadan önce elde edilen çerçeve üzerinden Protein Composition Server[5] ile ilgili çerçeve ve dolayısıyla hedef aminoasit hakkında karakteristik bilgiler elde edilmiştir. Bu bilgilerde öznitelik olarak daha önceki aşamalarda oluşturulan öznitelik vektörüne eklenmiştir. Bahsi geçen sunucu ile 4 farklı aminoasit karakteristik bilgisi elde edilmiştir. Bunlar sözde aminoasit kompozisyonu(PAAC), amfifil sözde aminoasit kompozisyonu(APAAC), beş çarpan skorlu aminoasit kompozisyonu(5FSS) ve fizikokimyasal özellikleridir(PC).

Bu karakteristik bilgiler ayrı ayrı öznitelik vektörüne eklenmiştir. Bunun sebebi bu dört bilginin de aynı farklılıkları ortaya çıkartacak özelliklere sahip olmasıdır. Yani sınıflandırma metotlarında verilen öznitelik vektörü yukarda bahsedilen dört karakteristik özniteliklerden sadece birini içerecektir.

Bu çalışmada yukarda bahsedilen dört öznitelik bilgisi ayrı ayrı sınıflandırma metotlarına verilmiş ve Şekil 2.9'deki gibi hepsi için ayrı sonuçlar elde edilmiştir.

1	PC	81010,7,42.8571428571429,14.2857142857143,28.5714285714286,28.5714285714286,57.1428571428571,14.2857142857143,28.5714285714286,14.2857142857143,14.2857142857143,71.4285714285714,14.2857142857143
2	PAAC	0,0.595876898920262,0,0,0,0.595876898920262,0,0.595876898920262,0.595876898920262,0,0,0.595876898920262,0,0,0,0.595876898920262,0,0,0,0.595876898920262
3	APAAC	0,0,0,0,0,-4.51943980707188,0,0,-4.51943980707188,0,-9.03887961414377,-4.51943980707188,0,0,0,-4.51943980707188,0,0,0,-4.51943980707188,3.17836642410698,3.66681714894065,-2.2188598448188,-7.43774066423208,-4.93079936152248,4.11602938640677,7.06303607740982,-2.02920857569027,26.4863848601219,22.8909036522514,26.9861326107747,18.558242423971,13.0666173110258,22.8909036522514,-1.19938364778212,0.659332954648802,13.0666173110258,22.8909036522514,-9.99486297399451,4.11602938640677
4	5FSS	0,0,0.27,0.15,0,0,0,0,0.048,0,0.261571428571429,0,-0.0947142857142857,0,0,0,0,0,0,-0.191,0,0,0.236571428571429,0.0431428571428571,0,0,0,0,-0.0595714285714286,0,-0.0801428571428572,0,-0.217714285714286,0,0,0,0,0,0,-0.0398571428571429,0,0,0.371142857142857,-0.522285714285714,0,0,0,0,-0.239,0,0.0761428571428572,0,0.317,0,0,0,0,0,0,-0.0777142857142857,0,0,-0.0482857142857143,-0.037,0,0,0,0,-0.210571428571429,0,-0.0395714285714286,0,-0.143571428571429,0,0,0,0,0,0,0.177428571428571,0,0,0.266571428571429,-0.463142857142857,0,0,0,0,-0.0111428571428571,0,0.235428571428571,0,0.173142857142857,0,0,0,0,0,0,-0.180285714285714

Şekil 2.9 Karakteristik Öznitelik Özellikleri

Şekil 2.9’ de gösterilen özniteliklerin kullanılmasıyla, toplam öznitelik sayısı Şekil 2.10’deki gibi olacaktır.

PAM + RP + 5FSS	301
PAM + RP + APAAC	241
PAM + RP + PAAC	221
PAM + RP + PC	214

Şekil 2.10 Toplam Öznitelik Sayıları

3. SONUÇLAR

3.1 Veri Kümesi

Bu çalışmada Passerini ve diğerlerinin hazırlamış olduğu ve çalışmalarında kullandığı veri kümesi kullanılmıştır [6]. Veri kümesi tekrarsız bir şekilde 2727 adet protein dizilimi içermektedir. Çalışma süresince hedef aminoasitler olarak CYS ve HIS seçilmiştir. Bu veri kümesi içerisinde 5635 adet CYS ve 13660 adet HIS bulunmaktadır. Tablo 1' de bu aminoasitlerin metal ile bağlanan toplam sayıları gösterilmiştir.

Tablo 3.1 Veri Kümesi Metal İle Bağlanma Dağılımı

	Metal İle Bağ Yapabilenler	Metal İle Bağ Yapmayanlar
CYS	933	4702
HIS	678	12982

3.2 Değerlendirme Yöntemi

Bu çalışmada performans değerlendirme ölçütü olarak hata matrisi ve eğrinin altında kalan alan (Area Under Curve-AUC) yöntemleri kullanılmıştır.

