T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HIV-1 PROTEAZ ENZİMİNİN KESME KONUMLARININ TESPİTİNDE YENİ ÖZNİTELİK VEKTÖRLERİ

DOKTORA TEZİ

Murat GÖK

Enstitü Anabilim Dalı : ELEKTRONİK-BİLGİSAYAR EĞİTİMİ

Tez Danışmanı : Yrd. Doç. Dr. Ahmet Turan ÖZCERİT

Mayıs 2011

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HIV-1 PROTEAZ ENZİMİNİN KESME KONUMLARININ TESPİTİNDE YENİ ÖZNİTELİK VEKTÖRLERİ

DOKTORA TEZİ

Murat GÖK

Enstitü Anabilim Dalı : **ELEKTRONİK-BİLGİSAYAR EĞİTİMİ**

Tez Danışmanı : **Yrd. Doç. Dr. Ahmet Turan ÖZCERİT**

Bu tez 27 / 05 / 2011 tarihinde aşağıdaki jüri tarafından Oyçokluğu ile kabul edilmiştir.

Prof. Dr.	**Prof. Dr.**	**Doç. Dr.**
İsmail ERTÜRK	**Akif KUTLU**	**Cabir VURAL**
Jüri Başkanı	**Üye**	**Üye**
Yrd. Doç. Dr.		**Yrd. Doç. Dr.**
Ahmet Turan ÖZCERİT		**Fahri VATANSEVER**
Üye		**Üye**

ÖNSÖZ

Bu tez çalışmasında, HIV-1 proteaz enziminin kesme yerleri tahmininin, makine öğrenmesi algoritmaları ile modellenerek geliştirilmesine yönelik çalışılmıştır. Bu kapsamda FTKY, BirTVD ve BirBOOL adları verilen üç adet öznitelik kodlama yöntemi geliştirilmiştir. Geliştirilen bu yöntemler HIV-1 proteaz enzimi özgünlüğü problemi üzerinde geçmişte yapılan çalışmalar ile başta doğrusal Destek Vektör Makineleri sınıflandırıcısı olmak üzere makine öğrenmesi yöntemlerine göre deneysel olarak karşılaştırılmıştır. Elde edilen sonuçlara göre geliştirilen yöntemler ve mevcut yöntemlerin başarım değerlendirmesi yapılmıştır.

Bu çalışmanın gerçekleşmesi sırasında gösterdiği ilgi ve emek ile beni her konuda destekleyen, bu çalışmanın sonuca ulaşmasını sağlayan değerli Hocam Yrd. Doç. Dr. Ahmet Turan ÖZCERİT'e teşekkürlerimi sunarım. Çalışmalarım süresince bilgilerini benimle paylaşan, yardımlarını esirgemeyen Doç.Dr. Cabir VURAL ve Yrd. Doç. Dr. Hasan OĞUL'a teşekkürlerimi sunarım.

Tez çalışmam boyunca her zaman her zorlukta yanımda olan, duaları ile manevi olarak destekleyen anneme, babama ve bana her zaman inanan ablalarıma teşekkürlerimi ve sevgilerimi sunarım.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

AIDS	: Acquired Immuno Deficiency Syndrome
HIV	: Human Immuno Deficiency Virus
TVD	: Taylor'un Venn Diyagramı
PVB	: Protein Veri Bankası
CA	: Capsid Proteini
TFP	: Transframe Proteini
PR	: Proteaz Enzimi
RTp51	: Ters Transkriptaz
RTp66	: Ters Transkriptaz-RNase H
IN	: İntegraz
TT	: Ters Transkripsiyon
FDA	: A.B.D. Yiyecek ve İlaç Kurumu (U.S. Food and Drug Administration)
DVM, SVM	: Destek Vektör Makineleri
TBA, PCA	: Temel Bileşen Analizi
DAA, LDA	: Doğrusal Ayırıcı Analizi
PR-1625	: Kontijevskis'e ait HIV-1 Proteaz Özgünlüğü Veri Seti
PR-3261	: Schilling'e ait HIV-1 Proteaz Özgünlüğü Veri Seti
BKY	: Birimdik Öznitelik Kodlama Yöntemi
ATKY	: Ağırlık Tabanlı Öznitelik Kodlama Yöntemi

KÇKY : Kalıntı Çiftleri Öznitelik Kodlama Yöntemi

FTKY : Fizikokimyasal Özellikler Tabanlı Öznitelik Kodlama Yöntemi

BirTVD : Birimdik Taylor'un Venn Diyagramı Öznitelik Kodlama Yöntemi

BirBOOL : Birimdik Boolean Öznitelik Kodlama Yöntemi

ÇDT, CV : Çapraz Doğrulama Tekniği

AİK, ROC : Alıcı İşletim Karakteristiği

AİKAA : AİK Eğrisi Altında Kalan Alan

$R(\alpha)$: Umulan risk

$R_{emp}(\alpha)$: Deneysel risk

$p=(x,y)$: Ortak olasılık dağılımı

$\Re$: Vektör uzayı

x_i : Eğitim örneklerinin değerleri

y_i : Eğitim örneklerine karşılık gelen etiketlerin değerleri

L_P : İlkel Lagrange ifadesi

L_d : İkili Lagrange ifadesi

α_i : i. Lagrange çarpanı

b : Bias

$d(\mathrm{x},\mathrm{w},b)$: Örüntü tanıma sistemi karar fonksiyonu

S : Sınır boşluğu

μ : Eğitim örneklerinin ortalaması

$\bar{x}$: Eğitim örneklerinin ortalama matrisi

$Cov(.)$: Eğitim örneklerinin kovaryans matrisi

W : Sütun özvektörleri

P : Peptit

↓ : Makas bağ

δ_{ij} : Kronecker delta sembolü

d_i : i. birimdik vektör

$\{\vec{y}\}_i^1$: P_i'nin BKY karşılığı olan vektör

$\{\vec{y}\}_i^2$: Amino asitlerin fizikokimyasal özellik vektörü

$\vec{\chi}$: $\{\vec{y}\}_i^1$ ve $\{\vec{y}\}_i^2$ vektörlerinin birleşimi olan peptit öznitelik vektörü

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: AIDS, HIV, HIV-1 Proteaz Enzimi, Proteaz Özgünlüğü, Örüntü Tanıma, Öznitelik Kodlama Yöntemi Yöntemleri, Destek Vektör Makineleri, Temel Bileşenler Analizi

Canlıların vücudunda bulunan proteaz enzimleri, pek çok yararlı biyolojik işlevi yerine getirirler. Bununla beraber, virüsler, parazitler gibi pek çok bulaşıcı mikroorganizmalar, proteazları enfekte olabilmek için kullanırlar. Proteazların temel görevi yeni sentezlenmiş çoklu proteinleri uygun yerlerinden keserek yapısal hale gelmelerini sağlamaktır. Böylece, ait oldukları mikroorganizmanın olgunlaşması ve çoğalmasında rol alırlar. Bu nedenle proteazların özgünlüklerini çözmek ilaç ve aşı geliştimek için çok önemlidir. Bununla beraber, proteaz enzimlerinin özgünlükleri konusunda yetersiz bilgi bulunmaktadır. Bu nedenle laboratuvar ortamlarında, proteaz verileri elde etmek ve proteazların özgünlüklerini karakterize etmek için uygun biyobilişim öznitelik kodlama yöntemleri ve algoritmaları geliştirmek hayati derecede önemlidir. Bu tezde, Human Immunodeficiency Virüs Tip 1 (HIV-1) proteazının proteinleri kesme konumlarının tespiti üzerine çalışılmıştır.

Proteinlerle çalışırken göz önününde bulundurulması gereken iki temel bilgi bulunmaktadır: kalıntıların birbirleri ile olan fizikokimyasal etkileşimleri ve protein dizilimi içindeki konumları. Bu iki temel bilgi, proteinin işlevini anlamada nirengi noktalarıdır ve HIV-1 proteazının çoklu proteinleri nereden keseceğinin tahmin edilmesinde kullanılabilir. Bu varsayımdan yola çıkarak, HIV-1 proteaz enzimi özgünlüğünün modellenmesinde Fizikokimyasal Tabanlı Kodlama Yöntemi (FTKY), Birimdik Taylor Venn Diyagramı (BirTVD) ve Birimdik BOOL (BirBOOL) olarak isimledirilen üç öznitelik kodlama yöntemi geliştirilmiştir.

HIV-1 proteazın kesme konumlarını tespit etmek için güncel iki HIV-1 proteaz veri setlerine ait peptit örüntüleri, öznitelik çıkarım yöntemleri ile kodlanmıştır. Bu kodlanan örneklerin öznitelikleri Temel Bileşenler Analizi (TBA) ve Doğrusal Ayırıcı Analiz (DAA) ile çıkarılmıştır. Ardından doğrusal Destek Vektör Makineleri (DVM) algoritması ile sınıflandırılmıştır. Elde edilen deneysel sonuçlara göre; BirTVD ve BirBOOL öznitelik çıkarım kodlama yöntemlerinde, başarım mevcut yöntemlere göre daha yüksek elde edilmiştir.

NEW FEATURE VECTORS ON PREDICTION OF HIV-1 PROTEASE ENZYME CLEAVAGE SITES

SUMMARY

Key Words: AIDS, HIV, HIV-1 Protease Enzyme, Protease Specificity, Pattern Recognition, Feature Encoding Schemes, Support Vector Machines, Principal Components Analysis

Protease enzymes which are inside the living organisms, implement many useful biological functions. However, many infectious microorganisms such as viruses and parasites use proteases to be infected as virulence factors. The main task of proteases is to cleave the polyproteins synthesized newly at the appropriate places to make them structural components. In this way, virulent proteases take role in maturation and replication of microorganisms. Hence, unravelling the specificities of proteases is of great importance to develop drugs and vaccines. However, little is known about the cleavage specificities of these proteases. It is therefore, an important challenge to collect experimental protease data and to develop appropriate bioinformatics feature encoding schemes, algorithms to characterize the specificities for all proteases. In this thesis, human immunodeficiency virus type 1 (HIV-1) protease site prediction has been studied.

When studying on proteins, there are two basic points considered: physicochemical relationships and the positions of the residues in protein sequnces. These two references are the keys to understand the functions of the proteins and can be used to predict where HIV-1 protease cleave the polyproteins. This hypothesis leads us to develop three feature encoding schemes namely FTKY, BirTVD and BirBOOL to model specificity of HIV-1 protease.

For the prediction of HIV-1 protease cleavage sites, peptide samples of two up-to-date HIV-1 protease datasets have been encoded with feature encoding techniques and extracted their features with Principal Components Analysis and Linear Discriminant Anaysis. Subsequently, they have been classified by Linear Support Vector Machines algorithm. According to empirical results obtained, BirTVD and BirBOOL methods have achieved better performance compared to hitherto methods.

BÖLÜM 1. GİRİŞ

Amino asitler; proteinler, peptitler, bazı hormonlar, vitaminler ve antibiyotikler gibi hayati öneme sahip bileşiklerin temel yapıtaşlarıdır. İnsan gen haritasında (DNA - Deoksiribonükleik Asit), kodonlar (üçlü nükleotit dizilimi) tarafından kodlanmış olan genetik kodlar önce Ribonükleik Asit'e (RNA) kopyalanırlar ve daha sonra amino asit bloklarına diğer bir ifade ile proteinlere çevrilirler. Gezegendeki tüm yaşam biçimlerinin ilk adımı olan genetik kod, DNA'nın keşfinin 10 yıl kadar sonrasında anlaşılabilmiştir 1. Harflerin kelimeleri oluşturması gibi amino asitler de birbirlerine bağlanarak proteinleri oluştururlar 2. Vücutta sentezlenen her protein molekülü fonksiyoneldir ve hiçbir zaman amino asit deposu değildir. Proteinler hücre içerisinde çeşitli biyokimyasal tepkimelere girerek canlı bünyesinde hayati görevler üstlenirler. Besinlerin sindirilmesinden kalıtsal özelliklerin yeni bir canlıya aktarılmasına varıncaya kadar bütün yaşam süreçleri biyokimyasal tepkimelere dayanır. Canlının yapısında bulunan elementlerin birbiriyle etkileşimi bütün kimyasal tepkimeler için geçerli olan temel yasalar çerçevesinde gerçekleşir. Canlı hücredeki biyokimyasal tepkimeler cansız ortamdaki kimyasal tepkimelere göre farklılıklar gösterir. Bu farklılıkların başında biyokimyasal tepkimelerde enzim adı verilen büyük proteinlerin kullanımı gelir. Enzimler, canlı organizmalar tarafından üretilen, farklı maddeler içeren, belirli bir kimyasal reaksiyonu kolaylaştıran, kendisi reaksiyondan bozulmadan ve değişikliğe uğramadan çıkabilen protein molekülleridir 3. Enzimler biyokimyasal reaksiyonların hızını kimyasal katalizli reaksiyonlara kıyasla 1 milyon kez artırırlar. Bütün biyokimyasal süreçler enzimlerce denetlendiği için, basit yapılı bir bitki yapısında bile yüzlerce enzim vardır. Canlının yapısı karmaşıklaştıkça, yapısındaki biyokimyasal tepkime miktarı da artar ve çeşitlenir. Dolayısıyla her biri ayrı bir biyokimyasal tepkimeye özgün (specificity) özelliklere sahip enzimlerin sayısı binleri bulur. Enzimlerin özgünlük şifrelerinin çözülmesi ilaç tasarımında ve geliştirilmesinde önemli rol oynar. Bu durum özellikle AIDS

hastalığına neden olan HIV gibi bulaşıcı mikroorganizmalar için baskılayıcı ilaç ve aşı gelişimi için hayati önemdedir.

HIV-1 proteaz, HIV-1'in yaşam döngüsü için hayati önem taşıyan bir enzimdir. HIV-1 proteaz, uzun protein dizilimlerini keserek işlevsel ve yapısal protein dizilimleri oluşmasını sağlar. HIV-1 proteaz enzimi, peptitlerde kesme işlemini herhangi bir motifsel ve basit bir yönteme göre yapmamaktadır, karmaşık bir yöntem uygulamaktadır 4. Proteaz enzimi, kesme işlemini gerçekleştirmezse virüs olgunlaşamaz ve enfekte olabilme kabiliyetini kaybeder. Kesme yerlerinin labaratuvar ortamlarında tespiti oldukça zor ve zaman alıcıdır. Bu nedenle bilgisayar ortamında HIV-1 proteaz kesme konumlarının tespitinde yapay zekâ tekniklerinden faydalanılması gerekmektedir. Bu çalışmada amaç, makine öğrenmesi algoritmaları için giriş olarak kullanılacak peptit örüntülerinin yeni bir öznitelik çıkarım yöntemi ile kodlanması ve böylece yüksek doğruluk oranları ile HIV-1 proteazın enziminin kesme konumlarını tespit etmektir. Bu kapsamda literatürde bulunan öznitelik kodlama yöntemleri araştırılmış ve daha yüksek başarım sağlayan üç öznitelik kodlama yöntemi geliştirilmiştir.

Tez çalışması sekiz bölümden oluşmaktadır. Giriş bölümünün ardından ikinci bölümde AIDS hastalığı ve HIV'in yapısı, yaşam çevrimi üzerinde durulmuştur. Ayrıca HIV-1 proteaz enziminin işlevi ve HIV için önemi anlatılmıştır. Üçüncü bölümde, örüntü tanıma sistemi aşamaları, yöntemlerin uygulanmasında kullanılan DVM ve TBA matematiksel çıkarımları ile beraber tartışılmıştır. Dördüncü bölümde mevcut HIV-1 proteazın özgünlüğünün modellenmesinde kullanılmış öznitelik kodlama yöntemlerinin üstünlükleri ve kısıtları ile açıklanmıştır. Beşinci bölümde, fiziko kimyasal özelliklere dayananılarak tez kapsamında geliştirilen FTKY tanıtılmıştır. Altıncı bölümde Birimdik Kodlama Yöntemi (BKY) ve Taylor'ın Venn Diyagramı (TVD) yöntemi temelinde geliştirilen BirTVD yöntemi detaylı olarak ele alınmıştır. Yedinci bölümde ise amino asitlerin fizikokimyasal özellikleri özgün bir sınıflandırmaya tabi tutularak BKY ile birleştirilmesi neticesinde geliştirilen BirBOOL yöntemi geliştirilmiştir. Son bölümde ise tez çalışması süresince geliştirilen ve literatürde bulunan öznitelik kodlama yöntemleri karşılaştırılmış ve elde edilen sonuçlar kapsamlı olarak değerlendirilmiştir.

BÖLÜM 2. AIDS HASTALIĞI ve HIV

AIDS (Acquired Immuno Deficiency Syndrome) hastalığı her yıl milyonlarca insanı etkileyen bulaşıcı bir hastalıktır. 22,5 milyonu Afrika kıtasında olmak üzere toplam 33,3 milyon HIV (Human Immunodeficiency Virus) bulaşmış insan vardır 5. HIV'in bulaşması ile vücudun bağışıklık sistemi hızla zayıflar ve AIDS hastaları bir takım ciddi sağlık sorunlarına maruz kalırlar. Bu sağlık sorunları basit bir grip virüsü olabileceği gibi çeşitli kanser hastalıkları (Kaposi's sarcoma, rahim ağzı ve bağışıklık sistemi kanserleri) da olabilir. Bağışıklık sistemi çok zayıfladığı için çok rahat atlatabilecek virüs, bakteri, mantar veya parazit enfeksiyonları dahi ölümcül olabilir. Günümüzde aşı çalışmaları ve tedavi araştırmaları son hızla devam etse de AIDS hastalığının tedavisi bulunamamıştır ve gelecekte de bulunacağına dair garanti yoktur 6.

Biyokimyanın yapıtaşı olan amino asitler HIV'in yapısı ve biyokimyasal süreçlere dayanan çoğalımını açıklamak için ilk adımdır.

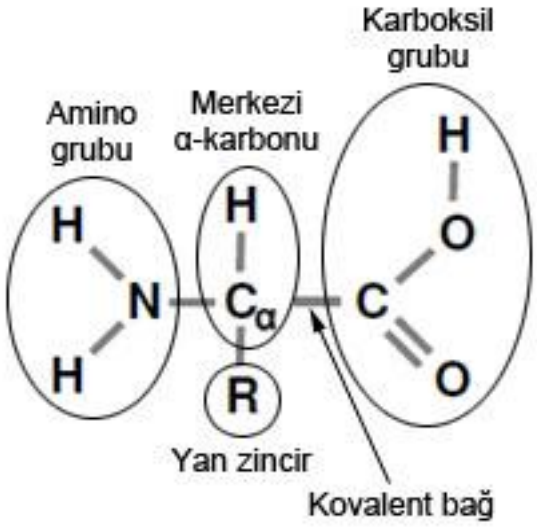
2.1. Amino Asitler

Amino asitler, yapılarında hem amino grubu ($-NH_2$) hem de karboksil grubu ($-COOH$) içeren bileşiklerdir. Doğada 300 kadar farklı amino asit bulunmaktadır. Tablo 2.1'de görülen 20 standart amino asit, DNA tarafından kodlanarak proteinler ve diğer biyomoleküllerin sentezinde kullanılırlar. Amino asitlerin fazlası atılmaz ve depolanmaz, bunlar hücre içinde yakıt metabolizmasına dahil olmak üzere yıkılırlar.

Tablo 2.1. 20 standart amino asit

Sıra	Amino Asit	1-harf	3-harf		Sıra	Amino Asit	1-harf	3-harf
1	Alanin	A	Ala		11	Lösin	L	Leu
2	Arginin	R	Arg		12	Lizin	K	Lys
3	Asparajin	N	Asn		13	Metiyonin	M	Met
4	Aspartik asit	D	Asp		14	Fenilalanin	F	Phe
5	Sistein	C	Cys		15	Prolin	P	Pro
6	Glütamin	Q	Gln		16	Serin	S	Ser
7	Glütamik asit	E	Glu		17	Treonin	T	Thr
8	Glisin	G	Gly		18	Triptofan	W	Trp
9	Histidin	H	His		19	Trozin	Y	Tyr
10	İzolösin	I	Ile		20	Valin	V	Val

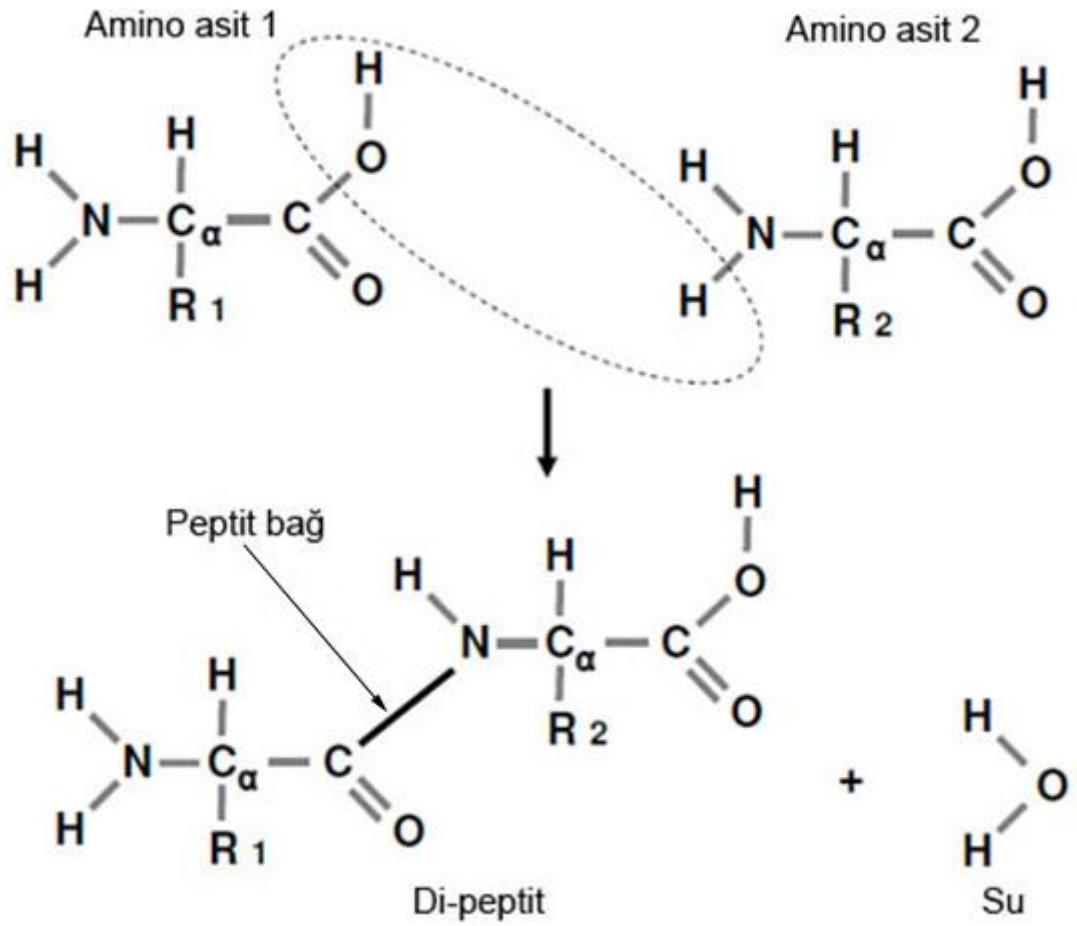
Şekil 2.1'de görüldüğü üzere her bir standart amino asit dört kısımdan meydana gelir: merkezi α-karbonu (C_α), amino ($-NH_2$) ve karboksil ($-COOH$) grupları ile yan zincir (R) grubu. Amino, karboksil, ve yan zincir grupları kovalent bağlar ile merkezi α-karbonuna bağlıdırlar.