3.2.1 Hata matrisi

Hata matrisi makine öğrenim alanında uygulanmış olan öğrenim yönteminin performansını ölçmek için yaygın olarak kullanılan bir yöntemdir. Çünkü basit bir şekilde sistemin ürettiği sonuçları görselleştirmektedir. Bu matrisin sütunları sistem tarafın üretilen tahmin durumunu gösterirken, satırları gerçek durumunu göstermektedir. Tablo 2 bu matrisin düzenini göstermektedir.

Tablo 3.2 Hata Matrisi Gösterimi

	Tahmin Pozitif Sınıflar	Tahmin Negatif Sınıflar
Gerçek Pozitif Sınıflar	TP	FN
Gerçek Negatif Sınıflar	FP	TN

TP (Doğru Kabul – True Positive): Sistemin sınıflandırdığı sınıf pozitif iken, ilgili örneğin gerçekte de pozitif olduğu durumdur.

FN (Yanlış Red– False Negative): Sistemin sınıflandırdığı sınıf negatif iken, ilgili örneğin gerçekte pozitif olduğu durumdur.

FP (Yanlış Kabul – False Positive): Sistemin sınıflandırdığı sınıf pozitif iken, ilgili örneğin gerçekte negatif olduğu durumdur.

TN (Doğru Red – True Negative): Sistemin sınıflandırdığı sınıf negatif iken, gerçekte de negatif olduğu durumdur.

Yukarda verilen tanımlar doğrultusunda; Doğruluk değeri, sistemin doğru bir şekilde tahminlemiş olduğu bütün pozitif ve negatif sayılarının bütün örneklem sayısına bölümüne eşittir. Denklem 15'de bu değer f(A) olarak gösterilmiştir.

$$f(A) = {(TP + TN)}\Big/{(TP + FP + TN + FN)} \tag{15}$$

Anma değeri ise, sistemin doğru bir şekilde sınıflandırdığı pozitif örnekler sayısının, doğru şekilde sınıflandırdığı pozitif örnek sayısı ile yanlış olarak pozitif sınıflandırdığı örneklem sayısının toplamına bölümüne eşittir. Denklem 16'da bu değer f(P) olarak gösterilmiştir.

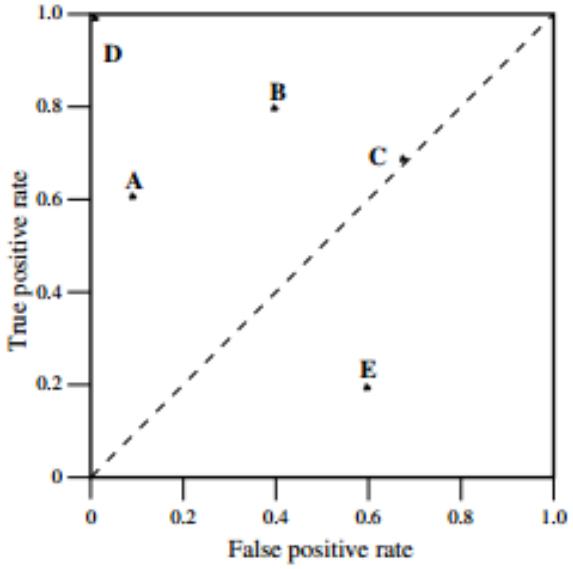
$$F(P) = {(TP)}\Big/{(TP + FP)} \tag{16}$$

Duyarlılık/Duyarlılık değeri ise, sistemin doğru tahminlediği pozitif örnek sayısının, yine bu sayı ile negatif olarak değerlendirilen pozitif örnek sayısının toplamının bölümüne eşittir. Denklem 17'de bu değer f(R) olarak gösterilmiştir.