Şekil 2.1. Standart bir amino asidin yapısı

Amino asitler birbirlerine peptit bağlar ile bağlanırlar. Bir peptit bağ oluşturabilmek için iki amino asidin, amino ve karboksil grupları tepkimeye girerler. Biyokimyasal

bir tepkimede, bir amino asitin karboksil grubu başka bir amino grubuna bağlanmasıyla peptit bağ oluşur. Şekil 2.2'de görüldüğü gibi tepkime sonrasında peptit bağ ile beraber su ortaya çıkar. Amino asitlerin peptit bağlar kurarak oluşturdukları bileşiklere peptit denir. Küçük bir peptit, 10 ila 50 arası amino asitten meydana gelebilir.



Şekil 2.2. Bir çift amino asidin bir peptit bağ oluşturması

Amino asitlerin fizikokimyasal özelliklerini her amino asitte bulunan özgün farklılıklar gösteren yan zincir grubu belirler 7.

2.1.1. Amino asitlerin fizikokimyasal özellikleri

Kimyanın bir dalı olan, fiziksel yöntemler üzerine kurulu fizikokimya, moleküllerin doğasını açıklamak için bu moleküllerin birbirleri ile olan etkileşimleri ve bu etkileşimler sırasında meydana gelen enerji alış verişlerini inceler. Bu etkileşimler moleküllerden oluşan amino asitlerin ait oldukları proteinlerin işlevlerini belirler. Amino asitler, hidrofobiklik, polarlık, moleküler ağırlık gibi pek çok birbirinden farklı fizikokimyasal özelliklere sahiptirler. Bu özellikler, amino asit indeksi adı verilen 20 sayısal değerden oluşan vektörler ile ifade edilebilirler. Proteinlerin sınıflandırılmasında amino asitlerin fizikokimyasal özelliklerinden sıkça faydalanılmıştır. Tablo 2.2'de amino asitlerin fizikokimyasal özelliklerinin niceliksel ifade edildiği indeks tablosu görülmektedir 8.

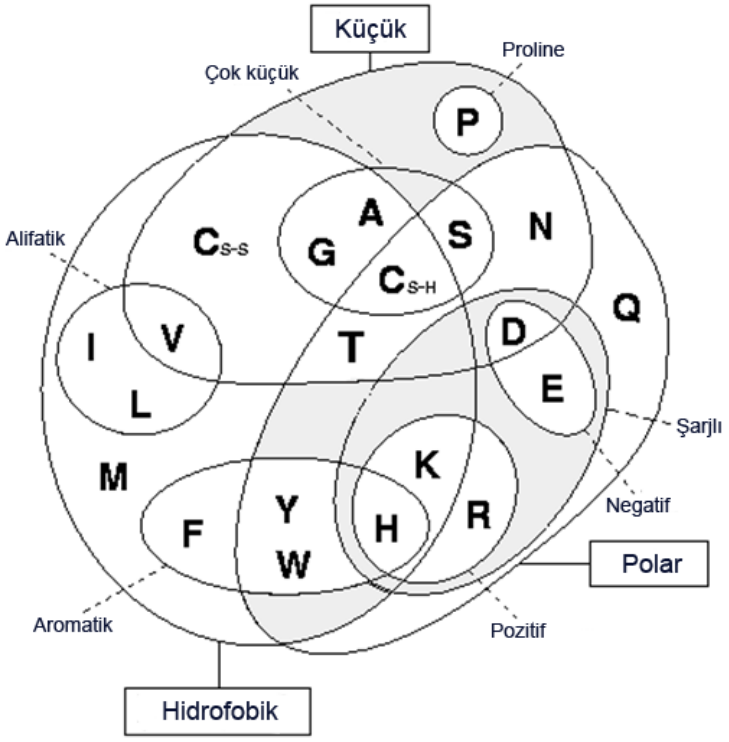
Örneğin 544. fizikokimyasal özellik olan hidrofobiklik, her bir amino asit için -6,04 ila 3,88 arasında değişen değerler almaktadır. Bu değerler amino asitlerin hidrofobiklik derecelerini belirlemektedir.

Amino asitlerin fizikokimyasal özellik indeks verilerini barındıran AAindex adında çevrim içi bir veri tabanı bulunmaktadır 8. AAindex'in en son versiyonu olan 9,1 - Ağustos 2006, 544 adet fizikokimyasal özelliğe ait veri içermektedir.

Tablo 2.2. Amino asitlerin fizikokimyasal özelliklerine ait indeks tablosu örneği

Amino asit	1. fizikokimyasal özellik	…	63. fizikokimyasal özellik	…	544. fizikokimyasal özellik
	Alpha-CH kimyasal kaydırma (Andersen et al., 1992)	…	Büyüklük (Dawson, 1972)	…	Hidrofobiklik indeksi (Fasman, 1989)
A	4,35	…	2,5	…	-0,21
R	4,38	…	7,5	…	2,11
N	4,75	…	5	…	0,96
D	4,76	…	2,5	…	1,36
C	4,65	…	3	…	-6,04
Q	4,37	…	6	…	1,52
E	4,29	…	5	…	2,3
G	3,97	…	0,5	…	0
H	4,63	…	6	…	-1,23
I	3,95	…	5,5	…	-4,81
L	4,17	…	5,5	…	-4,68
K	4,36	…	7	…	3,88
M	4,52	…	6	…	-3,66
F	4,66	…	6,5	…	-4,65
P	4,44	…	5,5	…	0,75
S	4,5	…	3	…	1,74
T	4,35	…	5	…	0,78
W	4,7	…	7	…	-3,32
Y	4,6	…	7	…	-1,01
V	3,95	…	5	…	-3,5

Taylor 9, amino asitleri fiziksel, kimyasal ve yapısal özelliklerine göre Şekil 2.3'de görüldüğü gibi sınıflandırılmıştır. Amino asitler Venn diyagrama, Dayoff'un yer değiştirme matrisinin çok boyutlu ölçeklendirilmesi yapılarak konumlandırılmışlardır. Sınıflandırma amino asitlerin hidrofobiklik, polarlık ve boyut temel özelliklerine göre yapılmıştır.



Şekil 2.3. Taylor'un standart 20 amino asit venn diyagramı

Diyagramda amino asitlerin suda çözünmeye elverişli olmaları polar (hidrofilik) özelliği ile, sudan kaçınma hidrofobik (apolar) özelliği ile sınıflandırılmıştır. Hidrofobik özelliği olan amino asitlerin R grupları fizyolojik pH'da (canlı organizmasında) iyonlaşmaz, hidrojen ve iyonik bağların yapısında yer almaz. Protein yapılarında hidrofobik etkileşimle üç boyutlu yapılarının kazanılmasında rol oynarlar. Şarjlılık ise iki alt kategoriye ayrılmaktadır: pozitif ve negatif. Diğer bir ifade ile, D ve E amino asitleri fizyolojik pH'da negatif yüklüdürler ve asidik özellik gösterirler. H, K ve R amino asitleri fizyolojik pH'da pozitif yüklüdürler ve bazik özellik gösterirler. Diğer amino asitler ise yüksüzdür. Amino asitlerin boyut özelliği küçük ve çok küçük olmak üzere iki alt kategoriye ayrılmıştır. Ayrıca Venn diyagram yan zincirinde benzen, benzen türevleri ve naftalin gibi benzen halkası ihtiva eden amino asitleri tanımlayan aromatiklik ile alifatiklik özelliklerine ait sınıflandırmaları da içermektedir. Yapısında amino grubu yerine imino grubu (−NH) taşıyan P'de diyagramda ayrı bir alt kategoride bildirilmiştir. Ayrıca C amino asiti hücre içinde bir protein dizilimine dahil ise C_{S-H} ile hücre dışında bir protein

dizilimine dahil ise $C_{S\text{-}S}$ ile Venn diyagramda gösterilmiştir. Çünkü C amino asiti hücre içinde ve hücre dışında farklı kimyasal özellikler göstermektedir 7.

Taylor'un venn diyagramı (TVD), amino asitleri genel anlamda sınıflandırmada yeterli olsa da basit bir sınıflandırma tehlikesi söz konusudur. Örneğin amino asitlerin hidrojen bağı kurabilme kabiliyetleri bu sınıflandırmada iyi bir şekilde ele alınmamıştır 7.

2.2. Proteinler

Proteinler, amino asit çiftlerinin polimerleşmesi sonucunda sentezlenirler. Canlılarda DNA ve RNA ne zaman, hangi proteinin gerektiğini enzimler aracılığıyla hücreye bildirerek protein sentezini yönlendirirler. Hücre içerisinde ribozomlar, mesajcı RNA (mRNA) moleküllerini kalıp olarak kullanarak amino asitleri uç uca ekleyerek proteinleri sentezlerler ve bu işleme translasyon denir. Sentezlenen her bir proteindeki amino asit dizisinin sırası bir gen tarafından tanımlanır. Bir protein zincirindeki amino asitler bir dehidrasyon tepkimesi sonucu oluşan peptit bağı ile birbirlerine bağlanırlar. Protein zincirine dahil olmuş amino asit birimlerine kalıntı (residue) denir. Hücre içerisinde her bir süreçte görev alan proteinler canlı organizmaların temel bileşenlerindendir. Çoğu proteinler, biyokimyasal tepkimelerde katalizör işlevi gören ve canlı için yaşamsal öneme sahip olan enzimlerdir.

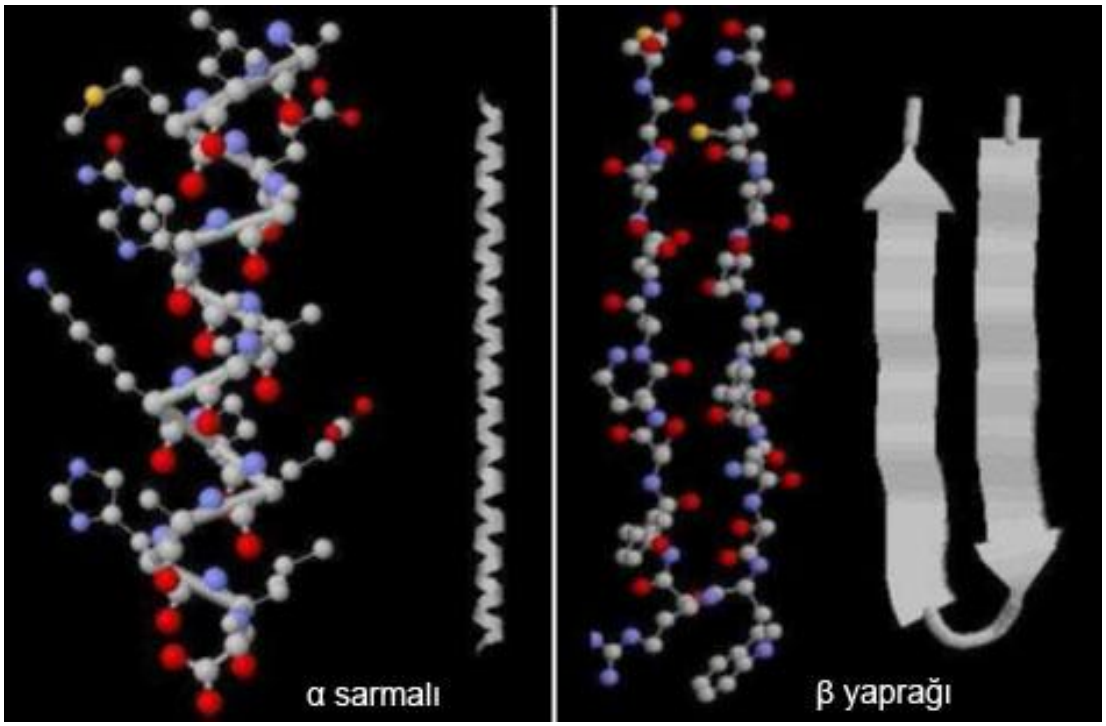
Protein dizilimleri, serbest bir amino grubu ucu olan N-terminalinden, serbest bir karboksil grubu ucu olan C-terminali doğrultusunda ifade edilirler.

Biyokimyagerler proteinlerin yapısını dört farklı şekilde ifade ederler 10. Bunlardan ilki olan birincil yapı (primary structure), proteinleri oluşturan amino asitlerin hangi sırayla birbirlerine bağlı olduklarını gösteren basit yapıdır. Diğer üçü ise proteinlerin üç boyutlu katlanma yapılarına dayalıdır.

İkincil yapı (secondary structure), hidrojen bağları ile kararlı kılınan, düzenli tekrarlanan geometrik yapılardır. α sarmalı (α helix) ve β yaprağı (β sheet) en yaygın ikincil yapılardır. Şekil 2.4'de ikincil yapı gösterimleri görülmektedir.

Üçüncül yapı (tertiary structure), proteinin üç boyutlu gösterimdir.

Dördüncül yapılar (quarternary structure) ise birden fazla çoklu peptit içeren karmaşık, büyük proteinler için geçerli yapılardır. Dördüncül yapılar, protein içindeki peptit dizilimlerinin birbirleri ile olan etkileşimlerini tanımlar. Proteinlerin yapıları ile ilgili bilgilere Protein Veri Bankası (PVB)[1] aracılığıyla ulaşılabilir. PVB, proteinler ve nükleik asitler gibi biyolojik makro molekül yapılarını barındıran büyük bir veri tabanıdır 11.



Şekil 2.4. Bir α sarmalı ve bir β yaprağı yapısı

Enzimlerin özgünlüğünün anlaşılması açısından protein-substrat etkileşiminin modellenmesi büyük önem taşımaktadır. Böylece bağlanma yerlerinin labaratuvar (in vitro) ortamlarında tespit edilmesi ve baskılayıcı ilaçların (inhibitör) geliştirilmesi yolu açılır.

[1] http://www.pdb.org

2.3. Proteaz Enzimleri ve Substratlar

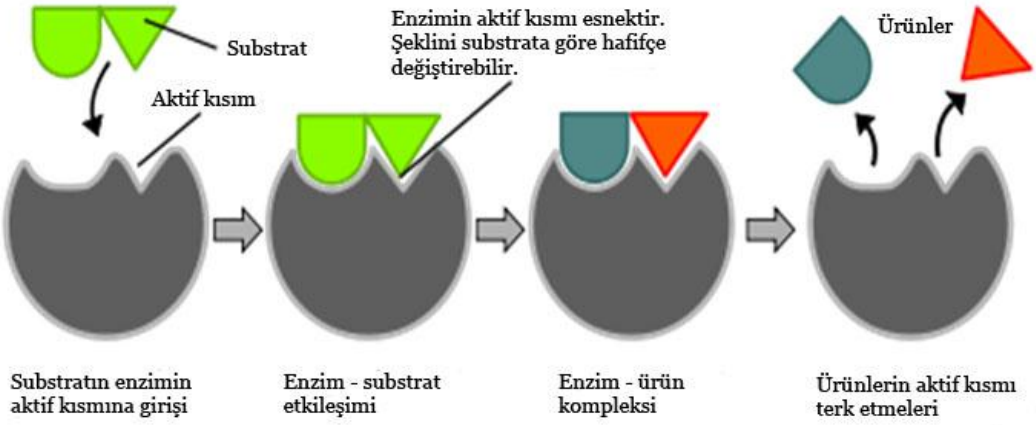
Çoklu proteinlerden oluşan proteazlar, proteinleri oluşturan amino asitler arasındaki peptit bağları hidroliz ile kesen enzimlere denir 12. Proteazlar, aktivasyon enerjisini düşürerek, zor ve uzun sürede gerçekleşecek olan hidroliz tepkimelerini çok kısa sürede ve az enerji ile gerçekleştirmeyi sağlarlar. Proteazlar, kestikleri peptit bağlara ve aktif bölgelerine göre farklılıklar gösterirler. Serin, threonin, sistein, aspartik, metallo ve glutamik asit olmak üzere altı çeşit proteaz vardır.

Peptit bağların, proteazlar tarafından hidrolizlenme ile kesilen bağ yerlerine makas (scissile) bağ denir. Enzimin etki ettiği bileşiğe substrat denir. Proteazın aktif kısmı substrata bağlanır. Briggs 13, bir enzimin substrata nasıl bağlandığını ve onun ürüne dönüşümünü enzim kinetiği ile açıklamıştır:

$$E + S \rightleftarrows E + S \longrightarrow E + P. \qquad (2.1)$$

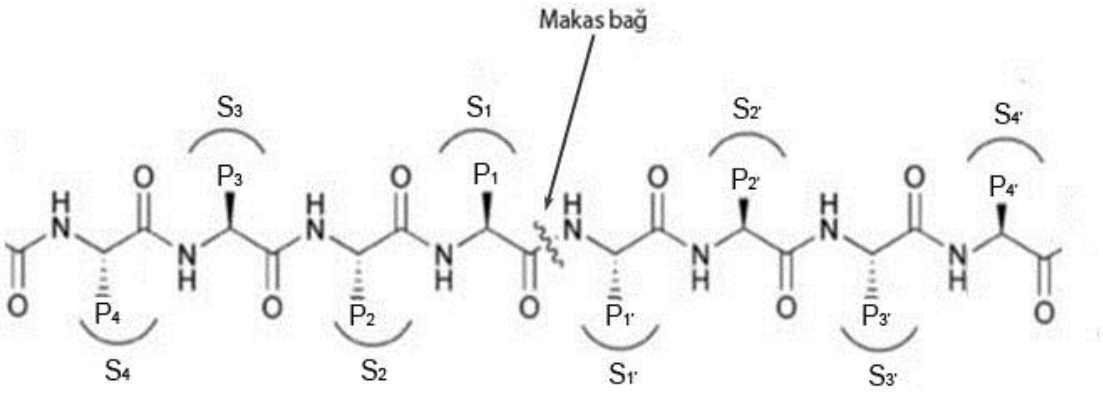
İşlevde E, enzimi; S, substratı ve P, ürünü temsil etmektedir.

1958 yılında Koshland 14, Şekil 2.5'de görülen "induced fit" model ile enzim-substrat etkileşimini açıklamıştır. Bu modele göre substrat bağlandıktan sonra proteazın aktif kısmı substratı tamamlayacak biçimi alır. Sonra proteaz tarafından kesilen substrat ürünlere dönüşür.



Şekil 2.5. Koshland'ın enzim-substrat etkileşim modeli

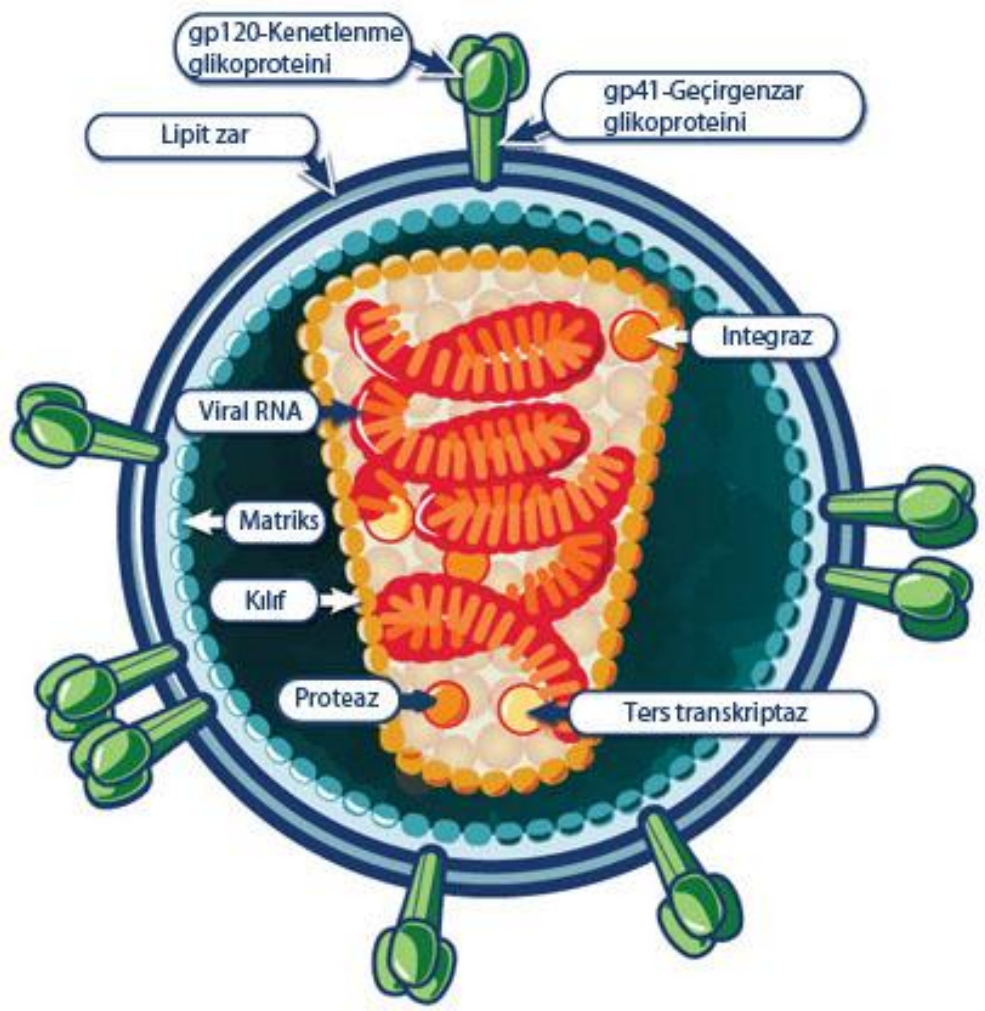
Proteaz/substrat bağlanmaları, substratın yüzey geometrisi ile proteazın aktif kısmına uyumluluğunu gerektiren özgün ve seçici bir süreçtir. Bağlanma sadece proteolotik olayların (bağlanma süreci) ilk adımıdır ve bağlanan her proteinin proteaz tarafından kesileceği anlamına gelmez. Şekil 2.5'de proteaza bağlanan bir subsratın, proteaz tarafından kesilmesi görülmektedir. N-terminal kısmındaki substrat kalıntıları P_1, P_2, P_3 ve P_4 ile, C-terminal kısmındaki substrat kalıntıları $P_{1'}$, $P_{2'}$, $P_{3'}$ ve $P_{4'}$ ile ifade edilir. İlgili amino asitlere karşılık gelen proteaz üzerindeki cepler ise S_1, $S_{1'}$, S_2, $S_{2'}$, S_3, $S_{3'}$ ve S_4, $S_{4'}$ ile belirtilmektedir. Şekil 2.6'da da görüldüğü gibi makas bağ P_1 ile $P_{1'}$ arasında meydana gelir.



Şekil 2.6. Substratın, proteaza bağlanması ve makas bağ

2.4. HIV-1'in Yapısal ve Genetik Özellikleri

Şekil 2.7'de görülen, 1/10,000 mm çapında olan HIV-1, retrovirüs ailesindendir. Retrovirüsler, genetik bilgilerini RNA formunda taşırlar 15.

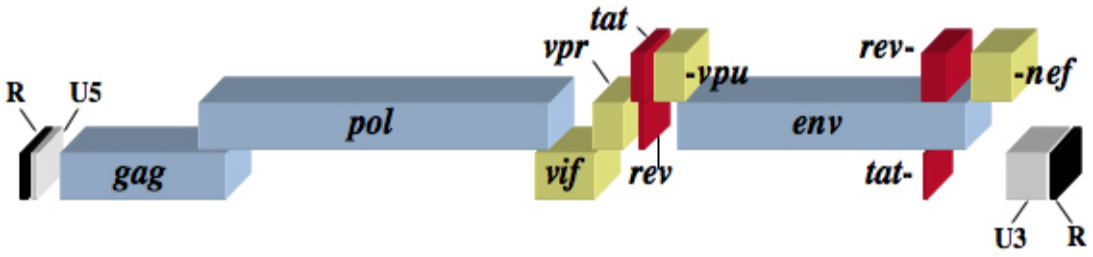


Şekil 2.7. HIV-1 virüsünün yapısı 16

Şekil 2.7'de görüldüğü gibi, genetik yapıyla birlikte proteaz (p9), ters transkriptaz & RNase H (p66), integraz gibi enzimler yapısal proteinlerden oluşan kılıf (p24) altında tek katmanlı bir tabaka içinde saklanırlar. Bu katmanın çevresinde ek yapısal proteinlerden oluşan bir matriks (p17) protein zarı bulunur. Virüsün en dışında

yukarıda saydığımız tüm yapısal elementleri kaplayan ve enfekte edilen hücrelerden alınan yağsı maddelerden oluşmuş bir lipit zar bulunur. Yağsı bu zarın üstünden dışarı doğru gelişmiş olan sayısız zarf glikoprotein uçları bulunmaktadır. Bu uçlar virüsün enfekte edeceği hücreye çeşitli hücre yüzey reseptörleri ile bağlanma işlemini kolaylaştırır. Zarf likoproteinleri iki ayrı kısımdan oluşur: kenetlenme glikoproteinleri (gp120) ve geçirgenzar glikoproteinleri (gp41) 16. Parantez içerisinde ifade edilen sayılar, proteinlerin dalton cinsinden yaklaşık atomik ağırlıklarını belirtirler.