$$F(R) = {}^{(TP)}\Big/_{(TP+FN)} \tag{17}$$

3.2.2 Eğrinin altında kalan alan (AUC)

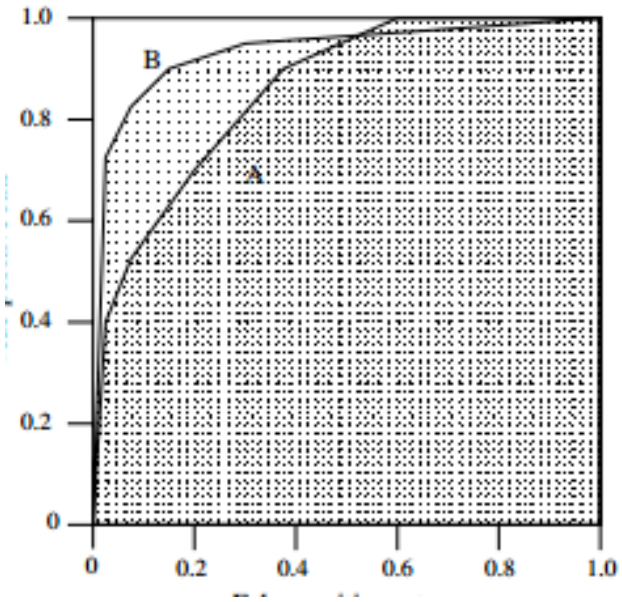
Eğrinin altında kalan alan, alıcı işletim karakteristiği (ROC) grafiğinin oluşturmuş olduğu alanı ifade etmektedir. ROC eğrisi, X ekseninde doğru kabul oranının (true positive rate), Y ekseninde ise yanlış kabul oranın (false positive rate) gösterildiği iki boyutlu bir grafiktir. ROC, bir sınıflandırıcı sistemin, elde edilen doğru kabul seçimleri için, üretilen yanlış kabul sonuçlarını gösterir. Gayri resmi bir ifadeyle, ROC eğrisi, Y eksenine yaklaştıkça izlenen sistemin performansı iyi olarak nitelendirilebilir. Şekil 3.1 de gösterilen ROC grafiğinde, belirtilen A noktası, Y eksenine yaklaşmış olmasına rağmen, düşük değerlerde doğru kabul oranına sahip olduğu görülmektedir. B noktası ise daha yüksek yanlış kabul oranına sahip olmasına rağmen, doğru kabul oranı A noktasından daha yüksek olduğu görülmektedir. Bu da A noktasına sahip bir sistemin doğru kabul işlemi için, belirgin özellikler üzerinden doğru sınıflandırma yaptığını, B noktasına sahip bir sistem ise, daha belirsiz özelliklerden daha doğru sonuçlar ürettiğini söyleyebiliriz. Böylece ROC grafikleri sayesinde iki sistem arasında seçim yaparken, nelerden vazgeçilip, nelerin elde edilebileceği görülebilmektedir. Örneğin A sistemini daha düşük yanlış kabul değerleri için seçildiğinde, düşük doğru kabul yüzdesine sahip bir sistem seçmiş oluruz.



Şekil 3.1 Örnek ROC Eğrisi I [15]

AUC değeri, ROC eğrisinden elde edilen sonuçları tek bir sayısal sonuca indirgenmiş halidir. AUC birim karenin altında ki herhangi bir alanın sayısal ifadesi olduğu için 0 ile 1 arasında bir değer olmaktadır. Rastgele sınıflandırma yapan bir sistemin oluşturduğu ROC eğrisi (0,0) (1,1) noktaları arasında bir köşegen çizecek şekilde olduğundan, 0.5 değerinin altındaki AUC değerleri, ilgili sistemin kıymetsiz olduğunu göstermektedir. AUC, ikili sınıflandırma uygulamalarında ayrım eşik değerinin farklılık gösterdiği durumlarda, duyarlılıkin/duyarlılığın kesinliliğe olan oranıyla ortaya çıkmaktadır. AUC değeri daha basit anlamda doğru pozitiflerin, yanlış pozitiflere olan kesri olarak da ifade edilebilir [7].

Örneğin Şekil 3.2'den yola çıkarak, B eğrisinin altında kalan alanın, A eğrisine oranla daha büyük olduğu kolaylıkla anlaşılmaktadır. Bu eğri aynı zamanda A eğrisini oluşturan sistemin, yanlış kabul oranı 0.6'dan büyük olduğu oranlarda B den daha iyi performans gösterdiğini göstermektedir.



Şekil 3.2 Örnek ROC Eğrisi II [15]

Her sınıflandırma işleminde yapıldığı gibi, metotlar anma ve duyarlılık/duyarlılık arasındaki dengeyi kurmakla uğraşmaktadır. Veri kümesindeki pozitif ve negatif örnekler, eşit bir şekilde dağılım göstermediğinden dolayı, doğrudan anma ve duyarlılık ölçütlerinden önce, AUC değeri, anma ve hasiyet arasındaki dengeyi değerlendirmek için kullanılmıştır. AUC değeri, değişen sınıflandırma eşik değerlerine göre doğru pozitiflerin sayısının, yanlış pozitiflerin bir fonksiyonu olarak çizilmesiyle oluşmaktadır. AUC değerinin 1 (bir) olduğunda anlamı, pozitifler

mükemmel bir şekilde negatiflerden ayrılmıştır, olmaktadır. AUC değerinin 0 (sıfır) olduğunda ise herhangi bir pozitif bulunamadı anlamına gelir.

Bu çalışmada performans ölçütü olarak hata matrisinin yanında AUC da kullanılmıştır. Bunun en önemli sebebi kullanılan veri kümesinde negatif olarak tanımlanmış örnek sayısının, pozitiflere göre oldukça çok olmasıdır. Bu tür dengesiz dağılım gösteren veri kümelerinde AUC çok daha iyimser sonuçlar gösterebilmektedir. Veri kümesindeki dengesiz dağılım Tablo 3.1' de gösterilmiştir.

3.3 Deneysel Sonuçlar

Bu çalışmada 10 farklı öznitelik kümesi oluşturulmuştur. Bu öznitelik kümeleri için hem SVM hem Naïve Bayes sınıflandırıcılarıyla sonuçlar alınmıştır. Bunlara ilaveten değişken uzunluklu Markov zincirleri (VLMC) ve bölgesel hizalama skoruna dayalı sınıflandırma metodu için sınıflandırma skorları elde edilmiştir. Bu dört farklı yöntemle alınan skorlar Bölüm 3.2 belirtilen yaklaşımlarla performans değerlendirmeleri yapılmıştır.

Bu bölümde yukarda bahsedilmiş olan yöntemlerin performans bilgileri okuyuculara aktarılacaktır.

3.3.1 Pam içeren öznitelik vektörleri için sonuçlar

Tablo 3.3' de öznitelik vektörü için sadece PAM 120 matrisi gösterimi kullanılarak yapılan sınıflandırma skorları gösterilmektedir. Anma ve duyarlılık değerleri bir birine yakın olmasına rağmen iyi sayılabilecek seviyenin aşağısındadır. Buna rağmen AUC değeri 0.5 in üstünde bir sonuç üretmiştir.

Tablo 3.3 SVM Sınıflandırıcısı PAM Öznitelikleri İçin Sonuçlar

Anma	Duyarlılık	AUC
0.22	0.24	0.58

Tablo 3.4' de yine öznitelik vektöründe sadece PAM 120 matrisi gösterimi yöntemi kullanılmıştır. Burada elde edilen sonuçlar SVM ye göre dramatik olarak artmıştır. Buda göstermektedir bu öznitelik seçiminde Naïve Bayes sınıflandırıcısı SVM ye göre çok daha iyi sonuç vermiştir.

Tablo 3.4 Naive Bayes Sınıflandırıcısı PAM Öznitelikleri İçin Sonuçlar

Anma	Duyarlılık	AUC
0.65	0.45	0.78

3.3.2 PAM ve 5FSS in birlikte kullanıldığı sonuçlar

Tablo 3.5' de SVM sınıflandırıcısına girdi olarak PAM ve 5FSS' lerden oluşan öznitelik vektörü verildiğinde alınan sonuçlar listelenmiştir. Anma ve duyarlılık sonuçları Tablo 3.4'e göre artmış olsa da AUC değeri sabit kalmıştır. Buda göstermektedir ki 5FSS öznitelikleri sonuca pek bir etkisi olmamıştır.

Tablo 3.5 SVM (PAM + 5FSS)

Anma	Duyarlılık	AUC
0.23	0.25	0.58

Tablo 3.6' da Naïve Bayes'e girdi olarak PAM ve 5FSS'lerden oluşan öznitelik vektörü verildiğinde elde edilen sonuçlar gösterilmiştir. Anma değeri Tablo 3.4'de gösterilen değere göre artmış olsa da duyarlılık değeri düşmüştür. AUC değeri ise bahsi geçen sonuca göre artış göstermiştir. Bu sonuçlarda göstermektedir ki, yanlış ret kararlarında(FN) düşüş sağlanmışken, yanlış kabul kararlarında artış oluşmuştur.

Tablo 3.6 Naïve Bayes (PAM + 5FSS)

Anma	Duyarlılık	AUC
0.75	0.35	0.80

3.3.3 PAM, 5FSS ve bağıl pozisyonun birlikte kullanıldığı sonuçlar

Tablo 3.7'de PAM, 5FSS ve hedef aminoasidin bağıl pozisyon bilgisinin birlikte kullanıldığı zaman SVM'nin üretmiş olduğu değerler gösterilmektedir. Sadece PAM özniteliği kullanılan sonuçlara yakın değerler elde edilmiştir. Bu da bağıl pozisyonunda SVM sınıflandırıcısı için çok fazla seçicilik getirmediğini göstermektedir.

Tablo 3.7 SVM (PAM+5FSS+R)

Anma	Duyarlılık	AUC
0.22	0.24	0.59

Tablo 3.8'de PAM, 5FSS ve R bilgisinin birlikte kullanıldığı zaman Naïve Bayes'in üretmiş olduğu değerler gösterilmiştir. Bu sonuçta göstermektedir ki hem SVM hem de Naïve Bayes sınıflandırıcısında bağıl pozisyon bilgisi seçicilik göstermemektedir.

Tablo 3.8 Naïve Bayes (PAM+5FSS+R)

Anma	Duyarlılık	AUC
0.72	0.36	0.76

3.3.4 PAM ve APAAC'ın birlikte kullanıldığı sonuçlar

Tablo 3.9'da PAM ve APAAC bilgilerinin SVM ye girdi olarak verildiği durumda üretilen sonuçlar gösterilmiştir. Bu sonuçlardan yola çıkarak APAAC bilgisinin seçiciliği düşürdüğü söylenebilir.

Tablo 3.9 SVM (PAM +APAAC)

Anma	Duyarlılık	AUC
0.11	0.18	0.51

Tablo 3.10'da PAM ve APAAC bilgilerinin Naïve Bayes e girdi olarak verildiğinde üretilen sonuçlar gösterilmiştir. Bu sonuçlar, APAAC bilgisinin seçiciliği düşürdüğünü ve performansı negatif yönde etkilediği görülmüştür.