Virüsün hücreye girişi (füzyonu), gelişmesi ve çoğalması için zarf proteinleri ile birlikte yapısal ve enzimatik proteinlere de ihtiyaç vardır. Şekil 2.8'de görüldüğü üzere virüs genomunda 15 protein env, gag ve pol genlerinde kodlanmış halde bulunur. Tablo 2.3'de HIV-1 genomuna ait genler ve kodladıkları proteinler görülmektedir.



Şekil 2.8. HIV-1 genomu[2]

[2] http://www.mimo.unige.ch/images/Genome.jpg

Tablo 2.3. HIV-1 Proteinleri 17

Gen	Protein
Gag	Matriks
	Kılıf
	Nükleokılıf
	p6
Pol	Proteaz
	Integraz
	Ters transkriptaz
Env	gp120
	gp41
Düzenleyici proteinler	Vif - Viral enfeksiyon faktör proteini
	Vpr - Viral Protein R
	Tat – Transaktivatör
	Rev - Viral protein regülatörü
	Vpu - Viral Protein U
	Nef - Negatif faktör proteini

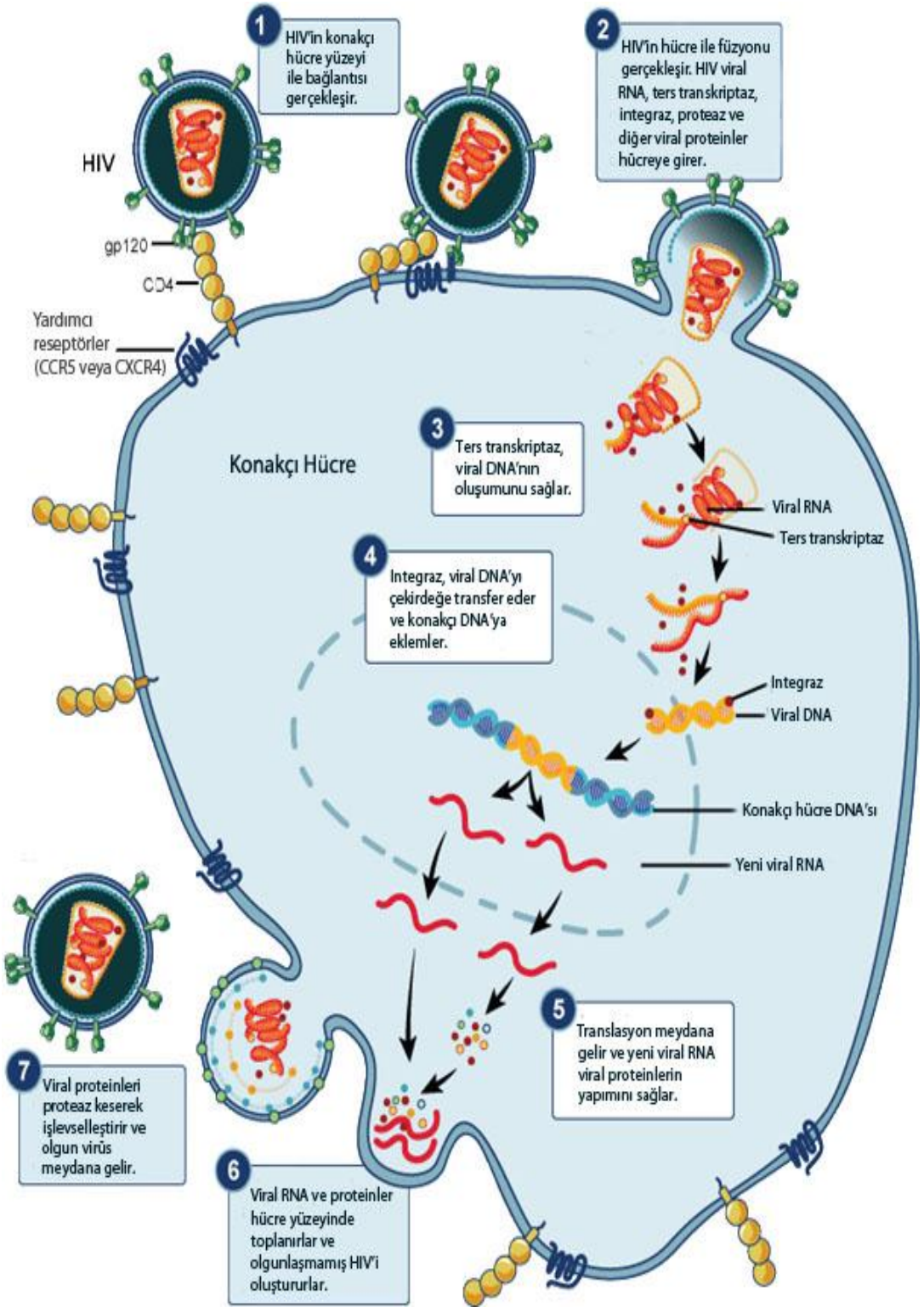
2.5. HIV-1'in Konakçı Hücrede Gelişimi

HIV, çok yüksek çoğalım kapasitesine sahiptir. Örneğin bir AIDS hastasının vücudunda her gün yaklaşık 10 milyar virüs çoğalabilir. Virüsün girdiği hücreler belli bir süre sonra ölmeye başlar. Ortalama olarak her 6 saatte bir enfekte olan hücrelerin sayısı yarı yarıya azalır. HIV-1, $CD4^+$ yardımcı T hücreleri ve makrofaj gibi bağışıklık sistemi hücrelerine bulaşır. Şekil 2.9 daki adım 1'de görüldüğü gibi ilk olarak konakçı hücreye yaklaşan HIV-1'in yüzey gliko proteini gp120, konakçı hücrenin CD4 reseptörleri ile etkileşime girer ve bağlanır. Başarılı bir füzyon, gp120 glikoproteinlerinin CD4 reseptörlerine bağlanmasının yanısıra, hastalığın ilk safhalarında makrofaj hücrelerinde CCR5, hastalığın ilerleyen safhalarında $CD4^+$ yardımcı T hücrelerinde CXCR4 yardımcı reseptörleri ile etkileşimine de bağlıdır 18. CCR5, virüsün hücreye girişi için son derece önemlidir. Çünkü herhangi bir nedenle CCR5 proteini yoksa veya mutasyona uğramışsa, virüs vücuda girse bile hücrelerin içine giremez ve dolayısıyla AIDS'e neden olamaz 19. Virüs hücre zarında dışa bakan CD4 ve CCR5 moleküllerine bağlandıktan sonra yapısal bir değişime uğrar ve

virüsün kabuğu ile hücre zarı arasında füzyon gerçekleşir. Diğer bir ifade ile virüsün kabuğu hücre zarının bir parçası haline gelir. Bu arada Şekil 2.9'daki adım 2'de görüldüğü gibi virüsün genetik malzemesi hücre sitoplazmasına aktarılır.

Bir sonraki adımda, bu genetik malzemelerden olan ters transkriptaz enzimi, ters transkripsiyonu (TT) meydana getirir. TT'de, Şekil 2.9'daki adım 3'de görüldüğü üzere, ters transkriptaz enzimi, virüsün tek iplikli viral RNA olan genomunu çift iplikli DNA'ya dönüştürür. Adım 4'de, Viral DNA, integraz enzimi tarafından konakçı hücrenin çekirdeğine transfer edilir ve insan DNA'sına eklemlenir 20. Böylece hücre kendi DNA'sı ile virüs DNA'sı arasındaki farkı algılayamaz ve adım 5'de görüldüğü gibi kendi DNA'sının mRNA'ya kodlanan proteinleri ürettiği gibi virüs DNA'sının viral RNA aracılığı ile kodlanan Tablo 2.3'de belirtilen proteinleri de üretmeye başlar. Böylece virüs genleri, yeni virüsleri oluşturacak molekülleri üretir.

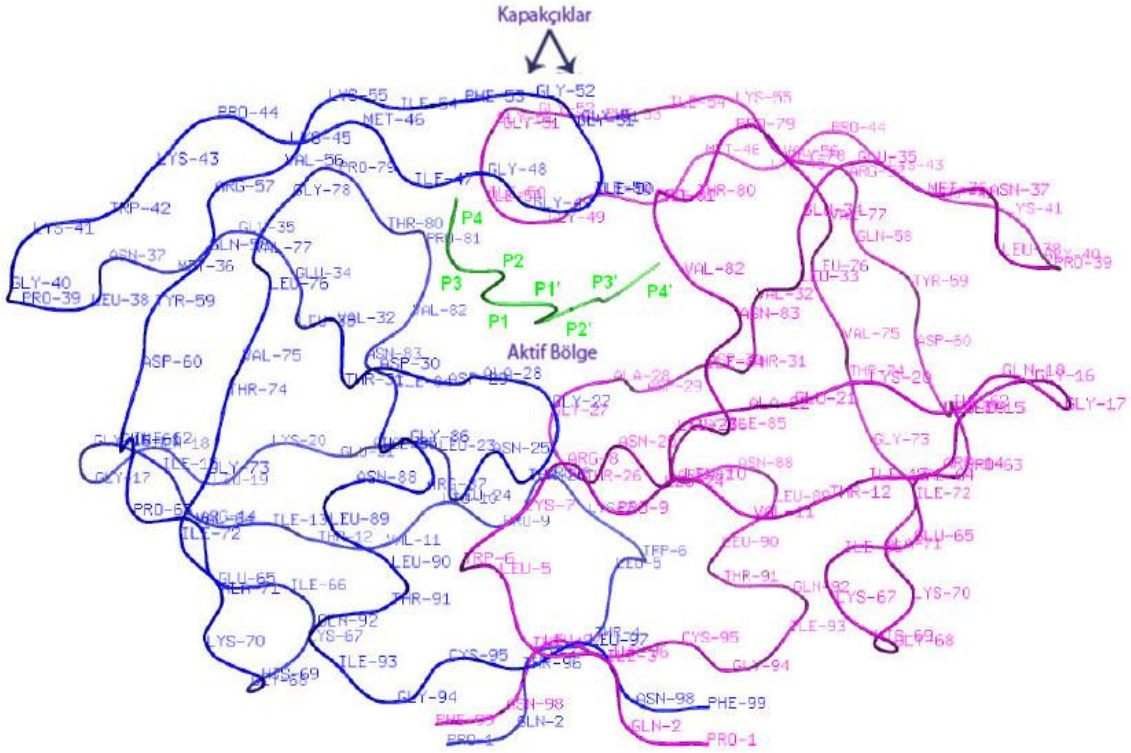
Adım 6'da görüldüğü gibi translasyon süreci sonunda sentezlenen proteinler, daha sonra yeni bir virüsü oluşturacak şekilde bir araya gelir ve hücre zarına doğru ilerler. Bir araya gelen viral proteinler hücre zarından dışarı çıkarken zardan bir parçayı da koparırlar ve beraberlerinde götürürler 21. Böylece koparılan bu parça hücreyi terk eder etmez virüsün dış yüzeyini oluşturan kabuğa dönüşür. Son olarak Adım 7'de görüldüğü gibi proteaz enzimi uzun protein dizilimlerini keserek işlevsel ve yapısal protein dizilimleri oluşmasını sağlar. Proteaz enzimi, kesme işlemini gerçekleştirmezse virüs olgunlaşamaz ve enfekte olabilme kabiliyetini kaybeder 22. HIV, genetik malzemesi çok küçük olmasına rağmen olağanüstü bir karmaşıklıkla yeni virüsü oluşturacak proteinleri ortaya çıkarır. HIV'nin karmaşık yapısına ve yaptıklarına bakınca onun diğer retrovirüslerden daha gelişmiş ve bir bakıma daha akıllı olduğunu söylemek mümkündür. Bu gerçek de HIV'nin yeni bir virüs olduğuna işaret etmektedir 19.



Şekil 2.9. HIV-1 yaşam döngüsü 23

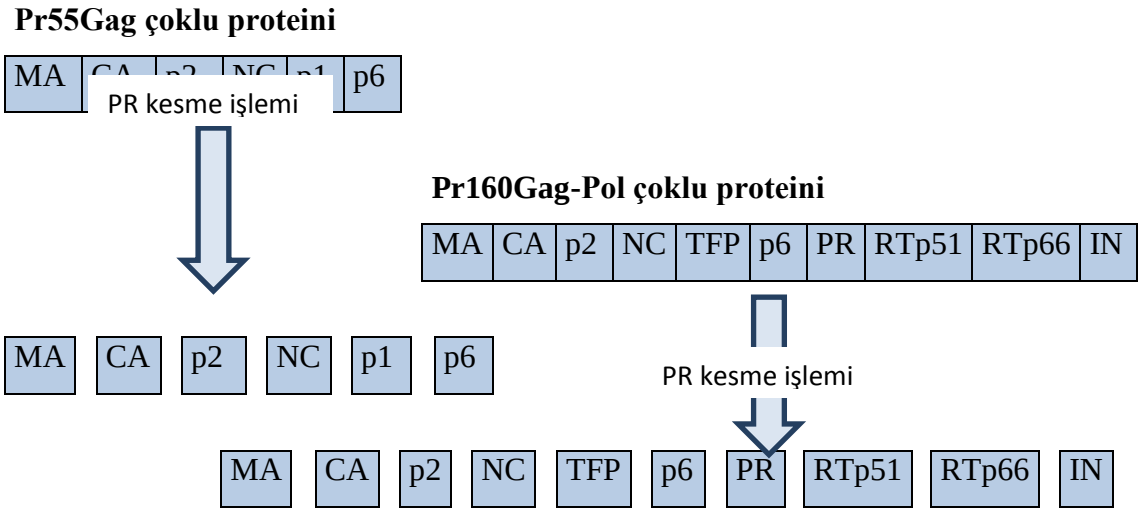
2.6. HIV-1 Proteaz Enzimi

HIV-1 proteaz, HIV-1'in yaşam döngüsü için hayati önem taşıyan aspartik bir enzimdir. Proteaz, Şekil 2.10'da görüldüğü gibi her biri 99 amino asitten meydana gelmiş birbirine özdeş iki amino asit zincirinden oluşan bir homodimer'dir 22.



Şekil 2.10. HIV-1 proteazın substrat ile homodimeri oluşturan kalıntı etiketleri görünümü 24

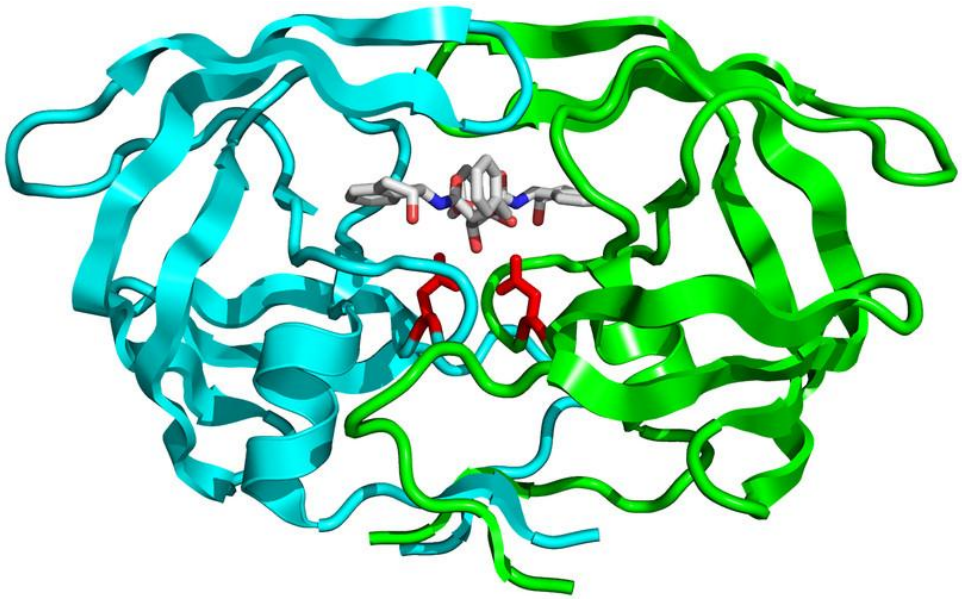
Translasyon sırasında Gag, Gag-Pol ve Env uzun çoklu proteinleri sentezlenir. HIV-1 proteaz, Şekil 2.11'de görüldüğü gibi Gag ve Gag-Pol çoklu proteinlerini keserek yapısal ve işlevsel proteinlere dönüşümlerini sağlar 22. Böylece virüs olgunlaşma evresini tamamlar.



Şekil 2.11. HIV-1 proteazın Gag ve Gag-Pol çoklu proteinlerini kesmesi

Şekil 2.11'de görüldüğü gibi proteaz, dimer arayüzlerinin oluşturduğu bir aktif bölge (katalitik bölge) ile iki esnek kapakçığa sahiptir. Aktif bölge tarafından tutulan substrat, belirli bir mekanizmaya göre sekizli peptitler halinde kesilir (cleavage) veya kesilmez (noncleavage) 25. Kapakçıklar çok esnek yapıdadırlar ve kataliz boyunca substrat veya baskılayıcı ilaç üzerine katlanırlar 26.

Şekil 2.12'de proteazın turkuaz ve yeşil renklerle gösterilen her bir monomeri kimyasal olarak özdeş ve birbirine simetriktir. Aktif bölgede bulunan iki aspartil kalıntı, kırmızı ile baskılayıcı ilaç, renkli tüpler (karbon atomu beyaz, azot atomu mavi ve oksijen atomu kırmızı) ile gösterilmektedir.



Şekil 2.12. HIV-1 proteazın baskılayıcı bir ilaç ile yapısının görünümü[3]

Aktif bölgede bulunan kalıntıların mutasyona uğraması veya baskılayıcı ilaçların bağlanması ile proteazın katalitik aktifliği durdurulabilir ve böylece virüsün enfekte olabilme kabiliyeti bloke edilebilir. HIV proteaz baskılayıcı ilaçlar peptidomimetik'tir. Başka bir ifade ile proteaz baskılayıcı ilaç proteazın kestiği kısımları taklit eder ancak kimyasal yapısı değiştirildiğinden aktif bölgede makas (scissile) bağlar proteaz tarafından kesilemez 27. Böylece aktif bölge tıkanır ve proteaz görevini yapamadığı için virüs olgunlaşma şansı bulamaz.

1987 yılında, the U.S. Food and Drug Administration (FDA) tarafından onaylanan ilk baskılayıcı TT ilacı olan azidothymidine'den beri FDA yedi TT ve sekiz proteaz baskılayıcı ilacı onaylamıştır. Günümüzde AIDS hastalığının tedavisinde kullanılan bu proteaz ilaçları Saquinavir, Amprenavir, Indinavir, Nelfinavir, Ritonavir, Atazanavir, Lopinavir ve Tipranavir'dir. TT ve proteaz baskılayıcı ilaçlarının birleşiminden oluşan tedaviler AIDS hastalarının yaşam sürelerinin uzamasına büyük katkı sağlamıştır 28. Bununla beraber, uzun süreli ilaç tedavilerinde HIV-1, baskılayıcı ilaçlara karşı mutasyon geçirerek direnç varyasyonları geliştirmektedir 29. Bu durum, HIV-1'e karşı kesin bir ilaç tedavisinin önündeki en büyük engeldir. Sonuç olarak günümüzdeki HIV ilaçları, tam manasıyla proteazın substrat

[3] PVB veri tabanından 1EBY tanımlama numarası ile elde edilmiştir.

özgünlüğünün karmaşık yapısına hâkim değillerdir. Bu nedenle proteaz özgünlüğünün işlevini tam olarak çözmeye yönelik yapılacak sistemli, daha açıklayıcı ve etkili çalışmalar mutasyondan daha az etkilenecek baskılayıcı ilaçların geliştirilmesine ön ayak olabilecektir.

HIV-1 proteaz, peptitlerde kesme işlemini herhangi bir motifsel ve basit bir yönteme göre yapmamaktadır, karmaşık bir yöntem uygulamaktadır. Kesme yerlerinin labaratuvar ortamlarında tespiti oldukça zor ve zaman alıcıdır. Bu nedenle bilgisayar ortamında HIV-1 proteaz kesme konumlarının tespitinde yapay zekâ tekniklerinden faydalanılması gerekmektedir. Makine öğrenmesi algoritmaları için giriş olarak kullanılacak peptit örüntülerinin, yeni bir öznitelik çıkarım yöntemi ile kodlanması ve böylece yüksek doğruluk oranları ile proteazın kesme konumları tespit edilebilir.

2.6.1. HIV-1 proteaz enzimi/substrat etkileşimi

HIV-1 proteaz, aktif kısmından substrata bağlanır ve etkileşime girer. Bu etkileşimde proteaz bağlandığı protein dizilimlerini sekizli peptitler halinde keser veya kesmez. 1980'lerden beri HIV üzerine çalışmalar yapılsa da hala HIV-1 proteaz/substrat etkileşimine dair kısıtlı bilgi bulunmaktadır. Peptitlerin substrat bilgisini öğrenmek için birkaç yol vardır. Birincisi, laboratuvar ortamında her bir peptit test edilerek kesilmiş peptit mi, kesilmemiş peptit mi olduğu öğrenilebilir. Diğer bir yöntem ise denatüre proteinlerden faydalanmaktır 30. Denatüre olmuş proteinler, ortam şartlarından kaynaklı (ısı, üre vb.) ikincil ve üçüncül yapıları bozulmuş ama peptit bağları hala sağlam olan proteinlerdir. Denatüre proteinler, HIV-1 proteaz kesim konumlarını tespit etmek için laboratuvar testlerine tabi tutulurlar. Eğer denatüre proteinde kesilmiş peptite rastlanmazsa, tüm dizilimler kesilmemiş kabul edilirler. Kayan pencere yöntemi ile tüm dizilimler, negatif örüntü örnekleri olarak elde edilirler. Kaydırma değişmezliği (shift invariance) olarak adlandırılan bu yöntemde, kesilmiş peptit olmadığı tespit edilen denatüre proteinde bir kalıntı, aralıklarla sağa veya sola doğru kaydırma yapılarak kesilmemiş peptitler belirlenir 25.

Denatüre proteinlerdeki kesilmiş kısımlar, doğal proteinlere göre farklılıklar gösterebilmektedirler. Örneğin bovin serum albümin proteini doğal ortamında HIV-1

proteaz enzimi aktivitesine direnç gösterirken denatüre olduğunda bu direnç ortadan kalkmaktadır 30. Aslında bu durum proteinin yapısının, proteaz enziminin işlevine etkisi olduğunun göstergesidir.

BÖLÜM 3. ÖRÜNTÜ TANIMA SİSTEMLERİ

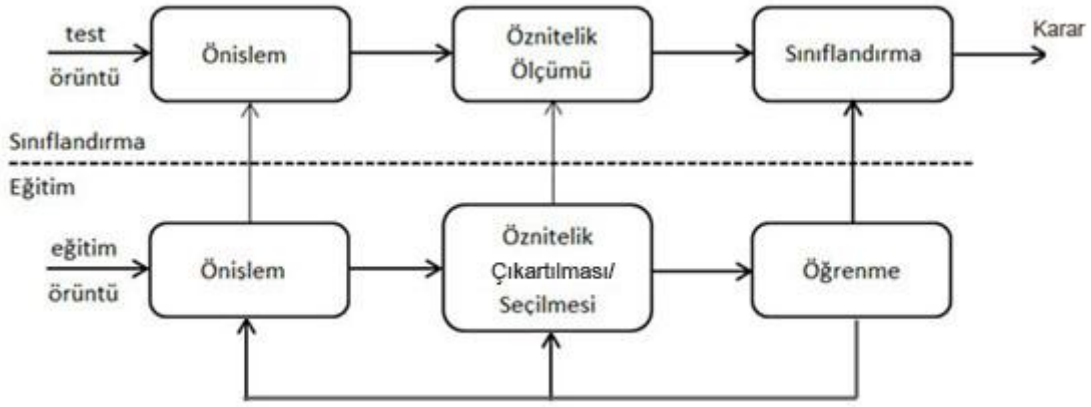
Örüntü tanıma, nesneleri değişik sayıda sınıflara veya kategorilere ayırma amacına dayalı bir bilim dalıdır. Sınıflandırılmak istenen nesneler biyomedikal verilerden, elektromanyetik işaret dalgalarına kadar çeşitlilik gösterebilir. Bu nesneler genel bir ifade ile örüntü olarak adlandırılır. Örüntü tanıma, mühendislik araştırma, geliştirme konularında en uygun kararın verilmesinde önemli bir rol oynamaktadır.

Günümüzde, biyomedikalde kalp ve beyin işaretlerinin incelenmesi, endüstriyel kontrolde robot kontrolü ve otonom cihazların kullanımı, haberleşmede uydu görüntülerinin işlenmesi, biyometride yüz, parmak izi tanıma sistemleri gibi bilim alanlarında örüntü tanıma uygulamaları etkin olarak uygulanmaktadır.

Teknolojinin ilerlemesine paralel olarak biyobilişim alanında da canlı bedeninde (in vivo) yapılması güç olan uygulamalar, bilgisayar destekli olarak (in silico) gerçekleştirilebilmektedir. Bu tezin konusu olan HIV-1 proteaz enziminin protein kesme konumlarının tespit probleminin çözümü de bilgisayar destekli olarak çalışılmıştır.

Şekil 3.1'deki örüntü tanıma sisteminde görüldüğü gibi tanıma süreci iki kısımdan oluşur: eğitim ve sınıflandırma. Önişlem kısmının amacı, örüntünün temsilini sonraki adımlar için daha anlaşılır ve işlenebilir hale getirmektir. Bu amaç doğrultusunda örneğin biyobilişimde protein verileri üzerine araştırma yapılıyorsa örüntü verisi normalize edilerek bir kodlamaya tabi tutulabilir (Öznitelik kodlama yöntemi) veya elektronikte işaret teknikleri üzerine araştırma yapılıyorsa işaretler üzerinde gürültü azaltma işlemleri uygulanabilir. Tüm bu işlemlerde, örüntünün uygun bir şekilde temsili amaçlanmaktadır 31.

Bir öğrenme modellemesi, sınıflandırma (classification) veya bağlanım (regression) problemi olabilir. Bağlanım analizi, iki veya daha çok değişken arasındaki ilişkiyi ölçmek için kullanılan istatistiki bir yöntemdir. Sınıflandırma ise örüntü tanıma sürecine dahil edilen giriş verilerinin, süreç sonunda, tanımlanmış olan sınıflardan hangisine ait olduğunun tahmin edildiği istatistiki bir yöntemdir. Tez çalışmasının amacı açısından söylemler ve çıkarımlar, sınıflandırma metodolojisi üzerine yapılacaktır.



Şekil 3.1. Örüntü tanıma sisteminin genel diyagramı 32

Öznitelik (feature), örüntüye ait ölçülebilir veya gözlenebilir bilgi olarak tanımlanabilir. Eğitim sürecinde, öznitelik çıkartılması/seçilmesi kısmında temsil edilen örüntü verileri için en uygun öznitelikler tespit edilir ve sınıflandırıcı öznitelik uzayını bu yönde çeşitli bölümlere ayırır. Geri besleme ile önişlem ve öznitelik çıkartılması/seçilmesi stratejilerinin optimize edilmesi sağlanır. Sınıflandırma kısmında, eğitilmiş sınıflandırıcı, giriş örüntülerini öznitelik ölçümlerine göre hangi sınıflara ait olduğuna karar verir. Unutmamak gerekir ki makine öğrenmesi ile gerçekleştirilen örüntü tanıma modellemelerinde amaç, deneysel gözlemlerin (eğitim verileri) tam olarak temsilini öğrenmek değil, temel (underlying) fonksiyonu üretebilmek ve eğitim verilerinden farklı yeni örnekler (test verileri) üzerinde başarılı biçimde genelleme yapabilmesini sağlamaktır.

3.1. Genelleme

Sınıflandırıcının eğitim seti üzerindeki başarımını artırma yolu ile optimizasyon yapma her zaman test seti üzerinde istenen performansı vermeyebilir. Bir sınıflandırıcının genelleme yeteneği, o sınıflandırıcının test seti üzerindeki performansını belirtir. Zayıf bir genelleme yeteneği sebepleri için aşağıdaki nedenler sayılabilir 32:

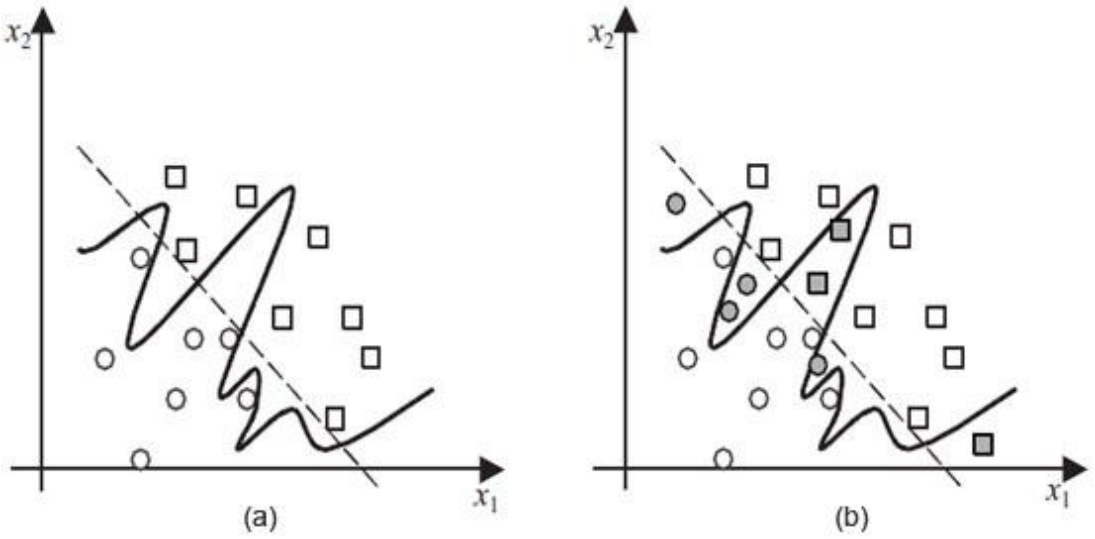
- Sınıflandırıcının eğitim veri seti üzerinde aşırı derecede eğitilmesi (Aşırı öğrenme – Overtraining veya overfitting),
- Öznitelik sayısının, eğitim veri örneklerine göre aşırı büyük olması (Boyut problemi – Curse of dimensionality).

3.1.1. Aşırı öğrenme

Eğitim verilerini öğrenmek bir sınıflandırıcı için zor değildir. Ama bunun anlamı, sınıflandırıcının başarılı bir temel fonksiyon çıkaracağı ve test verileri üzerinde başarılı bir genelleme yapacağı anlamına gelmez. İyi bir genelleme performansı, eğitim verilerini optimum şekilde, aşırı ve yetersiz öğrenme olmaksızın öğrenmeyi ve test verileri üzerinde düşük riski gerektirir. Test verileri üzerinde risk, eğitilmiş bir model için test hata beklentisidir ve umulan risk (expected risk), $R(\alpha)$, olarak adlandırılır. Eğitim verileri üzerindeki risk ise deneysel risk (empirical risk), $R_{emp}(\alpha)$ olarak adlandırılır 33. Öğrenmenin amacı, en düşük umulan riski elde etmektir. Eğitim veri sayısı sonsuza ulaştığında, $\lim_{N\to\infty} R_{emp}(\alpha) = R(\alpha)$ olur. Bununla beraber sonlu bir eğitim veri setinde model, eğitim setine kısmen uyar. Bu nedenle $R_{emp}(\alpha) < R(\alpha)$ olarak kabul edilir.

Şekil 3.2'de daire ve karelerle temsil edilen iki sınıfın doğrusal olarak ayrılabildiği basit bir sınıflandırma örneği görülmektedir. Şekil 3.2a'da, 2 boyutlu bir uzayda örüntüler hem doğrusal bir yaklaşımla (kesik çizgiler) hem de yüksek kapasiteye sahip, doğrusal olmayan bir yaklaşımla (sürekli eğri çizgisi) hatasız risk değeri, $R_{emp}(\alpha) = 0$, ile eğitilmiştir. Şekil 3.2b'de, düzleme gri daire ve karelerden

oluşan test verileri yerleştirilmesi sonrasında her iki doğrusal ve doğrusal olmayan sınıflandırıcıların genelleme performansları görülmektedir.



Şekil 3.2. Doğrusal olarak ayrılabilen bir örüntü problemi üzerinde aşırı öğrenme 34

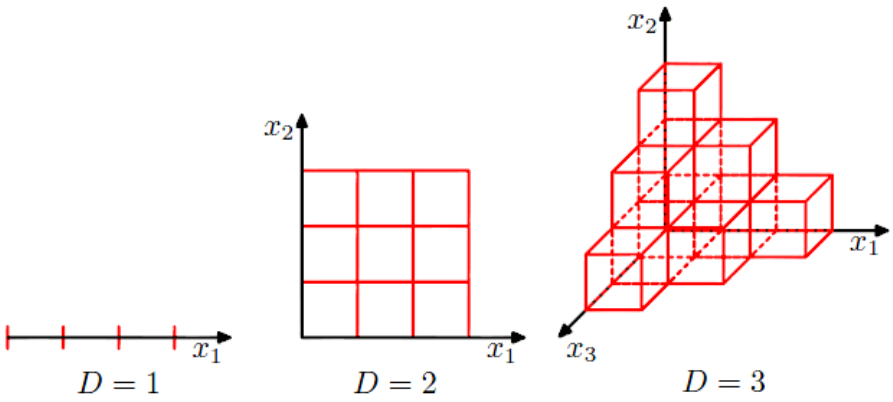
Doğrusal modelleme hatasız bir genelleme sonucu verirken doğrusal olmayan modelleme ise kötü bir genelleme performansı sergilemektedir. Bunun nedeni doğrusal olmayan modellemenin eğitim verilerini aşırı öğrenmesidir.

3.1.2. Boyut problemi

Örüntü tanıma sistemlerinde genellemenin istenen performansı verememesinin nedenlerinden biri olan boyut problemi, sabit sayıda eğitim verilerine karşılık gelen öznitelik sayısının aşırı fazla olmasından kaynaklanır. Bu durumda seyrek sayıda öznitelikler vektörleri ile temsil edilen eğitim verileri, sınıflandırıcının veriminin düşmesine neden olurlar.

Boyut problemi, giriş uzayının hücrelerle temsil edildiği bir örnekle açıklanabilir 35. Şekil 3.3'de görüldüğü gibi hücre sayısı, boyut arttıkça üstel olarak hızla artmaktadır. Örüntülerin, doğrusal olmayan, gereksiz yere karmaşık bir karar fonksiyonu ile sınıflandırıldığı varsayılsın. Bu durumda her bir hücreyi doldurabilmek için üssel sayıda artan eğitim verisine ihtiyaç vardır. Bu kadar eğitim verisi ise mümkün

değildir. Eldeki kısıtlı sayıdaki veri ile her bir eğitim verisi bir hücreye gelecek şekilde yerleştirildiğinde ise yetersiz bir öğrenme meydana gelir.



Şekil 3.3. Boyut problemi 35

Eğer eksen boyunca hücre sayısı eğitim veri sayısı oranında artırılırsa kararlılık da artar. Fakat hücre boyutu artışı üssel olursa tüm hücrelerin dolması için bu oranda da eğitim veri sayısı artırmalıdır. Aksi halde sabit veya yetersiz eğitim veri sayısı ve üssel olarak artan boyut, modelin iyi temsil edilememesine ve buna bağlı olarak genellemenin performansının gerilemesine neden olur.

3.2. Sınıflandırma

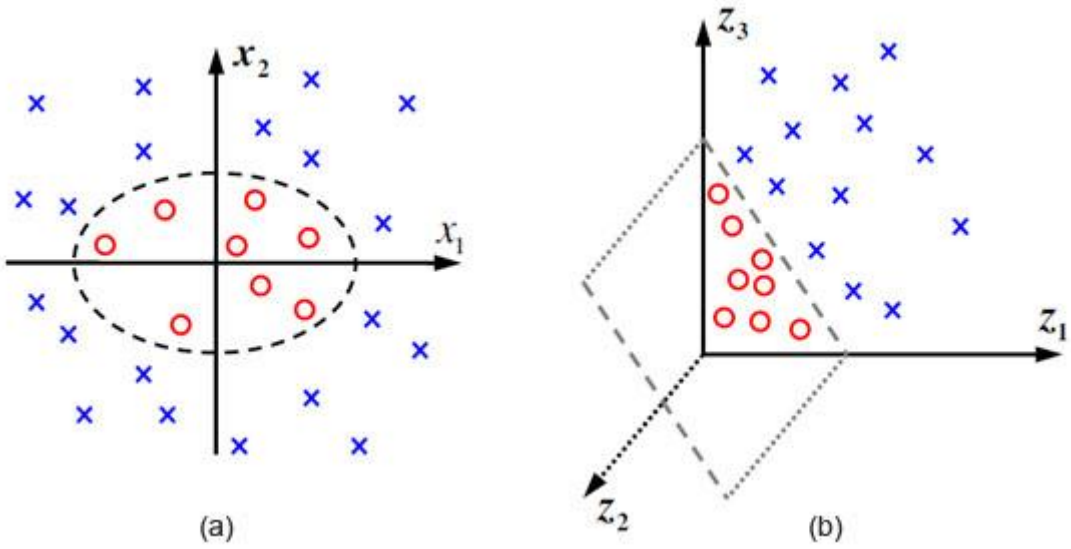
Örüntü tanımanın eğitim sürecinde, öznitelik çıkarımından sonra karar verme aşaması olan sınıflandırma aşaması başlar. Bu aşamada, nesnelerin sınıflandırılması için bir karar fonksiyonu elde edilir ve bu fonksiyona göre nesneler sınıflandırılır. N adet setten oluşan bir deneysel gözlem kümesi için $D = \{x_i, y_i\}_{i=1,2,\ldots,N}$ ve $x_i \in \Re^d$. Kümede bulunan tüm deneysel veriler birbirinden bağımsız ve özdeş dağılımlara sahiptirler. Sınıflandırma sonucunda her bir x_i'nin sınıfına karar verilir. Burada y_i, karşılık gelen x_i örüntüsünün sınıfını bildiren etikettir. Örneğin, HIV-1 proteaz kesme kısımlarının sınıflandırması ele alındığında, x_i gözlemlenmiş peptit dizilimlerini, y_i ise karşılık gelen sınıf etiketlerini temsil eder. Eğer ilgili peptit üzerinde kesme işlemi gerçekleşmiş ise $y_i = +1$, gerçekleşmemiş ise $y_i = -1$ olur.

Sınıflandırmada iki tür yöntem kullanılır 36: üretici (generative) yöntem, ayırıcı (discriminative) yöntem. Üretici modelde, sınıflandırıcı ortak olasılık dağılımı (joint probability distribution) modelini öğrenir ve Bayes kuralı ile $p=(y|x)$'i mümkün tüm y'ler içinde en yüksek olasılığa sahip y'yi doğru sınıf olarak alır. Üretici modellemede, tüm değişkenler için olasılıklara dayalı tahmin süreci gerekir. Buna karşın, ayırıcı modellemede ise gözlemlenmiş değişkenlerden hedef değişkenlerin olasılığı hesaplanır. Ayırıcı öğrenme, gözlemlenen değişkenlerin dağılımının modellenmesine ihtiyaç duyulmadığı için gözlemlenen ve hedef değişkenler arasında daha karmaşık ilişkileri ifade edebilir. Bu nedenle ayırıcı öğrenme modellerinde sınıflandırma ve bağlanım analizinde üretici modellere nazaran daha iyi performans elde edilir 36. Ayırıcı modeller, doğrusal ve doğrusal olmayan yöntemler olmak üzere ikiye ayrılır.

25 ve 54'de HIV-1 proteaz enziminin çoklu proteinleri kesme konumlarını tespiti probleminin doğrusal bir problem olduğu ve çözüm için en uygun sınıflandırıcının üretici bir model olan doğrusal DVM olduğu belirtilmiştir.

3.2.1. Destek vektör makineleri

DVM, 1979 yılında Vapnik tarafından geliştirilmiştir. DVM, eğitim örneklerinin, bir üst düzlem (hyperplane) ile doğrusal olarak ayrılabilmeleri üzerine kurulu bir makine öğrenmesi yöntemidir. Eğitim için kullanılacak N örüntüden oluşan verinin, $D=\{x_i, y_i\}_{i=1,2,...,N}$, olduğunu varsayalım. Burada $x_i \in \Re^d$ eğitim örnekleri ve $y_i \in \{-1,+1\}$ etiket değerleridir. Doğrusal olarak ayrılabilme durumunda, iki sınıfa ayrılabilen örüntüler direkt olarak bulundukları orijinal uzayda bir üst düzlem ile ayrılabilirler. Doğrusal DVM'nin amacı ayırıcı üst düzlemin iki eğitim sınıfına eşit uzaklıkta olmasını sağlayarak eğitim örneklerini ayırmaktır. Eğer eğitim örüntüleri giriş uzayında doğrusal olarak bir üst düzlem ile ayrılamıyorlarsa, Şekil 3.4a'da doğrusal olarak ayrılamayan eğitim örnekleri giriş uzayında görülmektedir. DVM, Şekil 3.4b'de görüldüğü gibi bu eğitim örneklerine ait öznitelikler vektörlerini, yüksek boyutlu bir öznitelikler uzayına taşıyarak, bir üst düzlem ile doğrusal olarak iki sınıfa ayrılabilmelerini sağlamaktadır.



Şekil 3.4. Öznitelikler vektörlerinin, ayırıcı bir üst düzlem ile ayrılması 33

DVM, orijinal uzaydan öznitelikler uzayına dönüşümü, doğrusal olmayan haritalama (polinom, sigmoid vb.) ile yapar. Bir sonraki adımda, en uygun doğrusal ayırıcı üst düzlem bulunur 34.

Doğrusal olarak ayrılabilen, (3.1)'de belirtildiği gibi eğitim örnekleri olsun.

$$(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \ldots, (\mathrm{x}_m, y_m), \qquad \mathrm{x} \in \Re^2, \;\; y \in \{+1, -1\}. \tag{3.1}$$

Verileri, iki sınıfa ayırabilen birden fazla üst düzlem bulunmaktadır. Bu üst düzlemlerden deneysel riski en aza indiren, en yüksek sınır boşluğuna (marjin) sahip olanı seçilir. Küçük sınır boşluğuna sahip olan bir sınıflandırıcı daha yüksek değerde umulan riske neden olabilir.

Örüntü tanıma sistemi öğrenme sürecinde, (3.2)'de görülen karar fonksiyonunun ($d(\mathrm{x},\mathrm{w},b)$), ağırlık ($\mathrm{w} = [w_1 w_2 \ldots w_n]^T$) ve bias ($b$) parametrelerini bulur.

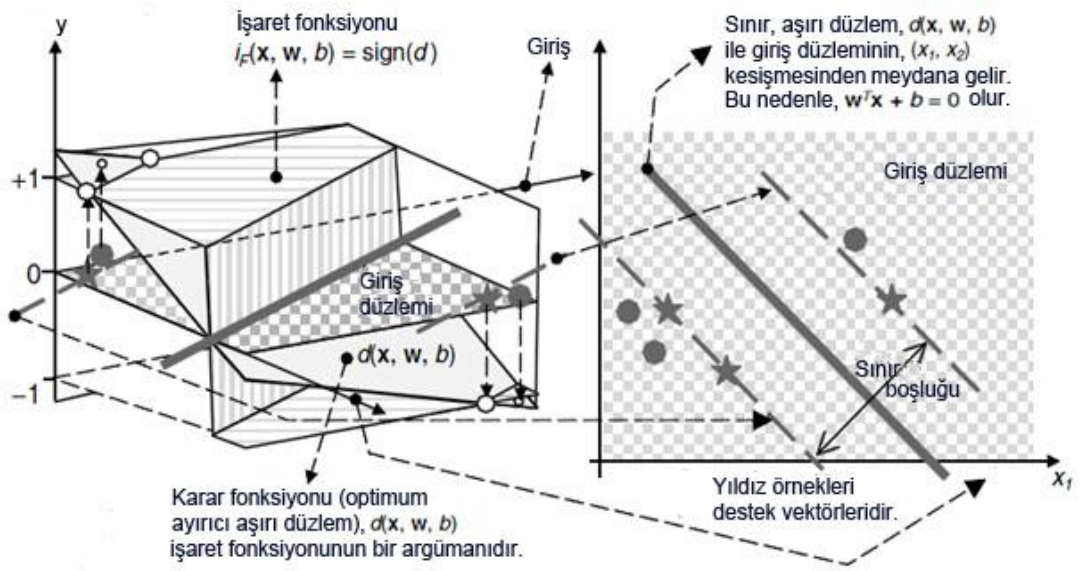
$$d(\mathrm{x},\mathrm{w},b) = \mathrm{w}^T\mathrm{x} + b = \sum_{i=1}^{n} w_i x_i + b, \qquad \mathrm{x},\mathrm{w} \in \Re^n. \tag{3.2}$$

Başarılı bir eğitim sürecinden sonra, öğrenme makinesi, (3.3)'de görülen karar vermeyi sağlayan,

$$i_F = sign(d(x_p, w, b)), \tag{3.3}$$

işaret fonksiyonuna göre Şekil 3.5'de görüldüğü gibi sınıflandırma yapar. Diğer bir ifade ile üst düzlemi ifade eden karar fonksiyonunun, $d(\mathrm{x_p},\mathrm{w},b)$ iki kuralı vardır:

- Eğer $d(\mathrm{x_p},\mathrm{w},b) > 0$ ise örüntü $\mathrm{x_p}$, sınıf 1'e ($y_1 = +1$) aittir.
- Eğer $d(\mathrm{x_p},\mathrm{w},b) < 0$ ise örüntü $\mathrm{x_p}$, sınıf 2'ye ($y_1 = -1$) aittir.



Şekil 3.5. Ayırıcı üst düzlem ($d(\mathrm{x},\mathrm{w},b)$), sınır ($d(\mathrm{x},\mathrm{w},b)=0$) ve işaret fonksiyonunun ($\mathrm{sign}(d(\mathrm{x_p},\mathrm{w},b))$) tanımlanması 34

Sınıflandırmada, giriş vektörleri (x) ile aynı uzayda bulunan ayırma sınırı, vektörleri iki sınıfa ayırır. Şekil 3.5'de görüldüğü gibi bu sınır, karar fonksiyonu ($d(\mathrm{x},\mathrm{w},b)$) ile giriş öznitelikler uzayının kesişmesinden oluşur. Sınırın matematiksel tanımlaması,

$$d(\mathrm{x},\mathrm{w},b) = 0, \tag{3.4}$$

ile ifade edilir.