Tablo 3.10 Naïve Bayes (PAM + APAAC)

Anma	Duyarlılık	AUC
0.43	0.18	0.62

3.3.5 PAM, APAAC ve bağıl pozisyonun birlikte kullanıldığı sonuçlar

Tablo 3.11'de göstermektedir ki, APAAC bilgisi sonuçları dramatik olarak

düşürmektedir. Bu öznitelik seçiminde bağıl pozisyon bilgisi de pozitif bir artış sağlamamıştır.

Tablo 3.11 SVM (PAM + APAAC + R)

Anma	Duyarlılık	AUC
0.11	0.18	0.55

Tablo 3.12'de gösterilen değerler Tablo 3.10'un nerdeyse aynısınıdır. APAAC özniteliği hem SVM hem de Naïve Bayes için negatif bir etki yapmıştır.

Tablo 3.12 Naïve Bayes (PAM + APAAC + R)

Anma	Duyarlılık	AUC
0.43	0.18	0.61

3.3.6 PAM ve PAAC' ın birlikte kullanıldığı sonuçlar

Tablo 3.13'de gösterilen sonuçlardan yola çıkarak yanlış Kabul sonuçlarının düştüğünü söyleyebiliriz. SVM için alınmış en iyi sonuçlardan birini göstermektedir. Rahatlıkla PAM değerlerine pozitif bir etki sağladığını söyleyebiliriz.

Tablo 3.13 SVM (PAM + PAAC)

Anma	Duyarlılık	AUC
0.24	0.27	0.60

Tablo 3.14'de PAM ve PAAC bilgilerinin Naïve Bayes'e girdi olarak verildiğinde, üretilen sonuçlar gösterilmiştir. Bu sonuçlar, PAM özniteliği tek başına kullanıldığında da yaklaşık eşit sonuçlar elde edildiğini göstermektedir. PAAC bilgisinin Naïve Bayes ile alınan sonuçlara da artı bir etkisi gözlenmemiştir.

Tablo 3.14 Naïve Bayes (PAM + PAAC)

Anma	Duyarlılık	AUC
0.64	0.44	0.77

3.3.7 PAM, PAAC ve bağıl pozisyonun birlikte kullanıldığı sonuçlar

Tablo 3.15'deki sonuçlardan yola çıkarak bağıl pozisyonun AUC değerlerini

belirgin bir şekilde düştüğünü söyleyebiliriz. Buda yanlış ret sonuçlarının artmış olduğunu söyleyebiliriz.

Tablo 3.15 SVM (PAM + PAAC + R)

Anma	Duyarlılık	AUC
0.24	0.27	0.51

Tablo 3.16'daki sonuçlar, hedef aminoasidin bağıl pozisyonunun sonuca pozitif bir etkisinin olmadığını göstermektedir. Çok küçük oranlarda anma ve duyarlılık değerleri değişmiş olsa da AUC değeri sabit kalmıştır.

Tablo 3.16 Naive Bayes (PAM + PAAC + R)

Anma	Duyarlılık	AUC
0.66	0.45	0.77

3.3.8 PAM ve PC' nin birlikte kullanıldığı sonuçlar

Tablo 3.17'de PAM ve PC özniteliklerinin birlikte kullanılması sonucu SVM'nin üretmiş olduğu sonuçlar listelenmiştir. Bu öznitelikler ile elde edilen sonuçlar APAAC kullanılarak elde edilenlere çok benzediği gözlemlenmiştir. Bunun bir sonucunun fizikokimyasal özelliklerin APAAC değerlerini etkilediği söylenebilir. Sonuç olarak SVM sonuçlarına negatif bir etkisi olmuştur.

Tablo 3.17 SVM (PAM + PC)

Anma	Duyarlılık	AUC
0.11	0.14	0.45

Tablo 3.18'de PAM ve PC özniteliklerinin beraber kullanıldığı durumda, Naïve Bayes sınıflandırıcısının üretmiş olduğu sonuçlar gösterilmiştir. Sonuçların diğer durumlara göre dramatik olarak düştüğünü söyleyebiliriz. Duyarlılık skorundaki dramatik düşüş ise, yanlış kabul seçiminin arttığını göstermektedir.

Tablo 3.18 Naïve Bayes (PAM + PC)

Anma	Duyarlılık	AUC
0.51	0.11	0.59

3.3.9 PAM, PC ve bağıl pozisyonun beraber kullanıldığı sonuçlar

Tablo 3.19'da, PAM, PC ve hedef aminoasidin bağıl pozisyon bilgisinin birlikte kullanılarak oluşturulan öznitelik vektörü ile SVM sınıflandırıcı aracılığı ile elde edilen sonuçlar listelenmiştir. Bu sonuçlar göstermektedir ki, PC değeri net bir şekilde seçiciliği düşürmektedir. Bağıl pozisyon bilgisinin ise etkisi bulunmamaktadır.