Şekil 3.5'de görülen yıldız örüntüleri ile temsil edilen destek vektörleri için hem i_F ve hem de d'nin değeri |1|'e eşittir. Diğer tüm eğitim örüntüleri için |d|>| İ_F|'dir.

Eğitim örüntülerini ayıran birden fazla üst düzlem vardır. DVM'nin altında yatan öğrenme teorisinin temel amacı, en büyük sınır boşluğuna sahip ayırıcı üst düzlemi bulmaktır. Böylece yeni verileri daha iyi sınıflandırabilir. Sınır boşluğu (S), (3.5)'de görüldüğü gibi,

$$S = \frac{2}{\| w \|}, \tag{3.5}$$

işlevi ile hesaplanır. S'nin en büyük olabilmesi, ||w|| değerinin en küçük olması ile mümkündür. (3.6)'da görüldüğü gibi öğrenme problemi,

$$\min \frac{1}{2} \| w \|^2, \tag{3.6}$$

işlevi ile hesaplanır. Bu fonksiyon en küçüklenmek istenen fonksiyon, (3.7) ise çözümün sağlaması gereken kısıtlardır.

$$y_i[w^T \mathrm{x}_i + b] \geq 1, \qquad i = 1, 2, ..., m. \tag{3.7}$$

Dolayısıyla problem ikinci dereceden sınırlamalı bir optimizasyon problemidir. Problemin çözümü için (3.8)'de görüldüğü gibi Lagrange formülasyonu,

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i \left(w^T x_i + b \right) + \sum_{i=1}^{N} \alpha_i, \tag{3.8}$$

yapılarak Lagrange çarpanlar hesaplanır. Bu formülasyonda $\alpha_i \geq 0$ değerleri pozitif Lagrange çarpanlardır. (2.7)'de ifade edilen formülasyonun çözümü için (3.9) ve (3.10)'da gösterilen Karush Kuhn Tucker (KKT) kısıtları,

$$\frac{\partial L_P}{\partial w} = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i, \tag{3.9}$$

$$\frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_i \alpha_i y_i = 0, \tag{3.10}$$

kullanılarak formülasyon ikili probleme dönüştürülür. KKT kısıtları (3.8)'de yerlerine yazılırsa,

$$\begin{aligned} L_d &= \frac{1}{2}(w^T w) - w^T \sum_i \alpha_i y_i + \sum_i \alpha_i, \\ &= -\frac{1}{2}(w^T w) + \sum_i \alpha_i, \\ &= \sum_i \alpha_i - \frac{1}{2}\sum_i \alpha_i \alpha_j y_i y_j x_i^T x_j, \end{aligned} \tag{3.11}$$

formülasyonu elde edilir. (3.11)'de görülen bu optimizasyon probleminin çözümü için gerekli olan koşullar (3.12)'de ifade edilmektedir.

$$\alpha_i \geq 0, \forall i. \tag{3.12}$$

Çözümden elde edilecek Lagrange çarpanlarının çoğunluğunun değeri sıfır olacaktır. Geriye kalan $\alpha_i > 0$ değerli x_i örnekleri, destek vektörleridir.

3.3. Öznitelik Çıkartılması

Örüntü modellemede öznitelik çıkartılması aşamasının amacı, örüntülere ait öznitelikler uzayının boyutu indirgenirken veya normalize edilmiş eğitim örneklerinin varyansı en büyüklenirken ayırıcı bilgilerini olabildiğince korumaktır 37. Öznitelik çıkartımı işlemi, örüntü modellemede zaman ve bilgisayar hafızasından tasarruf sağlamanın yanı sıra boyut problemi etkisini de iyileştirir. En yaygın doğrusal öznitelik çıkartımı yöntemleri TBA ve DAA'dır.

3.3.1. Temel bileşenler analizi

Karhunen–Loeve dönüşümü olarak da bilinen TBA, 1901 yılında Pearson tarafından geliştirilmiş, örüntü tanımada en sık kullanılan boyut indirgeme tekniklerindendir. TBA'nın dayandığı temel fikir özniteliklerin varyansıdır. Bu doğrultuda, öznitelik uzayının varyansı ne kadar yüksekse, o öznitelik uzayı o kadar çok bilgi taşır. Bu nedenle TBA, örüntü verilerinin varyansını olabildiğince yüksek, boyutunu ise olabildiğince düşük olacak şekilde doğrusal dönüşüme tabi tutar 38. Aralarında bağımlılık bulunan değişkenler, TBA uygulandıktan sonra, doğrusal, birbirine dik ve birbirinden bağımsız yeni değişkenlere dönüşürler.

TBA yöntemi üç adımda gerçekleşir 37:

Adım 1: Eğitim örnekleri, x_i, i = 1,2,...,N örnek veriler ve her bir x_i örneği için $x_i = [x_i^1 x_i^2 \cdots x_i^M]^T$ olmak üzere (3.13)'de görüldüğü gibi önce ortalaması (μ),

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{3.13}$$

ile hesaplanır. Sonra ortalama (3.14)'da görüldüğü gibi her bir x_i eğitim örneğinden çıkarılarak,

$$\overline{x}_i = x_i - \mu_i, \tag{3.14}$$

(3.15)'da görülen ortalama matrisi,

$$\overline{x} = [\overline{x}_1 \overline{x}_2 \cdots \overline{x}_N]_{MxN}, \tag{3.15}$$

elde edilir. Bu aşamada ortalama matrisi normalize edilebilir.

Adım 2: $\overline{x}$ matrisi, (3.16)'da görüldüğü gibi devriği (transpozu) ile çarpılarak kovaryans matrisi,

$$Cov(\overline{xx}^T), \tag{3.16}$$

ile hesaplanır. Kovaryans matrise ait özdeğerler ve bu özdeğerlere karşılık gelen özvektörler hesaplanır. Bulunan bu özvektörler aranan temel bileşenlerdir. En yüksek n özdeğere karşılık gelen n özvektör seçilir. Bu seçim kriterinde n sayısı, örneğin seçilen özvektörlerin toplamı tüm özvektörlerin toplamının % 95'ini oluşturacak şekilde belirlenebilir. Bu durumda sütun özvektörleri olan $W = [w_1, w_2, \cdots w_n]$ matrisi elde edilir.

Adım 3: Yeni öznitelik vektörleri (3.17)'de görüldüğü gibi,

$$\tilde{y}_i = W^T \overline{x}_i, \tag{3.17}$$

ile bulunur. Dönüşümü yapılan yeni uzayda, her bir yeni örnek vektörü n adet girişe sahip olur. Böylece orijinal örnek uzayı (*d*), *n*'ye indirgenir.

3.3.2. Doğrusal ayırıcı analiz

DAA, örüntü modelleme de sınıfların ayırımını sağlayan üst düzlemin belirlenmesi üzerine kurulu analiz yöntemidir. Fisher, 1936 yılında sınıfların ayrılmasını metadolojik olarak açıklayan DAA'yı, diğer adı ile Fisher'in ayırıcı analizi yöntemini geliştirmiştir. DAA, aynı sınıfa ait örüntüler arasındaki dağılımı olabildiğince en küçük, sınıflar arasındaki dağılımı olabildiğince en büyük yapmayı hedefler 39. Böylece verilerin dağılımı daha iyi anlaşılabilir.

Kittler ve Devijver 40, ayırıcı bilginin örüntünün dağılımı ile ilgili olduğunu, sınıf dağılımları aynı veya birbirine yakın ise DAA'nın ayırıcı bir düzlemle ayırma da başarısız olduğunu belirtmişlerdir.

DAA ve TBA'nın temel farkı, DAA'nın verileri sınıflandırırken, TBA'nın öznitelik sınıflandırması yapmasıdır. DAA'da orijinal örüntülerin konumu değiştirilmeksizin

örüntüler arasında ayırım yapılır. Fakat TBA'da örüntüler öznitelik uzayına indirgendiğinde hem şekilleri hem de konumları değişir.

DAA yönteminde ilk olarak, dönüştürülen uzayda sınıflar arası dağılım (D_a) ve sınıf içi dağılım (D_b) hesaplanır 41. D_a, (3.18)'de görüldüğü gibi,

$$D_a = \sum_{i=1}^{s} N_i (m_i - m)(m_i - m)^T, \tag{3.18}$$

fonksiyonu ile hesaplanır. Bu fonksiyonda N_i, S_i *(i=1,2,..,s)* sınıfının örüntü sayısıdır. Ayrıca m_i, S_i sınıfına ait örüntülerin, m ise tüm sınıflarda bulunan örüntülerin ortalamasıdır. D_b, (3.19)'da görüldüğü gibi,

$$D_b = \sum_{i=1}^{s} \sum\nolimits_{k \in si} (x_k - m_i)(x_k - m_i)^T, \tag{3.19}$$

fonksiyonu ile hesaplanır. Üçüncü adımda (3.20)'de görüldüğü gibi toplam dağılım matrisi yani tüm örneklerin kovaryans matrisi,

$$D_t = D_a + D_b = \sum_{i=1}^{N} (x_i - m)(x_i - m)^T, \tag{3.20}$$

hesaplanır. Burada N tüm sınıflarda bulunan örüntü sayısıdır. DAA, (3.21)'de görülen Fisher ölçüt fonksiyonu,

$$J_{DAA}(W_{opt}) = \max \frac{|D_a|}{|D_b|} = \max \frac{|W^T D_a W|}{|W^T D_b W|}, \tag{3.21}$$

çerçevesinde sütun özvektörleri, $W = [w_1, w_2, \cdots w_{s-1}]$ matrisini hesaplar.

(3.21)'de |A| notasyonu A matrisinin determinantını temsil etmektedir. Hesaplanan determinant değişkenleri ve elde edilen özvektörler kullanılarak, örüntü örneklerinin hangi sınıfa ait oldukları belirlenir.

3.3.3. Normalizasyon

Normalizasyon, örüntü tanıma sisteminin performansını artırmak amacıyla, veri setinin bulunduğu uzaydan başka bir uzaya taşınması işlemidir. Bu taşıma işleminde veri setinin sınır noktaları (en büyük ve en küçük nokta değerleri) değişmekle beraber boyutunda değişme olmaz. Yapılan normalizasyon makine öğrenmesi algoritmasının transfer karakteristiğine uygun ise elde edilen sonuçlar daha başarılı olur 42. Bu nedenle veriler üzerinde ne tür normalizasyonun uygulanacağı çözülmek istenen problem ve kullanılan makine öğrenmesi yöntemi ile ilgilidir. Literatürde yaygın olarak kullanılan üç tür normalizasyon yöntemi bulunmaktadır. Bu yöntemler; minumum-maksimum, ondalık ölçme ve z-skor yöntemleridir. Testlerde AAindex'te yer alan fizikokimyasal özelliklere ait değerler z-skor yöntemi kullanılarak normalize edilmiştir.

İstatistiki değerlendirilmelerde sıklıkla kullanılan z-skor yöntemi, verilerin ortalama değeri ile bu değerlerin ortalamadan olan uzaklıkları ile ilgili yeni değerler üretir 43. Değerlerin normalize edilebilmesi için önce (3.22)'da görüldüğü gibi ortalama değer (μ),

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{3.22}$$

sonra (3.23)'de görüldüğü gibi standart sapma (σ),

$$\sigma = \left(\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2 \right)^{1/2}, \tag{3.23}$$

işlevi ile hesaplanır. Burada x_i, i. veriyi, n veri sayısını temsil etmektedir.
Son aşamada (3.24)'de görüldüğü gibi verilen değerler z-skor değerlerine çevrilir.

$$x_i' = \frac{x_i - \mu}{\sigma}, \tag{3.24}$$

(3.24)'de x_i', x_i değerinin normalize edilmiş halidir.

BÖLÜM 4. ÖZNİTELİK ÇIKARIMI

Özniteliğin çıkarımı yani temsili doğru sınıflandırma açısından önemlidir. Çünkü örüntüye ait karakteristik özellikler ne kadar net temsil edilirse sınıflandırıcı örüntüyü o kadar iyi tanır. Öznitelik çıkarımı, Şekil 3.1'de görülen örüntü tanıma sisteminde, önişlem adımında gerçekleştirilir. Eğer bir örüntü birden fazla öznitelik ile temsil ediliyorsa, tek bir öznitelik yerine bir "öznitelik kümesi" söz konusudur. n adet özniteliğe sahip öznitelik kümesi ise n-boyutlu "öznitelik vektörü" ile temsil edilir. Özniteliklerin içinde bulunduğu n-boyutlu $\mathbb{R}^n$ uzayı ise "öznitelik uzayı" olarak isimlendirilir 44.

Proteinleri oluşturan amino asitlerin, sayısal olarak temsil edilmesi örüntü modellemenin bir ayağıdır. Proteinlerin temsilinde, amino asitlerin protein içindeki yeri, sayısı veya fizikokimyasal özellikleri, öznitelik vektörler kümesi olarak ifade edilebilir.

HIV-1 proteaz enziminin kesilme konumlarının tespiti problemi, protein dizilimlerinin proteaz enzimi tarafından makas bağ yerlerinden kesilmesi veya kesilmemesi durumlarının tespit edildiği bir sınıflandırma problemidir. HIV-1 proteaz enzimi, protein dizilimlerini sekiz kalıntıdan oluşan peptitler halinde keser. Bir peptit, P = $P_4P_3P_2P_1{\downarrow}P_{1'}P_{2'}P_{3'}P_{4'}$ ile ifade edilir. Burada her P_i bir amino asittir. ↓ işareti makas bağı temsil etmektedir. Diğer bir ifade ile kesme işlemi makas bağın olduğu, P_1 ile $P_{1'}$ arasında meydana gelir 25. P'nin kesilme konumlarının tespitinin tahmini için oluşturulan modelde, öncelikle P örüntülerinin öznitelik çıkarım yöntemleri ile kodlanması gerekmektedir.

4.1. Birimdik Öznitelik Kodlama Yöntemi

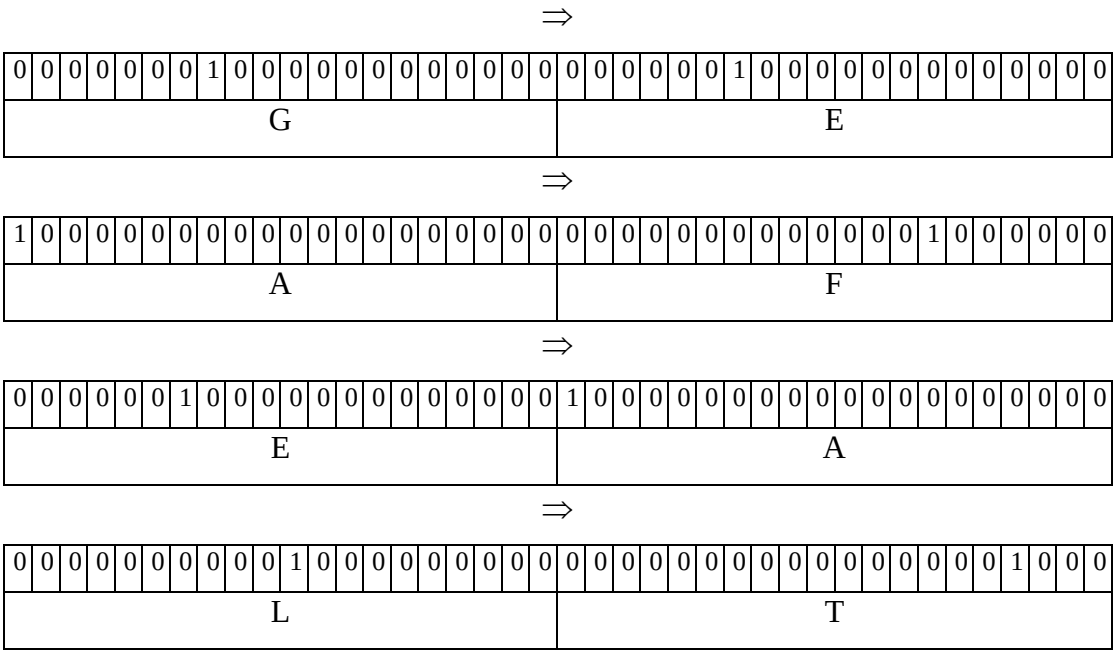
HIV-1 proteaz enzimi bölünme kısımlarının tespitinde birimdik vektörlerle (birbirine dik birim vektör – orthonormal vector) kodlama (BKY), en sık uygulanan yöntemlerdendir 45, 46. BKY, P peptitini oluşturan her bir P_i amino asit sembolü, birbirine dik, $d_i = (\delta_{i1}, \delta_{i2}, ..., \delta_{i20})$ vektörlerle ile ifade edilir. Burada δ_{ij} Kronecker delta sembolüdür. Tüm öznitelik vektörleri birbirine diktir.

Tablo 4.1. Amino asitlerin standart BKY ile temsil edilmeleri

Sıra	Amino Asitler	20-bit Vektör																			
1	A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
8	G	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	H	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
10	I	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
11	L	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
12	K	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
13	M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
14	F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
15	P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
16	S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
17	T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
18	W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
19	Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
20	V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Tablo 4.1'de görüldüğü gibi, BKY'de her bir kalıntı 20 bit uzunluğunda vektör ile temsil edilir. Bu temsilde, her bir kalıntının sırasına karşılık gelen bit, 1 ile geri kalan değerler ise 0 ile temsil edilir 30,47. Böylece her bir peptit dizilimi 1x160 bit büyüklüğünde birimdik vektörlerle temsil edilir.

GEAFEALT şeklinde verilen örnek peptit için BKY'yi açıklayalım. Peptitin ilk kalıntısı G'nin BKY öznitelik vektörü Tablo 4.1'de görüldüğü üzere [00000001000000000000]'dir. Her bir kalıntı için öznitelik vektörleri soldan sağa doğru Şekil 4.1'de görüldüğü gibi yerleştirilir.



Şekil 4.1. GEAFEALT peptit diziliminin BKY ile kodlanması

BKY'nin başlıca sakıncası, temsil ettiği peptit içindeki amino asitleri sadece ikili sayı sisteminde ifade eder ve dolayısıyla peptitin içeriğine dair herhangi bir bilgi taşımaz. Diğer bir ifade ile BKY, amino asitlerin birbirleri ile olan fizikokimyasal benzerlikleri veya farklılıkları hakkında bilgiler içermez. Bu durum verinin içeriğinin tam olarak temsil edilememesine neden olur.

BKY'nin ikinci sakıncası ise kısıtlı veri üzerinde, boyut problemine neden olabilecek büyük boyut gerektirmesidir.

BKY'nin son sakıncası ise mümkün olan tüm sekizli peptit sayılarına ulaşılsa bile, $20^8 << 2^{160}$ olacağından dolayı öz nitelik uzayı yine de aşırı derecede seyrek kalacaktır 25.

4.2. Ağırlık Tabanlı Öznitelik Kodlama Yöntemi

Ağırlık Tabanlı Öznitelik Kodlama Yöntemi (ATKY), peptitte bulunan kalıntıların ağırlıkları üzerine kurulu bir öznitelik çıkarım yöntemidir. Yöntemin uygulanmasında ilk olarak peptit dizilimini oluşturan her bir amino asitin ağırlığı (w_i),

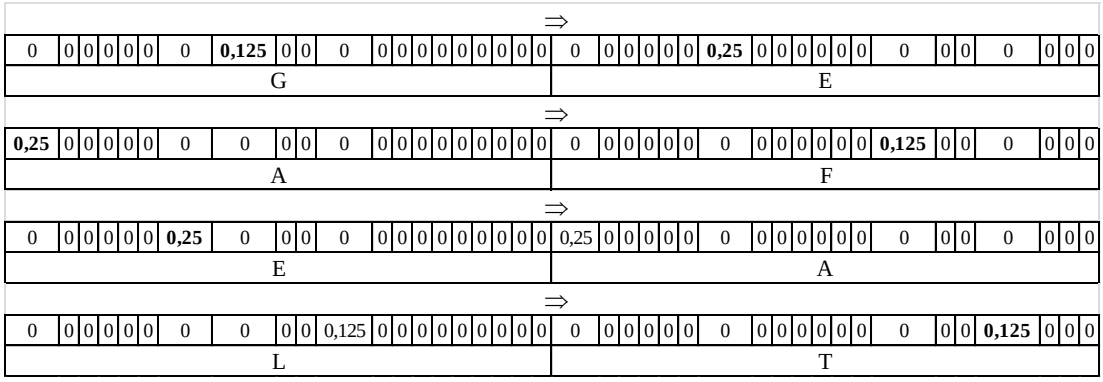
$$w_i = \frac{P_T}{N}, \tag{4.1}$$

hesaplanır. Burada P_T, P_i kalıntısının peptit içindeki sayısı, N ise peptitte bulunan kalıntı sayısıdır. Sonraki adımda BKY ile ATKY birleştirilir. Buna göre peptit içinde yer alan her bir amino asit ağırlığı ile amino aside karşılık gelen birimdik vektörü çarpılır. Böylece her bir peptit dizilimi, 1x160 bit büyüklüğünde vektörlerle temsil edilir 48.

GEAFEALT şeklinde verilen örnek peptit için ATKY'yi açıklayalım. İlk adımda (4.2)'deki gibi GEAFEALT peptitindeki kalıntıların ağırlıkları hesaplanır. Örneğin E amino asidi için ağırlık,

$$w_E = \frac{2}{8} = 0,25, \tag{4.2}$$

olarak hesaplanır. Bu işlem dizilim içindeki diğer amino asitler için de tekrarlanır. İkinci adımda, elde edilen kalıntı ağırlık değerleri, Şekil 4.2'de görüldüğü gibi BKY öznitelik vektöründe karşılık gelen ilgili kalıntı öznitelik vektörü ile çarpılır.



Şekil 4.2. GEAFEALT peptit diziliminin ATKY ile kodlanması

ATKY'nın üstünlüğü, amino asitlerin peptit içindeki ağırlıkları bakımından niceliklerini temsil etmesidir. Bununla beraber BKY'de olduğu gibi amino asitlerin birbirleri ile olan fizikokimyasal benzerlikleri veya farklılıkları hakkında bilgiler içermez. Örneğin, farklı dizilimlere sahip ancak benzer fizikokimyasal özellikler içeren polimerlerin, farklı fizikokimyasal özellikler içeren polimerlere göre öznitelikler uzayında birbirine yakın olması beklenir. Ama bu tür bilgileri ATKY içermez. Ayrıca ATKY'de kısıtlı veri üzerinde, boyut problemine neden olabilecek büyük boyut gereksinimi doğabilir.

4.3. Yer Değiştirme Matrisleri Tabanlı Öznitelik Kodlama Yöntemleri

Eşit iki uzunluktaki protein dizilimlerinin karşılaştırılmalarında ve özellikle protein hizalama problemlerinde, yer değiştirme matrislerine sıkça başvurulur. Yer değiştirme matrislerindeki sayılar, amino asitlerin birbirleri yerine geçme eğilimlerine dair bilgi verir. İki amino asit arasındaki yer değiştirme sayısının göreceli büyük olması demek birbirleri ile olan göreceli benzerliklerinin artması demektir. Bunun anlamı, bu iki amino asidin tabiatta işlevsel ve yapısal olarak benzer oldukları anlamına gelir. Yer değiştirme matrisleri, proteinlerin işlevsellikleri ile dizi hizalamaları arasındaki korelasyonları gözlemlenerek elde edilmişlerdir 49.
Her ne kadar yer değiştirme matrisleri protein dizilimlerinin hizalanması problemleri için çok faydalı olsa da dizilime özgün, belirli yerlere ait yer değiştirmelerin

benzerliklerini ve etkilerini belirlemede yetersiz kalmaktadırlar 50. HIV-1 proteaz enzimi kesme konumları da kendine özgün, belirli konumlardır.
Burada asıl problem, hangi yer değiştirme matrisinin kullanılacağıdır. Çünkü AAindex veritabanında 94 tane yer değiştirme matrisi bulunmaktadır. PAM, BLOSUM olmak üzere çeşitlendirilebilen yer değiştirme matrisler çok farklılıklar içermektedirler.

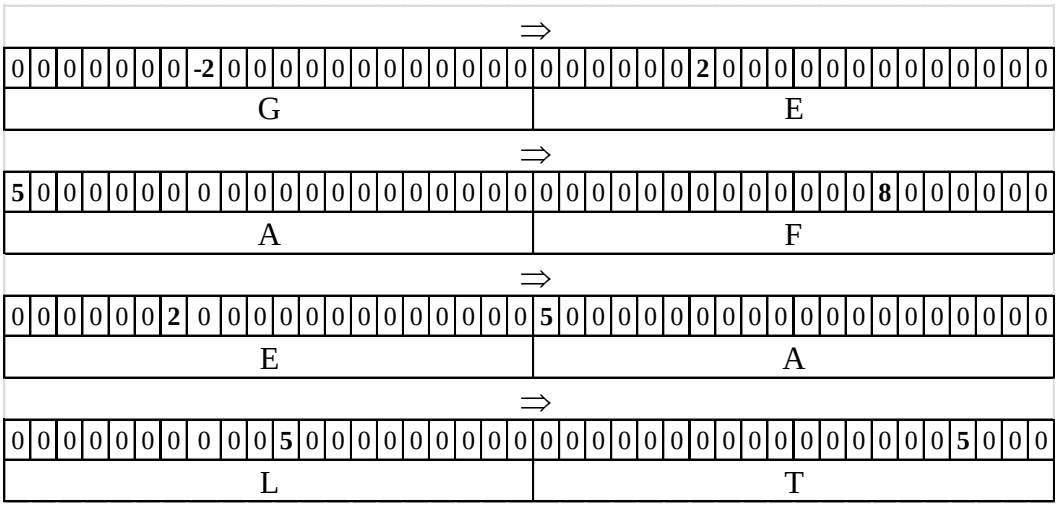
Yer değiştirme matrislerinde P_i ve P_j amino asitlerinin birbirlerine benzerlik oranı, i. satır ve j. sütunun kesişimi ile elde edilir. Yer değiştirme matrisi ile yapılan kodlamada önce P peptiti birimdik olarak kodlanır. Sonra her bir P_i amino asidinin yer değiştirme matrisi içindeki i. satır ve i. sütundaki değeri ile d_i birimdik vektörü çarpılır 51.

GEAFEALT şeklinde verilen örnek peptit üzerinde Şekil 4.2'de görülen BLOSUM50 yer değiştirme matrisine göre öznitelik kodlamasını açıklayalım. İlk adımda GEAFEALT peptitindeki her bir kalıntının kendi satır ve sütun kesişim değerleri (K_i, $i=1,2,...,8$) elde edilir. Buna göre K_G = -2, K_E = 2, K_A = 5, K_F = 8, K_Q = 7, K_T = 5 olur.

Tablo 4.2. BLOSUM50 amino asit yer değiştirme matrisi

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	0	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	0	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-4	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

İkinci adımda elde edilen K_i değerleri, Şekil 4.3'de görüldüğü gibi BKY öznitelik vektöründe karşılık gelen ilgili kalıntı öznitelik vektörü ile çarpılır.



Şekil 4.3. GEAFEALT peptit diziliminin BLOSUM50 yerdeğiştirme matrisine göre kodlanması

4.4. n-grams Öznitelik Kodlama Yöntemi

n-grams öznitelik kodlama yöntemi, kalıntı çiftlerinin sıklığını tespit üzerine kurulu bir öznitelik temsili yöntemidir. n-grams yönteminde P peptit diziliminde ardışık n tane kalıntı sayısı aranır. Bu sayılar n ardışık kalıntı çiftinin, üçlüsünün vb. meydana gelme sıklığını verir. Bu durumda öznitelik sayısı 20^n olur 52.

GEAFEALT şeklinde verilen örnek peptit üzerinde n = 2'ye göre n-grams öznitelik kodlamasını açıklayalım. Bu durumda 20^2 uzunlukta bir öznitelik vektörü ile mevcut tüm ikili kalıntı ihtimalleri temsil edilebilir. Peptit dizilimi meydana getiren 7 adet ikili, ardışık kalıntı çiftinin sayısı elde edilir. Buna göre GE = 1, EA = 2, AF = 1, FE = 1, AL=1, LT = 1 olur. Elde edilen bu değerler oluşturulan 1x400 büyüklüğündeki öznitelik vektöründe karşılık gelen yere yerleştirilir.

n-grams kodlamanın protein diziliminden bağımsız olması önemli bir üstünlüğüdür. Proteini oluşturan kalıntı sayısı artsa bile öznitelik sayısı daima 20^n dir. n-grams yönteminin başlıca kısıtı ise kalıntıların fizikokimyasal özelliklerine dair bilgiler içermemesidir. Ayrıca boyut problemi n-grams kodlamada da karşılaşılabilecek bir sorundur.

4.5. Kalıntı Çiftleri Öznitelik Kodlama Yöntemi

Kalıntı Çiftleri Öznitelik Kodlama Yöntemi (KÇKY), n-grams kodlama yöntemi gibi proteini oluşturan amino asit çiftlerinin, dizilim sırasına bağlı sıklık bilgisi üzerine kuruludur.

Her bir peptit dizilimi için tüm kalıntı çiftleri, çift derecesine (m, coupling-degree), bağlı olarak bir öznitelik vektörüne dönüştürülür 53. Birinci derece çiftler,

$$\{\vec{x}\}_1 = \frac{1}{N-1}\sum_{i=1}^{N-1} P(i, i+1), \tag{4.3}$$

vektörü ile ifade edilir. Burada P, proteini N, protein diziliminde bulunan kalıntı sayısını ifade eder. İkinci derece çiftler,

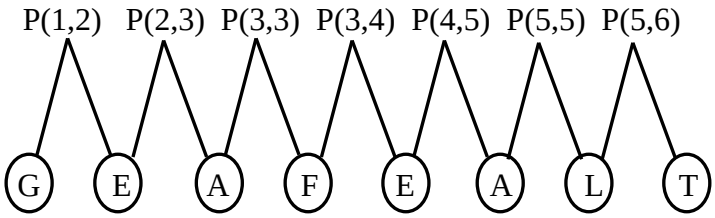
$$\{\vec{x}\}_2 = \frac{1}{N-2}\sum_{i=1}^{N-2} P(i, i+2), \tag{4.4}$$

vektörü ile ifade edilir. m. derece ise,

$$\{\vec{x}\}_m = \frac{1}{N-m}\sum_{i=1}^{N-m} P(i, i+m), \tag{4.5}$$

vektörü ile ifade edilir. Son aşamada bu vektörler birleştirilir.

KÇKY'yi, GEAFEALT şeklinde verilen örnek peptit üzerinde açıklayalım. 1. derece çiftler Şekil 4.4'de görülmektedir.



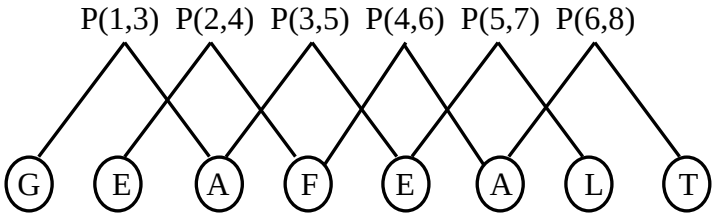
Şekil 4.4. GEAFEALT peptit dizilimi için 1. derece kalıntı çiftleri

Elde edilen kalıntı çiftlerinin dizilim içindeki sayısı, dizilimi oluşturan kalıntı sayısının bir eksiğine (*N-1*) bölünür ve kalıntı çiftinin ağırlığı (*w*) elde edilir. Örneğin GE için ağırlık (4.6)'da hesaplanmıştır.

$$w_{GE} = \frac{1}{7}1. \tag{4.6}$$

Bu işlem diğer kalıntı çiftleri için de tekrarlanır. Sonraki aşamada hesaplanan ağırlık, 1x400 büyüklüğünde 1. derece çiftler için oluşturulan öznitelikler vektöründe kalıntı çiftine karşılık gelen yere yerleştirilir.

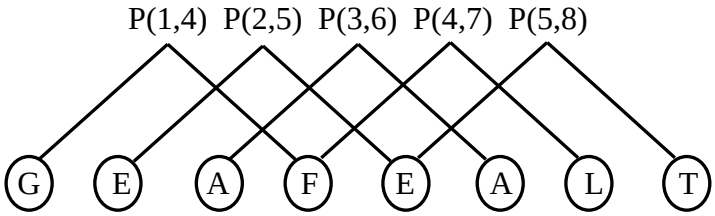
2. derece kalıntı çiftleri, Şekil 4.5'de görüldüğü gibi birer kalıntı atlanarak yapılır.



Şekil 4.5. GEAFEALT peptit dizilimi için 2. derece kalıntı çiftleri

1. derece çifler için yapılan işlem 2. derece kalıntı çiftleri için de tekrarlanır ve elde edilen ağırlıklar oluşturulan 1x400 büyüklüğünde 2. derece öznitelikler vektörüne yerleştirilir.

3. derece kalıntı çiftleri ise Şekil 4.6'da görüldüğü gibi ikişer kalıntı atlanarak oluşturulur.



Şekil 4.6. GEAFEALT peptit dizilimi için 3. derece kalıntı çiftleri

1. derece çifler için yapılan kalıntı çiftlerine ait ağırlık hesapları 3. derece kalıntı çiftleri için de tekrarlanır ve elde edilen ağırlıklar 3. derece çiftler için oluşturulan 1x400 büyüklüğünde öznitelikler vektörüne yerleştirilir. Son aşamada her derece için elde edilen üç adet 1x400 büyüklüğündeki öznitelikler vektörü birleştirilerek 1x1200 büyüklüğünde toplam öznitelikler vektörü elde edilir.

Her ne kadar KÇKY, amino asitlerin protein içindeki dizilim sırasını göz önüne alıyor olsa da fizikokimyasal özelliklerine dair bilgiyi içermemesi en büyük kısıtıdır.

Ayrıca KÇKY boyut problemine neden olabilecek bir yöntemdir. Örneğin, m = 3 için 400 x 3 = 1200 büyüklüğünde bir vektör, boyut problemini tetikleyebilir. Dolayısıyla çift derecesi arttıkça boyut probleminin meydana gelme ihtimali de yükselmektedir.

4.6. BLOMAP Öznitelik Kodlama Yöntemi

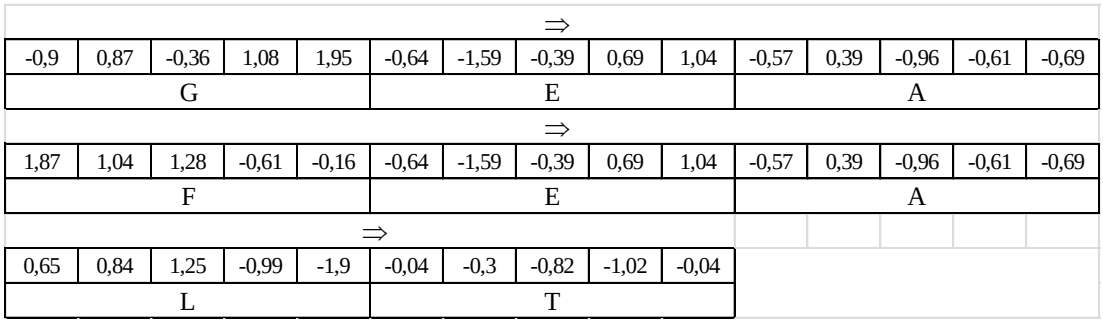
BLOMAP öznitelik vektörü kodlama yöntemi, yüksek boyutlu uzaydan göreceli düşük boyutlu uzaya haritalama yapılmasını sağlayan, doğrusal olmayan bir yöntem olan Sammon projection algoritmasının, BLOSUM 62 yer değiştirme matrisi ile geliştirilmesi üzerine kuruludur. Tablo 4.3'de BLOMAP yöntemine ait her bir amino asit için kod vektörleri görülmektedir. Kod vektörleri uzaklık bilgisinin en iyi temsil edildiği 5 boyutta ifade edilmiştir 46.

BLOMAP kodlamanın düşük boyut gerektirmesi ve BLOSUM 62 yer değiştirme matrisine bağlı olarak kalıntılar arasında benzerliklere dair yetersiz de olsa bilgi içermesi üstünlükleridir. Protein içindeki belirli yerlere ait yer değiştirmelerin benzerliklerini ve etkilerini belirlemede amino asit yer değiştirme matrislerinin yetersiz kaldığı unutulmamalıdır 50.

Tablo 4.3. BLOMAP yöntemi kod vektörleri

Sıra	Amino Asitler	BLOMAP62 kod vektörleri				
1	A	-0,57	0,39	-0,96	-0,61	-0,69
2	R	-0,4	-0,83	-0,61	1,26	-0,28
3	N	-0,7	-0,63	-1,47	1,02	1,06
4	D	-1,62	-0,52	-0,67	1,02	1,47
5	C	0,07	2,04	0,65	-1,13	-0,39
6	Q	-0,05	-1,5	-0,67	0,49	0,21
7	E	-0,64	-1,59	-0,39	0,69	1,04
8	G	-0,9	0,87	-0,36	1,08	1,95
9	H	0,73	-0,67	-0,42	1,13	0,99
10	I	0,59	0,79	1,44	-1,9	-0,93
11	L	0,65	0,84	1,25	-0,99	-1,9
12	K	-0,64	-1,19	-0,65	0,68	-0,13
13	M	0,76	0,05	0,06	-0,62	-1,59
14	F	1,87	1,04	1,28	-0,61	-0,16
15	P	-1,82	-0,63	0,32	0,03	0,68
16	S	-0,39	-0,27	-1,51	-0,25	0,31
17	T	-0,04	-0,3	-0,82	-1,02	-0,04
18	W	1,38	1,69	1,91	1,07	-0,05
19	Y	1,75	0,11	0,65	0,21	-0,41
20	V	-0,02	0,3	0,97	-1,55	-1,16

BLOMAP yöntemini, GEAFEALT şeklinde verilen örnek peptit üzerinde açıklayalım. Protein dizilimi içindeki her bir kalıntıya karşılık gelen kod vektörü Şekil 4.7'de görüldüğü gibi kalıntının proteinde karşılık gelen yerine yerleştirilir.



Şekil 4.7. GEAFEALT peptit diziliminin BLOMAP yöntemi ile kodlanması

BÖLÜM 5. FİZİKOKİMYASAL ÖZELLİKLERE GÖRE ÖZNİTELİK KODLAMA

HIV-1 proteaz enzimi, proteinlerin kesme konumlarını belirleme işlemini bir mekanizma içinde yapmaktadır. Bu kesme mekanizmasını anlamak için öncelikle proteinleri meydana getiren kalıntıların biyokimyasal etkileşimlerini anlamak gerekir. Diğer bir ifade ile kalıntıların fizikokimyasal özellikleri ne kadar doğru anlaşılırsa ait oldukları proteinin özelliklerini anlamak da o kadar açık olmaktadır. Bu nedenle HIV-1 proteaz enzimi özgünlüğünün modellenmesinde kalıntıların fizikokimyasal özellikleri, mutlaka modele aktarılmalıdır. Bu temel düşünceden yola çıkarak AAindex'te bulunan toplam 544 özellik kullanılarak FTKY geliştirilmiştir.

FTKY'nin gelişimi iki safhadır 54. Birinci safhada AAindex'de bulunan 544 fizikokimyasal özellikten sırasıyla en iyi 10, 20, 30, 40 ve 50 özellik dört adımda belirlenmiştir. İlk adımda 8 kalıntıdan oluşan her bir peptit dizilimi BKY ile kodlanarak 1x160 büyüklüğündeki öznitelik vektörleri elde edilmiştir. İkinci adımda dizilim içindeki kalıntılara ait her bir fizikokimyasal özellik değeri z-skor yöntemine göre normalize edilmiş ve ait olduğu kalıntının öznitelik vektöründe karşılık gelen sırasında 1 biti yerine yerleştirilmiştir. Bu işlem tüm eğitim örnekleri için tekrarlanmıştır. Üçüncü adımda elde edilen öznitelik vektörleri doğrusal DVM ile sınıflandırılarak her bir fizikokimyasal özellik için sınıf doğruluğu değerleri elde edilmiştir. Son adımda elde edilen sınıf doğruluğu değerleri büyükten küçüğe doğru sıralanarak en iyi 10, 20, 30, 40 ve 50 fizikokimyasal özellik belirlenmiştir.

İkinci safhada elde edilen en iyi fizikokimyasal özelliklerin indeks değerleri peptit diziliminde ait oldukları kalıntının yerine yazılarak öznitelik vektörleri elde edilmiştir. Buna göre sekiz kalıntında meydana gelen bir peptit diziliminin FTKY'ye göre öznitelik vektörlerinin büyüklükleri, en iyi 10-fk (fizikokimyasal) için 10 x 8 = 80, en iyi 20-fk için 20 x 8 = 160, en iyi 30-fk için 30 x 8 = 240, en iyi 40-fk için 40 x 8 = 320 ve en iyi 50-fk için 50 x 8 = 400 olmaktadır.

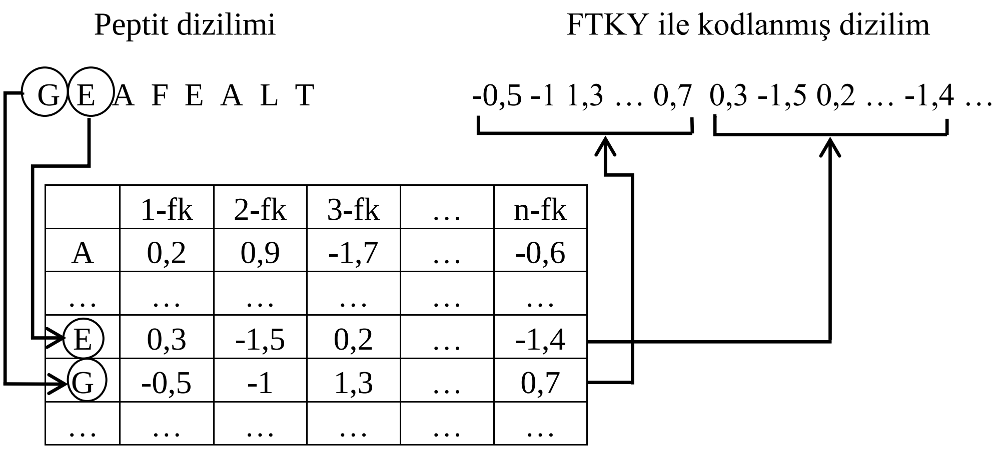
FTKY'yi, GEAFEALT şeklinde verilen örnek peptit üzerinde açıklayalım. İlk safhada en Şekil 5.1'de görüldüğü gibi tüm 544 amino asit indeks değerleri normalize edilerek sırasıyla BKY ile kodlanmış olan öznitelik vektörüne yerleştirilir.

⇒

0	0	0	0	0	0	0	**-1,75**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**-0,5**	0	0	0	0	0	0	0	0	0	0	0	0	0
G																				E																			
-0,26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0,95**	0	0	0	0	0	0
A																				F																			
0	0	0	0	0	0	**-0,5**	0	0	0	0	0	0	0	0	0	0	0	0	0	**-0,26**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E																				A																			
0	0	0	0	0	0	0	0	0	0	**-0,97**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**-0,26**	0	0	0
L																				T																			

Şekil 5.1. 544-fk özelliğin BKY özvektöründe yerleştilmesi

Her bir fizikokimyasal özelliğe göre kodlanan öznitelik vektörleri doğrusal DVM ile sınıflandırılarak sınıf doğrulukları elde edilir. Bu sınıf doğruluğu değerleri büyükten küçüğe sıralanarak en fizikokimyasal özellikler elde edilir. İkinci safhada Şekil 5.2'de görüldüğü gibi elde edilen en iyi fizikokimyasal özelliklere göre FTKY öznitelik vektörleri elde edilir.



Şekil 5.2. FTKY ile bir peptit diziliminin kodlanması

FTKY, amino asitlerin fizikokimyasal özelliklerinin modellemeye yansıtılması üzerine kuruludur. Böylece HIV-1 proteaz kesim konumlarının tahmininde, kesilmiş proteinlerin işlevleri sınıflandırıcıya daha iyi tanıtılmaktadır. Öte yandan FTKY, kalıntıların dizilimsel konumlarını modellemeye yansıtmakta zayıf kalmaktadır.

5.1. Deneysel Sonuçlar ve Analiz

Tez kapsamında yapılan deneysel çalışmalarda, FTKY, 10-fk, 20-fk, 30-fk, 40-fk ve 50-fk'ye göre sınıf doğruluğu, duyarlık ve Alıcı İşletim Karakteristiği Eğrisi Altında Kalan Alan (AİKAA - AUROC) değerleri bakımından test edilmiştir.

Sınıf doğruluğu değeri, modellemesi gerçekleştirilen HIV-1 proteaz enziminin kesme konumlarının tahmininde, doğru tahmin edilen peptit sayısının (kesme konumuna sahip olan ve olmayan), tüm peptit sayısına oranıdır. Duyarlık veya doğru pozitif değeri ise doğru tahmin edilen kesme konumuna sahip peptit sayısının, tüm kesme konumuna sahip peptit sayısına oranını ifade eder 55. Alıcı İşletim Karakteristiği (AİK - ROC) eğrisi, testin değişik kesim noktalarında doğru pozitif (y-ekseni) değerlerinin, yanlış pozitif değerlerine (x-ekseni) karşı noktalanması ile elde edilir. Her kesim noktasındaki doğru pozitif ve yanlış pozitife karşılık gelen noktalar birleştirilerek AİK eğrisi çizilir. AİKAA ise AİK eğrisi altında kalan alanın değeridir 55, 56.

Geliştirilen yöntemin testleri, matematik ve grafik fonksiyonları üzerine kurulu, etkileşimli bir programlama ortamı olan MatLab programında gerçekleştirilmiştir. Doğrusal DVM sınıflandırıcısı OSU Toolbox 57 ile uygulanmıştır. Ayrıca her bir öznitelik kodlama yöntemi hem doğrudan hem de TBA yöntemi ile boyutları indirgenerek testlere dâhil edilmişlerdir. TBA indirgeme yöntemi PRTools 58 ile gerçekleştirilmiştir.

Testler, 10-kat çapraz doğrulama tekniğine (ÇDT) göre gerçekleştirilmiştir. 10-kat ÇDT'de veri seti, 10 kümeye ayrılır. Kesme olan peptit ve olmayan peptitler her bir kümeye rastgele ve eşit olacak şekilde dağıtılır. Bir çapraz doğrulamada, 10 kümeden 9'u eğitim verisi, 1'si test verisi olarak modelleme gerçekleştirilir. Bir testte toplam

10 çapraz doğrulama gerçekleştirilir 31. Böylece her bir küme hem eğitim hem de test verisi olarak test sürecine dahil olur. Elde edilen sonuçlar 10 test üzerinden gerçekleştirilmiştir.