Tablo 3.19 SVM (PAM + PC + R)

Anma	Duyarlılık	AUC
0.11	0.13	0.45

Tablo 3.20'de Naïve Bayes sınıflandırıcısın bahsi geçen öznitelikler için üretmiş olduğu sonuçlar gösterilmiştir. PC özniteliğinin sonuçları dramatik olarak düşürdüğü net bir şekilde gözlenmektedir.

Tablo 3.20 Naïve Bayes (PAM + PC + R)

Anma	Duyarlılık	AUC
0.50	0.13	0.64

SVM ve Naïve Bayes ile yapılan çalışmalar, Naïve Bayes'in bu alanda gayet iyi sonuçlar ürettiğini göstermektedir. SVM bu alanda yaygın olarak kullanılan bir yöntem olmasına rağmen bu çalışmada Naïve Bayes sonuçları SVM e göre çok daha iyi olduğu gözlemlenmiştir

.

3.3.10 Değişken değerli markov zincirleri

Bu çalışma içerisinde Markov zincirleriyle geliştirilen sınıflandırma metodu şekil Tablo 3.21'de gösterilen performans değerleriyle sonuç üretmiştir.

Tablo 3.21 Değişken Değerli Markov Zincirleri

Anma	Duyarlılık	AUC
0.60	0.05	0.39

Bu çalışma içerisinde kullanılan diğer yöntemlerden daha düşük bir performans elde edilmiştir. Markov zincirleri yöntemi ile elde edilmiş sonuçlar, hedef aminoasit ve çevresinde sıralı halde bulunan komşu aminoasitlerden oluşan karakter dizileri ile sınıflandırma işlemi yapmanın performansının düşük olduğunu göstermiştir. Duyarlılık skorundaki düşüklük, geliştirilen sistemin, gerçek doğruları tahminlemede zorlandığını göstermektedir. Bunun en önemli sebeplerinden biri ise veri kümesindeki negatif ve pozitif değerlerinin dengesiz bir şekilde dağılmış olmasıdır.

3.3.11 İkili (Pairwise) yaklaşım

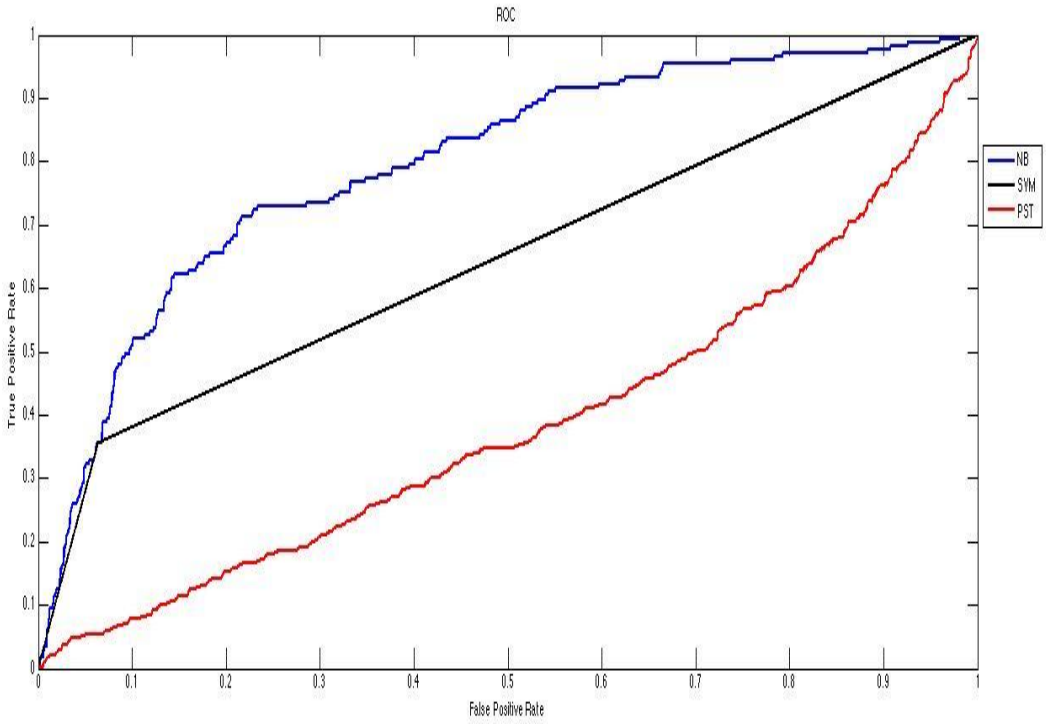
Tablo 3.22 Dizilim Hizalama Tabanlı Yaklaşım

Anma	Duyarlılık	AUC
0	0	-

Bu çalışma içerisinde geliştirilen bölgesel dizilim hizalama skoru üzerinden yapılan sınıflandırma işlemi düşük performans değerleri üretmiştir. Sistem doğru kabul sonucu üretmekte zorlanmıştır. Sistem her örnek için negatif sonucu oluşturmuştur. Bunun en önemli sebeplerinden biri, geliştirilen yöntem içerisinde son aşama olan oylama safhasında gerçekte negatif olan örneklerin çok sayıda çıkmasıdır. Buna yol açan durum ise veri kümesinde bulunan örneklerin dengesiz bir şekilde dağılmış olmasıdır. Elde edilen sonuçlar Markov zincirleri ile elde edilen sonuçlara benzemektedir.