HIV-1 proteaz/substrat etkileşimi için Kontijevskis 4 tarafından 2007 yılında 1625 peptit diziliminden oluşan bir veri seti (PR-1625) yayınlandı. Bu örüntü verilerinin 374'ü kesilmiş (cleavage) peptit, 1251'i kesilmemiş (noncleavage) peptittir. 2008 yılında Schilling 59 tarafından daha geniş bir veri seti (PR-3261) yayınlanmıştır. PR-3261 veri seti, 436 kesilmiş peptit, 2825 kesilmemiş peptitten oluşmaktadır. Her iki veri seti arasında % 7 oranında küçük bir benzerlik bulunmaktadır 60. Giriş uzayında 8 kalıntıdan oluşan toplam peptit sayısının 20^8 olduğu düşünülürse her iki veri setinin toplam peptit kümesinin çok az bir kısmını yansıttığı görülmektedir. Tez kapsamında yapılan deneysel çalışmalarda PR-1625 ve PR-3261 veri setleri kullanılmıştır.

FTKY'nin ilk safhasında en iyi 50 fizikokimyasal özellik, 10-kat ÇDT yöntemi ile elde edilmiştir. Buna göre toplam 10 test yapıldı ve PR-1625 veri seti altında Tablo A.1'de görülen fizikokimyasal özellikler elde edildi. Tablo A.2'de ise PR-3261 veri setinde yapılan testler sonucunda elde edilen en iyi 50 fizikokimyasal özellik görülmektedir. Tablolarda belirtilen sıklık değeri, yapılan 10 test içinde ilgili fizikokimyasal özelliğin sınıf doğruluğu performansının 544 özellik içinde ilk 20'ye girme sıklığıdır. Bu sıklıklar sıralamayı belirlemektedir. Elde edilen verilere göre her iki veri setinde en iyi 50-fk listesinde 13 fizikokimyasal özellik örtüşmüştür.

PR-1625 ve PR-3261 veri setleri üzerinde yapılan testlerde sırasıyla elde edilen Tablo A.1 ve Tablo A.2'de görülen en iyi 50 fizikokimyasal özelliklerin indeks değerleri ortalama 0, varyans 1 olacak şekilde z-skor yöntemine göre normalize edilerek FTKY'ye göre kodlanmıştır. Tablo 5.1'de, en iyi 10, 20, 30, 40 ve 50 fizikokimyasal özelliğe göre FTKY ile kodlanan PR-1625 ve PR-3261 verilerinin, doğrusal DVM algoritmasına göre sınıf doğruluk değerleri görülmektedir.

PR-1625 verileri üzerinde yapılan testlerde, FTKY ile kodlanan girişlere TBA uygulandığında sınıf doğruluğu değerleri 10-fk ve 20-fk'de düşerken 30-fk, 40-fk ve 50-fk'da artmıştır. TBA boyut indirgeme yöntemi öznitelik vektörü boyutu arttıkça

performansı olumlu yönde etkilemektedir. FTKY yöntemi hem TBA'lı hem TBA'sız PR-3261 veri setinde ise PR-1625 veri setine göre daha düşük başarım sergilemiştir. PR-1625 veri seti üzerinde en yüksek sınıf doğruluğu sonucunu 10-fk ile uygulanan TBA'sız kodlama verirken PR-3261 üzerinde ise 30-fk ile uygulanan TBA'sız kodlama vermiştir.

Tablo 5.1. FTKY'nin PR-1625 ve PR-3261 veri setleri üzerinde TBA'lı ve TBA'sız (doğrudan) sınıf doğruluğu başarımı

	PR-1625 Doğrudan	PR-1625 TBA	PR-3261 Doğrudan	PR-3261 TBA
10-fk	**95,15**	94,44	92,08	88,25
20-fk	94,91	94,79	94,04	91,41
30-fk	94,79	95,09	**94,17**	92,36
40-fk	94,4	94,93	94,12	93,65
50-fk	94,1	94,88	94,06	93,85

Tablo 5.2'de FTKY'nin PR-1625 ve PR-3261 veri setleri üzerinde duyarlık değerleri görülmektedir. Yapılan testlerde en yüksek sonuçlar yine PR-1625 üzerinde elde edilmiştir. En yüksek duyarlık değeri PR-1625'de % 90,68 ile 30-fk ile yapılan kodlamada, PR-3261'de ise % 76,63 değeri ile 50-fk'ya göre yapılan kodlamada elde edilmiştir. TBA yöntemi ile öznitelik vektörleri boyutu indirgendiğinde duyarlık performanslarında kayda değer bir artış olmamıştır. Hatta PR-3261 veri seti üzerinde düşüşler meydana gelmiştir. Özellikle 10-fk'ya göre yapılan kodlamada % 60,74'den % 27,28 gibi düşük bir orana erişilmiştir.

Tablo 5.2. PR-1625 ve PR-3261 veri setleri üzerinde FTKY'nin TBA'lı ve TBA'sız duyarlık başarımı

	PR-1625 Doğrudan	PR-1625 TBA	PR-3261 Doğrudan	PR-3261 TBA
10-fk	89,65	87,38	60,74	27,28
20-fk	90,22	89,38	73,63	60,67
30-fk	90,14	**90,68**	75,72	67,23
40-fk	89,46	90,24	76,28	74,19
50-fk	89,22	90,27	**76,63**	74,12

Tablo 5.3'de ise FTKY'nin AİKAA değerleri görülmektedir. Bu sonuçlara göre PR-1625 üzerinde en yüksek başarımı 0,99 değeri ile 20-fk, 30-fk ve 40-fk'ya göre yapılan öznitelik kodlama yöntemleri vermişlerdir. Yine PR-1625 üzerinde yapılan test performansları PR-3261'e göre daha yüksektir. En düşük değer ise PR-3261 üzerinde yapılan ve boyutu TBA ile indirgenen 10-fk'ya göre yapılan kodlamadan elde edilmiştir.

Tablo 5.3. FTKY'nin PR-1625 ve PR-3261 veri setleri üzerindeki karşılaştırmalı AİKAA sonuçları

	PR-1625 Doğrudan	PR-1625 TBA	PR-3261 Doğrudan	PR-3261 TBA
10-fk	0,98	0,98	0,93	0,88
20-fk	**0,99**	**0,99**	**0,96**	0,93
30-fk	0,98	**0,99**	**0,96**	0,95
40-fk	0,98	**0,99**	**0,96**	**0,96**
50-fk	0,98	0,98	**0,96**	**0,96**

10-fk'ya göre yapılan kodlamalarda, PR-1625 veri seti üzerinde en yüksek performans elde edilirken PR-3261'de ise en düşük performans elde edilmiştir. FTKY'nin fizikokimyasal özellikleri temel alan bir kodlama olması amino asitlerin biyokimyasal özelliklerinin modellemeye yansıtılması açısından üstünlüğüdür. Bununla beraber FTKY'nin önemli üç kısıtı bulunmaktadır. Öncelikle FTKY'de en birinci safhada iyi özellikler seçilirken kalıntılar arasındaki bağımlılık görmezden gelinmektedir ve her bir kalıntı bağımsız olarak kodlamaya dahil edilmektedir. Halbuki en iyi öznitelikler için,

$$\binom{544}{n}, \tag{5.1}$$

kadar seçenek denenmelidir. (5.1)' deki notasyonda 544 fizikokimyasal özelliğin n'li kombinasyonları hesaplanmaktadır. Burada *n,* en iyi fizikokimyasal özellik sayısıdır. Bu işlem sonunda en az sınıflandırma hatası yapan *n* adet fizikokimyasal özellik belirlenmelidir. Ayrıca FTKY kalıntıların dizilim içindeki konumları hakkında bilgi içermez. Dolayısıyla bu durum örüntü örneklerinin sınıflandırıcı tarafından açık

olarak tanınamamasına yol açar. Üçüncü kısıtı ise en iyi 30-fk, 40-fk ve 50-fk öznitelik vektörlerinin boyut problemine neden olabilecek büyük boyut gerektirmeleridir. Belirtilen bu kısıtların olumsuz etkileri test sonuçlarında elde edilen sınıf doğruluğu, duyarlık ve AİKAA değerlerinden de anlaşılmaktadır.

BÖLÜM 6. BirTVD ÖZNİTELİK KODLAMA YÖNTEMİ

Geçmiş yıllarda HIV-1 proteaz enzimi tarafından proteinlerin kesme konumlarının tespiti problemine çeşitli öznitelik temsili yöntemleri kullanılmıştır. 25 ve 45'de BKY, 61'de Ardışıl Kayar İleri Yönlü Seçme (Sequential Floating Forward Selection) yöntemi tarafından belirlenen fizikokimyasal özelliklerin BKY ile birleştirilmesi, 48'de ATKY, 51'de BLOSUM50 yer değiştirme matrisi tabanlı kodlama, 62'de Quasi-kalıntı çiftleri temsil yöntemleri HIV-1 özgünlüğü problemine uygulanmışlardır. Fakat problemin çözümüne uygulanan bu öznitelik vektörü temsili yöntemlerinin geliştirilmesinde, proteinlerin işlevlerini anlama ve akabinde örüntü temsilinde şu ölçütlerin aynı anda göz önünde bulundurulmadığı görülmektedir:

- Kalıntıların fizikokimyasal özellikleri,
- Kalıntıların protein içindeki konumları,
- Yöntemlerin füzyonu.

Bu nedenle, HIV-1 proteaz enziminin proteinleri kesme konumlarını tespit etmede geliştirilen BirTVD öznitelik temsili yönteminde bu üç ölçüt göz önünde bulundurulmuştur.

Temsil yöntemlerinde proteinleri oluşturan kalıntıların dizilim içindeki sırasını içeren bilgiler, öznitelik temsiline kolaylıkla dâhil edilememektedirler 63. Bu durum temsilin örüntüyü açık olarak ifade edememesine neden olmaktadır. Kalıntıların protein içindeki pozisyonlarının temsili açısından BKY en uygun yöntemlerdendir. BKY'de, tüm öznitelik vektörleri birbirlerine diktir. Böylece örüntü verileri öznitelik uzayında daha iyi temsil edilirler. Bununla beraber, ikili sayı sistemi ile temsil edilen BKY'de, kalıntıların birbirleri ile olan etkileşimlerine dair herhangi bir bilgi öznitelik temsil vektöründe yer almaz. Diğer bir ifade ile BKY, amino asitlerin birbirleri ile olan fizikokimyasal benzerlikleri veya farklılıkları hakkında bilgilerden yoksundur.

BKY'ye amino asitlerin fizikokimyasal özelliklerine ait sınıflandırılma bilgisi eklenirse, sınıflandırıcının eğitim verilerini daha iyi tanımasının yolu açılır.

TVD, amino asitleri temel 10 fizikokimyasal özelliklere göre bir venn diyagramında kategorize etmiştir. Bu fizikokimyasal özellikler: Hidrofobiklik, pozitiflik, negatiflik, polarlık, şarjlılık, küçüklük, çok küçüklük, alifatiklik, aromatiklik ve prolindir. Eğer BKY'nin kalıntıları konumlama, TVD'nin fizikokimyasal etkileşimlerini tanımlama yönleri birleştirilirse, peptit dizilimlerini daha iyi tanımlayan ve temsil eden bir öznitelik çıkarım yöntemi geliştirilebilir. Bu varsayımdan yola çıkarak, BKY ve ikilik sayı sistemi ile ifade edilen TVD birleştirilerek BirTVD kodlama yöntemi geliştirilmiştir.

Modellemede bir peptit dizilimini, $P = \{P_1P_2P_3,..,P_i\}$ ile ifade edelim. P_i, $1 \leq i \leq N$, olmak üzere P'ye ait *i*. amino asittir. P_i amino asiti BirTVD yönteminde $\{\vec{y}\}_i^1$ ve $\{\vec{y}\}_i^2$ vektörleri olarak kodlanır. $\{\vec{y}\}_i^1$ vektörü, P_i'nin standart amino asit alfabesine göre 20-bit boyunda BKY karşılığıdır. $\{\vec{y}\}_i^2$ vektörü ise Tablo 6.1'de görülen TVD'nin ikilik sayı sistemine dönüştürülmüş karşılığıdır. Buna göre amino asitler, TVD'de ait olduğu özellik kümelerinde 1 ile olmadıkları kümelerde ise 0 ile temsil edilirler. Her bir amino asit öznitelik vektörü, $\{\vec{y}\}_i^2$ 10-bit boyunda olur. $\{\vec{y}\}_i^1$ ve $\{\vec{y}\}_i^2$ vektörleri birleştirilerek 30 bit boyunda BirTVD öznitelik vektörü elde edilir.

Tablo 6.1. $\{\vec{y}\}_i^2$ vektörü için TVD'den elde edilen kod vektörleri

Amino Asit	Hidrofobik	Pozitif	Negatif	Polar	Şarjlı	Küçük	Çok küçük	Alifatik	Aromatik	Prolin
A	1	0	0	0	0	1	1	0	0	0
R	0	1	0	1	1	0	0	0	0	0
N	0	0	0	1	0	1	0	0	0	0
D	0	0	1	1	1	1	0	0	0	0
C	1	0	0	0	0	1	0	0	0	0
Q	0	0	0	1	0	0	0	0	0	0
E	0	0	1	1	1	0	0	0	0	0
G	1	0	0	0	0	1	1	0	0	0
H	1	1	0	1	1	0	0	0	1	0
I	1	0	0	0	0	0	0	1	0	0
L	1	0	0	0	0	0	0	1	0	0
K	1	1	0	1	1	0	0	0	0	0
M	1	0	0	0	0	0	0	0	0	0
F	1	0	0	0	0	0	0	0	1	0
P	0	0	0	0	0	1	0	0	0	1
S	0	0	0	1	0	1	1	0	0	0
T	1	0	0	1	0	1	0	0	0	0
W	1	0	0	1	0	0	0	0	1	0
Y	1	0	0	1	0	0	0	0	1	0
V	1	0	0	0	0	1	0	1	0	0

$\{\vec{y}\}_i^1$ ve $\{\vec{y}\}_i^2$ vektörleri birleştirilerek, Pi amino asiti için, (6.1)'de görüldüğü gibi öznitelik vektörü, $\{\vec{y}\}_i$ elde edilir.

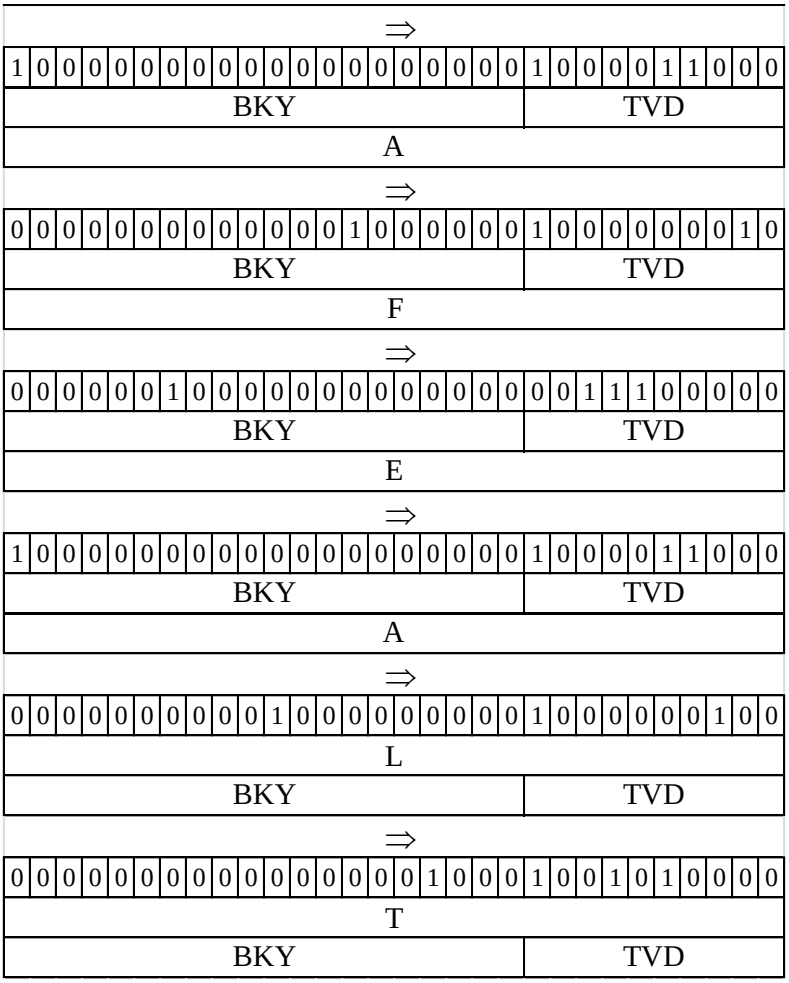
$$\{\vec{y}\}_i = \left(\{\vec{y}\}_i^1 \| \{\vec{y}\}_i^2\right) \tag{6.1}$$

Son olarak P peptidi için (6.2)'de görüldüğü gibi öznitelik vektörü, $\vec{\chi}$ elde edilir.

$$\vec{\chi} = \left(\{\vec{y}\}_1 \| \{\vec{y}\}_2 \| \cdots \| \{\vec{y}\}_N\right) \tag{6.2}$$

$\vec{\chi}$ öznitelik vektörü $30 \times N$ uzunluğundadır.

BirTVD yöntemini, GEAFEALT şeklinde verilen örnek peptit üzerinde açıklayalım. Peptit dizilimini oluşturan tüm kalıntıların BKY öznitelik vektörleri ile Tablo 6.1'de görülen TVD öznitelik vektörleri Şekil 6.1'de görüldüğü gibi birleştirilir.



Şekil 6.1. GEAFEALT peptitinin BirTVD yöntemine göre öznitelik vektörü

$\vec{\chi}$ öznitelik vektörü 8 kalıntından oluşan bir peptit için 30 x 8 = 240 bit büyüklüğünde olmaktadır. Bu vektör büyüklüğü boyut problemine neden olabilir.

6.1. Deneysel Sonuçlar ve Analiz

Tez kapsamında yapılan deneysel çalışmalarda, BirTVD öznitelik kodlama yöntemi, Bölüm 4'de anlatılan BKY, ATKY, BLOSUM50 yer değiştirme matrisi tabanlı, 2-grams, KÇTY, BLOMAP ve TVD yöntemleri ile doğrusal DVM'ye bağlı olarak sınıf

doğruluğu, duyarlık ve AİKAA değerleri bakımından PR-1625 ve PR-3261 veri setleri kullanılarak karşılaştırılmıştır.

Tablo 6.2'de öznitelik yöntemlerinin sınıf doğruluğu karşılaştırmalı sonuçları görülmektedir. BirTVD kodlama, PR-1625 veri seti üzerinde varyansı 0,98 olacak şekilde TBA ile indirgendiğinde 240 olan öznitelik vektörlerinin büyüklüğü 127'ye düşmüştür ve % 95,15 değeri ile en iyi performansı sergilemiştir. TBA'nın varyansı daha düşük veya yüksek tutulduğunda tüm yöntemlerin hem PR-1625 hem de PR-3261 veri setleri üzerindeki sınıf doğruluğu, duyarlık ve AİKAA başarımları düşmüştür. PR-1625 veri seti üzerinde en düşük sınıf doğruluğu değeri ise % 85,23 değeri ile DAA ile öznitelik girişleri indirgenen KÇKY'ye aittir. BirTVD yöntemi PR-3261 veri seti üzerinde de TBA ile indirgendiğinde % 94,84 değeri ile en iyi sonucu vermiştir. PR-3261 veri seti üzerinde performansı en düşük olan öznitelik kodlama yöntemi ise yine DAA ile indirgenen KÇKY yöntemine ait % 86,68 değeridir. DAA öznitelik indirgeme yönteminin her iki PR-1625 ve PR-3261 veri setleri üzerinde yapılan testlerde hemen hemen hiçbir kodlamanın başarımını artırıcı yönde etkisi olmaması kayda değerdir.

Tablo 6.2. Öznitelik kodlama yöntemlerinin PR-1625 ve PR-3261 veri setleri üzerindeki sınıf doğruluğu başarımları

Öznitelik Kodlama Yöntemleri	PR-1625 (%)			PR-3261 (%)		
	Doğrudan	TBA	DAA	Doğrudan	TBA	DAA
BKY	94,73	94,69	94,12	94,39	94,40	93,03
ATKY	92,17	92,13	92,93	89,70	89,62	91,53
BKY+BLOSUM50	93,69	93,31	94,38	93,99	93,8	93,33
2-grams	94,56	94,41	94,23	88,82	88,29	87,77
KÇKY	92,98	92,81	85,23	86,77	86,77	86,68
BLOMAP	92,10	91,51	90,99	90,70	90,59	90,75
TVD	93,37	93,01	92,71	92,18	92,36	91,24
BirTVD	94,90	**95,15**	93,63	94,27	**94,84**	91,87

Tablo 6.3'de yöntemlerin duyarlık değerine bağlı performansları görülmektedir. Elde edilen Deneysel Sonuçlar ve Analiza göre hem PR-1625 hem de PR-3261 veri setleri üzerinde TBA yöntemi ile boyutu indirgenen BirTVD yöntemi ile kodlanan örnekler, sırasıyla % 90,31 ve % 77,02 değerleri ile en iyi sonuçları vermişlerdir. Tüm

yöntemlerde elde edilen duyarlık değerlerine bakıldığında, PR-1625 üzerinde tahmini kesilmiş kısımların, PR-3261'e nazaran daha başarılı olduğu görülmektedir.

Tablo 6.3. Öznitelik kodlama yöntemlerinin PR-1625 ve PR-3261 veri setleri üzerindeki duyarlık başarımları

Öznitelik Kodlama Yöntemleri	PR-1625 (%)			PR-3261 (%)		
	Doğrudan	TBA	DAA	Doğrudan	TBA	DAA
BKY	89,86	89,41	87,27	76,58	75,84	68,37
ATKY	76,97	76,73	83,95	27,74	26,81	56,79
BKY+BLOSUM50	89,11	87,73	88,05	76,93	74,65	69,77
2-grams	83	83,14	84,54	39,37	35,07	36,91
KÇKY	73,24	72,76	43,68	0	0	0,07
BLOMAP	82,16	80,19	78,92	51,14	48,81	49,26
TVD	86,19	85,08	83,54	60,67	61,05	53,95
BirTVD	90,27	**90,31**	86,43	76,21	**77,02**	62,09

Tablo 6.4'de tüm öznitelik çıkarım yöntemlerinin AİKAA sonuçları görülmektedir. Bu deneysel sonuçlara göre, PR-1625 üzerinde BKY ve BirTVD yöntemleri 0,99 AİKAA değeri ile en yüksek başarıma ulaşmışlardır. PR-3261 üzerinde ise BirTVD kodlama 0,97 AİKAA değeri ile diğer yöntemlere göre en yüksek başarımı göstermiştir.

Tablo 6.4. Öznitelik kodlama yöntemlerinin PR-1625 ve PR-3261 veri setleri üzerindeki karşılaştırmalı AİKAA sonuçları

Öznitelik Kodlama Yöntemleri	PR-1625			PR-3261		
	Doğrudan	TBA	DAA	Doğrudan	TBA	DAA
BKY	**0,99**	**0,99**	0,98	0,96	0,96	0,95
ATKY	0,97	0,97	0,97	0,94	0,94	0,94
BKY+BLOSUM50	0,98	0,97	0,98	0,96	0,95	0,96
2-grams	0,96	0,96	0,93	0,85	0,83	0,84
KÇKY	0,96	0,96	0,86	0,87	0,86	0,48
BLOMAP	0,96	0,96	0,96	0,92	0,91	0,91
TVD	0,97	0,97	0,97	0,93	0,94	0,92
BirTVD	**0,99**	**0,99**	0,97	**0,97**	**0,97**	0,94

Deneysel sonuçlar ve analiz, BirTVD yönteminin diğer yöntemlere göre HIV-1 proteaz kesme yerlerinin tahmininde başarımı en yüksek öznitelik yöntemi olduğunu göstermektedir. Ayrıca öznitelik vektörlerinin boyutlarının indirgenmesinde TBA'nın, DAA'ya göre daha üstün olduğu anlaşılmıştır.