3.3.12 ROC eğriler

Uygulanan yöntemlerden üretici yaklaşım ve ayırt edici yaklaşımlar ile alınan sonuçlar için ROC eğrileri çizdirilmiştir. Şekil 3.3 de gösterilen grafikte SVM ve Naive Bayes yöntemleriyle elde edilmiş en yüksek performans değerleri için ROC eğrileri bulunmaktadır. ROC eğrilerinden de anlaşılacağı üzere, Naive Bayes yöntemi diğer sınıflandırıcılardan daha iyi bir performans göstermiştir.



Şekil 3.3 Elde Edilen En İyi Sonuçların ROC Eğrileri

4. TARTIŞMA

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. Proteinler biyolojik süreçler içerisinde kritik rol üstlenmektedir. Fotosentezden, kan hücreleri ile oksijen taşınmasına kadar birçok hayati sistemin eksiksiz işleyişini sağlamaktadır. Proteinlerin bu çok önemli görevlerini eksiksiz bir şekilde yerine getirmeleri ise üç boyutlu yapılarının korunması ile mümkündür. Üç boyutlu yapılarında ki bozulmalar, almış oldukları görevleri yerine getirmemesine sebep olabilmektedir.

Bu sebeplerden dolayı proteinlerin üç katmanlı yapısının belirlenmesi önemli bir çalışma alanıdır. Bu çalışma alanın bir parçası da, protein dizilimleri üzerinden metal ile bağlanma noktalarının tespitidir. Eğer bu noktalar tespit edilebilir ise, belirlenen noktalardan üç boyutlu yapıya geçiş için zemin hazırlanmış olur.Bu tez çalışmasında, metal iyonlarının protein dizilimleri üzerinde bağlanma noktalarını tespit eden sistemler geliştirilmiştir. Bu sistemlerin geliştirilmesinde farklı metotlar kullanılmıştır.

Protein dizilimleri üzerinden yapılan çalışmalarda sıkça kullanılan ve yüksek performanslı sonuçlar üreten SVM yöntemi tez çalışmamız süresincede kullanılmış ve sonuçlar paylaşılmıştır. SVM yöntemi ile elde edilen sonuçların ortalama değerlere yakın veya üstünde olduğu gözlenmiştir. Çalışma alanımızda pek sık kullanılmayan Naive Bayes yöntemi tez çalışması süresince kullanılmış ve sonuçlar paylaşılmıştır. Naive Bayes yöntemi ile yapılan çalışmaların ürettiği sonuçlar, ilgili yöntemin bu alanda rekabetçi bir metot olabileceğini göstermiştir. Elde edilen sonuçlar SVM metodu ile elde edilen sonuçlara göre daha iyi olduğu gözlenmiştir.

Bu çalışma süresince Değişken uzunluklu Markov zincirleri de kullanılmış ve elde edilen sonuçlar paylaşılmıştır. Bu yenilikçi yaklaşım ile gerçekleştirdiğimiz çalışma, düşük seçicilik ile çalıştığı gözlenmiştir. Bu durumun en büyük sebeplerinden biri olarak örneklem kümesinde gerçekte negatif olan örneklerin pozitif olanlara ziyade

oldukça çok olmasıdır. (Pozitif örneklem sayısının, negatige olan oranı 0.09 dur.) Yukarda bahsedilen çalışmaların dışında, bu çalışma içerisinde dizilim hizalama skoru üzerinden sınıflandırma gerçekleştiren bir yöntem geliştirilmiştir. Bu çalışma ile elde edilen sonuçlar paylaşılmıştır. Tasarlanan sistemin gerçek pozitif değerleri bulamadığı gözlemlenmiştir.

Bu çalışma içerisinde geliştirilen farklı yöntemler üzerinden alınan sonuçların karşılaştırılması dışında, bir protein dizliminin nasıl modelleneceği konusunda da çalışmalar yapılmıştır. SVM ve Naive Bayes yöntemlerinde sadece protein diziliminden elde edilen bilgiler üzerinden farklı farklı öznitelik vektörleri modellenmiştir. Çıkartılan bu öznitelik vektörleri SVM ve Naive Bayes yöntemleri ile eğitilmiş daha sonrasında performans değerleri elde edilmiştir. Elde edilen bu skorlar üzerinden karşılaştırmalar yapılmıştır. Bu karşılaştırmalar sonucunda, performansa pozitif ve negatif etki yapan öznitelikler tespit edilmiştir.

Tez çalışması süresince yapılan bu çalışmanın, alanında pozitif etkiler sağlayacağına inanılmaktadır. Sınıflandırma metotları için yaratılan öznitelik vektörleri bu alanda yapılacak olan yeni çalışmalarda kullanılabileceğine ve seçiciliği artıracağını düşünmekteyiz. Buna ilaveten bu çalışma içerisinde geliştirilen ve yenilikçi bir yöntem olarak ifade ettiğimiz Değişken Uzunluklu Markov Zincirleri yönteminin, ilerde yapılacak olan çalışmalarda tek başına veya birden çok aşamalı sınıflandırma sisteminde sonuçlara pozitif etkiler sağlayacağına inanmaktayız. Aynı zamanda bu alanda pek örneğine rastlanılmayan, dizilim hizalama skorundan yola çıkarak üretilen sınıflandırma yönteminin ilgili araştırma alanına pozitif bir etkisinin olacağını düşünmekteyiz.

Bu çalışmanın geliştirilebilmesi amacı ile elde edilen sonuçların farklı örnek kümeleri üzerinde birleşik etkilerini görmek için, protein veri bankası aracılığı ile yeni örneklem kümeleri yaratılabilir. Yaratılan örneklem kümelerini, bu çalışma içerisinde geliştirilen yöntemler ile eğitip sonuçlar karşılaştırılabilir. Böylece elde edilen yeni örneklem kümeleri bu alanda yapılacak olan yeni çalışmalarda kullanılabilir. Yine bu çalışmanın ilerletilmesi doğrultusunda, aminoasidin metal ile bağla durumunu etkileyebilecek çevresel faktörlerin bu bağlanma durumuna olan etkisi araştırabilir. Böylece özellik vektörleri oluşturulurken ilgili aminoasidin doğal

sık bulunma ortamını belirli bir sayısal değer veya değerler ile modellenebilir. Bu modelleme sonucunda elde edilen sonuçlarda seçiciliği artırıcı sonuçlar gözlemlenebilir. Aynı zamanda bu çalışma, Apache Spark gibi büyük veri işleme teknolojileri kullanılarak tekrarlanabilir. Bunun sonucunda hem yığın veri işleminin gücü hem de paralel işlem yapabilme kabiliyeti ile veri boyutundan neredeyse bağımsız sürelerde sonuçlar elde edilebilir.

KAYNAKLAR LİSTESİ

[1] Smith, Temple F. ve Waterman, Michael S. "Identification of Common Molecular Subsequences". Journal of Molecular Biology 147: pp. 195–197, 1981

[2] G. Bejenora ve G. Yona, “Variations on probabilistic suffix trees: statistical modelling and prediction of protein families”, Bioinformatics Vol.17 No.1, pp. 23-43, 2000

[3] C. Cortes ve V. Vapnik, "Support-vector networks". Machine Learning 20 (3): pp 273, 1995

[4] Russell, Stuart; Norvig, Peter [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall, 2003

[5] L. Rishishwar, N. Mishra, B. Pant, K. Pant, ve K. R. Pardasani, ProCoS - PROtein COmposition Server, Bioinformation, 5(5): 227. PMC: 3040505, 2010.

[6] A. Passerini, M. Punta, A. Ceroni, B. Rost, ve P. Frasconi,“Identifying Cysteines and Histidines in Transition-Metal-Binding Sites Using Support Vector Machines and Neural Networks,” Proteins, vol. 65, no. 2, pp. 305-316, 2006

[7] Swets, John A.; Signal detection theory and ROC analysis in psychology and diagnostics : collected papers, Lawrence Erlbaum Associates, Mahwah, NJ, 1996

[8] M. Boulle, “Parsimonious Naïve Bayes”, 2014 Federated Conference on Computer Scşence and Information System (FedCSIS), pp. 355-359, 2014

[9] A. Passerini, M. Lippi ve P. Frasconi, “Predicting Metal-Binding Sites from Protein Sequence”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 9 No.1, 2012.

[10] Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

[11] M. Lippi, A. Passerini, M. Punta, B. Rost ve P. Frasconi, “MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence”, Bioinformatics Applications Note, Vol.24, no.18, pp.2094-2095, 2008.

[12] A. Passerini, M. Lippi ve P. Frasconi, “MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence”, Nucleic Acids Research Advance Access, pp.1-5, 2011.

[13] N. Shu, T. Zhou ve S. Hovmöller, “Prediction of zinc-binding sites in proteins from sequence”, Bioinformatics Original Paper, Vol. 24, no.6, pp. 775-782, 2008.

[14] H. Oğul and E. Mumcuoğlu, “SVM-based detection of distant protein structural relationships using pairwise probabilistic suffix trees”, Computational Biology and Chemistry Vol.30, pp. 292-299, 2006.

[15] T. Fawcett, “An introduction to ROC analysis”, Pattern Recognition Letters 27, pp. 861-874, 2006.