Tez çalışması boyunca (5 ve 7. bölümler dahil), PR-3261 veri seti üzerinde, doğrusal DVM sınıflandırıcısına göre yapılan testlerde en iyi performansı BirTVD kodlama yöntemi vermiştir.

BÖLÜM 7. BirBOOL ÖZNİTELİK KODLAMA YÖNTEMİ

6. bölümde anlatılan BirTVD öznitelik kodlama yönteminde, kalıntıların birbirleri ile olan biyokimyasal etkileşimlerini örüntü modeline yansıtmak amacıyla TVD kullanıldı. TVD, 1986 yılında geliştirilmiştir. Hâlbuki AAindex'i oluşturan 544 adet fizikokimyasal özelliğin hemen hemen yarısı daha o tarihte indekste yoktu. Dolayısıyla TVD göreceli eski bir amino asit sınıflandırma diyagramıdır. Barnes 7'de TVD'nin, amino asitleri genel anlamda sınıflandırmada yeterli olsa da basit bir sınıflandırma yaptığını bildirmiştir. Bu varsayımlardan hareketle, 5. bölümde anlatılan FTKY gelişimi için belirlenen en iyi 10, 20, 30, 40 ve 50 fizikokimyasal özellikler, ikilik sayı sistemi ile ifade edilmiş ve BKY ile birleştirilerek BirBOOL öznitelik çıkarım yöntemi geliştirilmiştir.

BirBOOL yönteminde bir peptit dizilimini, P = {P_1P_2 P_3…P_i}, ile ifade edilsin. P_i, $1 \le i \le N$, olmak üzere P'ye ait i. kalıntıdır. P_i kalıntısının, BirBOOL yönteminde $\{\vec{y}\}_i^1$ ve $\{\vec{y}\}_i^2$ vektörleri hesaplanır. $\{\vec{y}\}_i^1$ vektörü, P_i'nin standart amino asit alfabesine göre 20-bit boyundaki BKY karşılığıdır.

$\{\vec{y}\}_i^2$ vektörü için 5. bölümde FTKY'nin gelişiminde elde edilen en iyi 10, 20, 30, 40 ve 50 fizikokimyasal özellikler kullanılarak her bir amino asit için ikilik sayı sisteminde kod tablosu elde edilir. AAindex kümesini, AAi={a_1,a_2,a_3,..,a_d}, AAi'ye karşılık gelen kod vektörü, $\{\vec{y}\}_i^2$={b_1,b_2,b_3,..,b_d}, (d=1,2,3,..,N) ile ifade edelim. Bu durumda (7.1)'de görüldüğü gibi önce ilgili kalıntıya ait fizikokimyasal özelliğin ortalaması (a_o) hesaplanır.

$$a_o = \frac{1}{20}\sum_{d=1}^{N} a_d, \qquad (7.1)$$

Eğer $a_d \leq a_o$, ise $b_d = 0$,

Eğer $a_i \leq a_o$, ise $b_d = 1$ olur.

Böylece en iyi fizikokimyasal özelliklere ait indeks değerleri ikilik sayı sistemine dönüştürülerek kod tablosu elde edilir.

Tablo 7.1'de, PR-1625 veri seti üzerinde yapılan testlere göre belirlenen $\{\vec{y}\}_i^2$'e ait kod tablosu görülmektedir. PR-3261 veri seti üzerinde de 5. bölümde elde edilen en iyi 50 fizikokimyasal özelliğe göre kod tablosu oluşturulmuştur.

Tablo 7.1. PR-1625 veri seti üzerinde $\{\vec{y}\}_i^2$ için belirlenen kod tablosu

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	0	0	1	0
2	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	1
3	1	0	0	0	1	0	0	1	1	1	1	0	1	1	0	0	0	1	0	1
4	1	0	0	1	1	0	0	1	0	1	1	0	1	1	1	1	0	0	0	1
5	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	1
6	0	1	0	0	0	1	0	1	0	1	1	0	1	1	0	1	1	1	1	1
7	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	0	1	0
8	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	0	1	0	1	1
9	0	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1	0	1	0
10	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
11	1	0	0	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	0	1
12	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	1	1	1	1
13	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1
14	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	1
16	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	0	0	0	1	0
17	1	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	1	0	0
18	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
19	0	1	0	0	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	1
20	1	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1
21	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	1	0	1
22	0	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1
23	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	1
24	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	0	0	1	0
25	0	0	0	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1

Tablo 7.1. (Devam) PR-1625 veri seti üzerinde $\{\vec{y}\}_i^2$ için belirlenen kod tablosu

26	0	1	1	1	0	1	1	1	1	0	0	1	0	0	0	1	0	0	0	0
27	1	1	1	1	0	1	1	1	0	0	0	1	0	0	1	1	1	0	0	0
28	1	1	1	0	1	1	1	0	1	0	0	1	1	0	0	1	0	0	0	0
29	0	1	0	0	1	1	0	0	0	1	1	0	0	1	0	1	1	1	1	1
30	1	0	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0	1	0	1
31	1	1	1	1	0	1	1	1	0	0	0	1	0	0	1	1	1	0	0	0
32	0	1	0	0	1	1	0	0	0	1	1	0	1	1	0	0	1	0	1	1
33	1	0	0	0	1	0	0	1	1	1	1	0	1	1	1	0	0	1	0	1
34	0	1	1	1	0	1	1	0	1	0	1	1	1	1	0	0	0	1	1	0
35	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
36	0	0	1	1	0	0	0	1	0	0	0	1	1	0	1	1	1	1	1	0
37	0	1	1	1	1	0	0	1	0	0	0	1	0	0	1	1	1	1	1	0
38	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
39	1	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0
40	1	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
41	0	0	0	0	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1
42	1	0	1	0	0	1	0	1	1	0	0	1	0	0	1	1	0	1	1	0
43	0	1	1	1	0	1	1	1	1	0	0	1	0	0	0	1	0	0	1	0
44	0	0	1	1	1	0	1	0	0	1	0	0	0	1	1	0	1	1	1	1
45	1	1	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1	1	1
46	0	0	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1
47	0	1	0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	1
48	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
49	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0
50	1	0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1

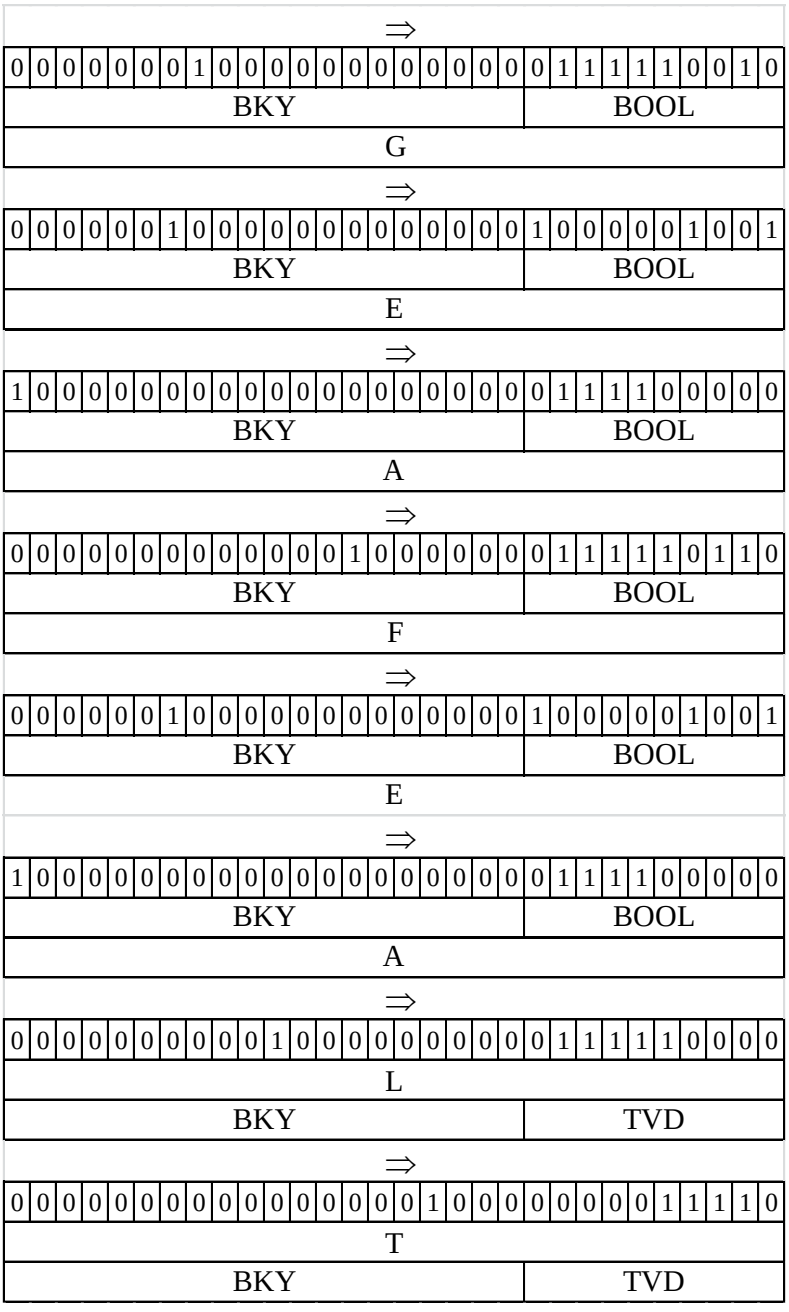
PR-3261 veri seti üzerinde, $\{\vec{y}\}_i^2$ için belirlenen kod tablosu Tablo B.1'de verilmiştir. $\{\vec{y}\}_i^1$ ve $\{\vec{y}\}_i^2$ vektörleri birleştirilerek, P_i amino asiti için, (7.2)'de görüldüğü gibi öznitelik vektörü, $\{\vec{y}\}_i$ elde edilir.

$$\{\vec{y}\}_i = \left(\{\vec{y}\}_i^1 \| \{\vec{y}\}_i^2\right) \tag{7.2}$$

Son olarak P peptidi için (7.3)'de görüldüğü gibi öznitelik vektörü, $\vec{\chi}$ elde edilir.

$$\vec{\chi} = \left(\{\vec{y}\}_1 \| \{\vec{y}\}_2 \| \cdots \| \{\vec{y}\}_N\right) \tag{7.3}$$

BirBOOL yöntemini, GEAFEALT şeklinde verilen örnek peptit üzerinde 10-fk için açıklayalım. Peptit dizilimini oluşturan tüm kalıntıların BKY öznitelik vektörleri ile Tablo 7.1'de görülen $\{\vec{y}\}_i^2$ öznitelik kod vektörlerinden ilk 10 değer Şekil 7.1'de görüldüğü gibi birleştirilir.



Şekil 7.1. GEAFEALT peptitinin BirBOOL yöntemine göre öznitelik vektörü

$\vec{\chi}$ öznitelik vektörü en iyi 10-fk için 30 x 8 = 240 bit, 20-fk için 40 x 8 = 320 bit, 30-fk için 50 x 8 = 400 bit, 40-fk için 60 x 8 = 480 bit ve 50-fk için 70 x 8 = 560 bit büyüklüğünde olmaktadır. Bu vektör büyüklükleri boyut problemine neden olabilir.

7.1. Deneysel Sonuçlar ve Analiz

Tablo 7.2'de görüldüğü gibi PR-1625 verileri üzerinde yapılan testlerde, 10-fk'ya göre BirBOOL ile kodlanan örüntüler, doğrusal DVM altında TBA'sız, % 95,12 ve TBA'lı % 95,28 sonuçları ile en yüksek sınıf doğruluğu değerlerini vermişlerdir. Bu sonuçlar 4. ve 5. bölümde anlatılan öznitelik çıkarım yöntemlerinin sonuçları ile kıyaslandığında PR-1625 verileri üzerinde yapılan testlerde elde edilen en yüksek sonuçlardır. PR-3261 verileri üzerinde yapılan testlerde de $\{\vec{y}\}_i^2$ vektörü 10-fk'ya göre kodlandığında ve TBA ile boyutu indirgendiğinde daha yüksek değerler elde edilmiştir.

Tablo 7.2. BirBOOL yönteminin PR-1625 ve PR-3261 veri setleri üzerinde TBA'lı ve TBA'sız (doğrudan) sınıf doğruluğu başarımı

	PR-1625 Doğrudan (%)	PR-1625 TBA (%)	PR-3261 Doğrudan (%)	PR-3261 TBA (%)
10-fk	95,12	**95,28**	94,17	**94,21**
20-fk	94,83	94,9	94,06	94,05
30-fk	94,44	94,53	94,04	93,82
40-fk	94,3	94,53	93,96	93,94
50-fk	94,03	94,6	93,93	94,17

Ayrıca diğer öznitelik kodlama yöntemlerinin 5. Bölüm'de gerçekleştirilen, Tablo 6.2'de görülen sınıf doğruluğu test sonuçlarına göre, PR-1625 üzerinde en yüksek sınıf doğruluğu oranını % 95,14 ile BirTVD yöntemi vermiştir. Buna karşın BirBOOL yöntemi PR-1625 veri seti üzerinde % 95,28 oranı ile hepsinden daha yüksek sınıf doğruluğu sonucu vermiştir.

Tablo 7.3'de en iyi BirBOOL yönteminin PR-1625 ve PR-3261 veri setleri üzerinde duyarlık değerleri görülmektedir. Yapılan testlerde en yüksek sonuçlar yine PR-1625

veri seti üzerinde elde edilmiştir. En yüksek duyarlık değeri hem PR-1625 hem de PR-3261 veri seti üzerinde sırasıyla % 90,65 (doğrudan) ve % 76,12 ile 10-fk ile yapılan kodlamada elde edilmiştir. TBA yöntemi ile öznitelik vektörleri boyutu indirgendiğinde duyarlık performanslarında 10-fk hariç düşüşler meydana gelmiştir.

Tablo 7.3. PR-1625 ve PR-3261 veri setleri üzerinde BirBOOL yönteminin TBA'lı ve TBA'sız duyarlık başarımı

	PR-1625 Doğrudan	PR-1625 TBA	PR-3261 Doğrudan	PR-3261 TBA
10-fk	90,51	**90,65**	**76,12**	74,77
20-fk	90,16	89,68	76,02	74,53
30-fk	89,51	89,32	75,7	73,37
40-fk	89,43	88,97	75,79	75
50-fk	89,41	89,46	75,93	75,65

Ayrıca diğer öznitelik kodlama yöntemlerinin 5. Bölüm'de gerçekleştirilen, Tablo 6.3'de görülen duyarlık test sonuçlarına göre, PR-1625 üzerinde en yüksek duyarlık oranını % 90,31 ile BirTVD yöntemi vermiştir. Buna karşın BirBOOL yöntemi PR-1625 veri seti üzerinde, % 90,65 duyarlık oranı ile hepsinden daha yüksek performans göstermiştir.

Tablo 7.4'de ise BirBOOL yönteminin PR-1625 ve PR-3261 veri setleri üzerinde, AİKAA değerleri görülmektedir. Bu sonuçlara göre PR-1625 üzerinde en yüksek performansı 0,99 değeri ile 10-fk ve 20-fk'ya göre yapılan yöntemlerde vermişlerdir. PR-1625 üzerinde yapılan test performansları PR-3261'e göre daha yüksektir. PR-3261 üzerinde elde edilen deneysel sonuçlara göre tüm girişler 0,96 AİKAA değerini vermiştir.

Tablo 7.4. BirBOOL yönteminin PR-1625 ve PR-3261 veri setleri üzerindeki karşılaştırmalı AİKAA sonuçları

	PR-1625 Doğrudan	PR-1625 TBA	PR-3261 Doğrudan	PR-3261 TBA
10-fk	**0,99**	0,98	0,96	0,96
20-fk	0,99	0,98	0,96	0,96
30-fk	0,98	0,98	0,96	0,96
40-fk	0,98	0,98	0,96	0,96
50-fk	0,98	0,98	0,96	0,96

Diğer öznitelik kodlama yöntemlerinin 5. Bölüm'de yapılan Tablo 6.5'de görülen AİKAA test sonuçlarına göre BirBOOL yöntemi, PR-1625 üzerinde 0,99 değeri ile BKY ve BirTVD yöntemleri ile aynı AİKAA değerini elde etmiştir. Fakat PR-3261 üzerinde BirBOOL yöntemi 0,96 değeri elde ederken 0,97 değeri ile en yüksek AİKAA başarımını BirTVD yöntemi göstermiştir.

BirBOOL öznitelik kodlama yöntemi ve TBA boyut indirgeme yöntemi birlikteliği sınıflandırma performansını artırıcı yönde etkilediği görülmektedir. Bu durum BirBOOL kodlamanın boyut indirgenmesine elverişli bir yöntem olduğunu göstermektedir.

Tez çalışmaları süresince PR-1625 veri seti üzerinde yapılan testlerde en yüksek başarımı BirBOOL kodlama yöntemi vermiştir.

BÖLÜM 8. SONUÇLAR

Bu tez çalışmasında, HIV-1 proteaz enziminin kesme konumlarının tahmin probleminin çözümü için literatürde bulunan öznitelik çıkarım yöntemleri araştırılmış ve daha yüksek başarım sağlayan BirTVD ve BirBOOL yöntemleri ile FTKY geliştirilmiştir.

Geliştirilen ilk yöntem olan FTKY her ne kadar diğer yöntemlere göre yüksek başarım verememiş olsa da fizikokimyasal özelliklerin yöntem de kullanılması ve en iyi özelliklerin belirlenmesinde önemli rol oynamıştır. Bu belirlenen fizikokimyasal özellikler, daha sonra geliştirilen yöntemlerde de kullanılmıştır.

Geliştirilen ikinci yöntem BirTVD, BKY'nin kalıntıları konumlama ve TVD'nin kalıntıları sınıflandırma bilgileri birleştirilerek elde edilen güçlü bir öznitelik çıkarım yöntemidir. Nitekim elde edilen deneysel sonuçlar BirTVD yönteminin başarımının diğer yöntemlere göre hem PR-1625 hem de PR-3261 veri setleri üzerinde daha yüksek olduğunu göstermiştir.

BirTVD yönteminde kalıntıların birbirleri ile olan biyokimyasal etkileşimlerini örüntü modeline yansıtmak amacıyla 1986 yılında geliştirilen TVD kullanılmıştır. Dolayısıyla TVD kalıntıları kategorize etmede oldukça eski bir sınıflandırmadır. Bu varsayımdan hareketle FTKY'nin gelişimi için belirlenen en iyi 10, 20, 30, 40 ve 50 fizikokimyasal özellikler ikilik sayı sistemi ile ifade edilmiş ve BKY yöntemi ile birleştirilerek BirBOOL öznitelik çıkarım yöntemi geliştirilmiştir. BirBOOL yöntemi, PR-1625 veri seti üzerinde tez süresince yapılan tüm çalışmalarda en yüksek başarımı vermiştir.

HIV-1 proteaz özgünlüğü probleminin çözümüne uygulanan tüm öznitelik kodlama yöntemleri, boyutları hem doğrudan kullanıldığında hem de TBA ile indirgendiğinde PR-1625 veri seti üzerinde, PR-3261 veri setine nazaran daha yüksek başarım göstermişlerdir. Bu durum PR-3261 veri setinin varyansının, PR-1625 veri setinden daha yüksek olduğunu ispatlamaktadır. Diğer bir ifade ile PR-1625 veri seti benzer özelliklere sahip peptitler içerme ihtimali çok yüksektir.

Ayrıca tüm öznitelik çıkarım yöntemleri HIV-1 proteaz özgünlüğü problemi çözümü için PR-1625 ve PR-3261 veri setleri üzerinde doğrusal lojistik (linear logistic), doğrusal perseptron (linear perceptron), doğrusal ayırıcı (linear discriminant) sınıflandırıcıları ile quadratik (quadratic), parzen, naive bayes doğrusal olmayan sınıflandırıcıları ile de test edilmişlerdir. Bu algoritmaların hiçbirisi doğrusal DVM kadar başarılı sonuçlar verememişlerdir. Ayrıca geliştirilen BirTVD ve BirBOOL öznitelik kodlama yöntemleri, ikili, üçlü ve dörtlü olarak birleştirilmiş sınıflandırıcılarla da modellenerek HIV-1 proteaz özgünlüğü problemine uygulanmışlardır. Fakat birleştirilmiş sınıflandırıcılar da doğrusal DVM kadar yüksek başarım verememişlerdir.

Tezin kapsamında olmamakla beraber BirTVD yöntemi MHC Class I bağlanma konumlarının tespiti problemine de uygulanmıştır. Güncel üç adet MHC Class I veri seti kullanılarak BirTVD ile diğer BKY, ATKY, BLOSUM50 yer değiştirme matrisi tabanlı, 2-grams, KÇKY, BLOMAP ve TVD yöntemleri AİKAA değerleri bakımından kıyaslanmıştır. Elde edilen deneysel sonuçlara göre, BirTVD yöntemi her üç veri seti üzerinde en yüksek başarımı göstermiştir. Ayrıca BirTVD yöntemi literatürde yer alan, MHC Class I bağlanma konumlarının tespiti konusunda geliştirilmiş olan DynaPredPOS, NetMHC, SVMHC ve YKW yöntemleri ile de deneysel olarak kıyaslanmıştır. Bu başarım sonuçlarına göre BirTVD yöntemi, iki veri sette NetMHC yönteminin ardından ikinci, son veri sette üçüncülüğü elde etmiştir.

Gelecekte yapılacak olan çalışmalarda, geliştirilen BirTVD ve BirBOOL yöntemlerine kalıntıların fizikokimyasal etkileşimlerini daha iyi tanımlayan bilgilerin eklenmesi ile daha yüksek başarım değerleri elde edilebilir. Ayrıca geliştirilen

yöntemlerden elde edilen yüksek değerli deneysel sonuçlar, bu yöntemlerin başka protein örüntü tanıma çözümlerine de uygulanabileceğini göstermektedir.

KAYNAKLAR

1. WATSON, J.D., CRICK, F.H.C., Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, 171,pp. 737-738, 1953.

2. The Structures of Life. National Institute of General Medical Sciences. http://publications.nigms.nih.gov/structlife/chapter1.html (Erişim tarihi: Mayıs 2011)

3. KEHA, E.E., KÜFREVIOĞLU, Ö.I., Biyokimya, Aktif Yayınevi, Erzurum, 2004.

4. KONTIJEVSKIS, A., WIKBERG, J.E., Computational proteomics analysis of HIV-1 protease interactome. Proteins-Structure Function and Bioinformatics 68(1): pp. 305-312, 2007.

5. UNAIDS 2010 Report on the Global AIDS Epidemic. http://www.unaids.org/globalreport/Global_report.htm (Erişim tarihi: Mayıs 2011)

6. BAROUCH, D., Challenges in the development of an HIV-1 vaccine. Nature. 455: pp. 613-619, 2008.

7. BARNES, M., GRAY, I., Bioinformatics for Genetics. John Wiley & Sons Inc, 2003.

8. KAWASHIMA, S., KANEHISA, M., AAindex: amino acid index database, Nucleic Acids Res. 20 (1): 374, 2000. (www.genome.jp/aaindex/)

9. TAYLOR, W.R., The Classification of Amino-Acid Conservation. Journal of Theoretical Biology 119(2): pp. 205-218, 1986.

10. BERG, J.M., TYMOCZKO, J.L., STRYER, L., Biochemistry. W. H. Freeman, New York, USA, 2001.

11. BERMAN, H.M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T.N., WEISSIG, H., SHINDYALOV, I.N., BOURNE, P.E., The protein data bank. Nucleic Acids Research, 28, pp. 235-242, 2000.

12. BUGG, T., Introduction to Enzyme and Coenzyme Chemistry, Blackwell Science, Great Britain, 2005.

13. BRIGGS, G.E., A Further Note on the Kinetics of Enzyme Action. Biochem J, 19(6), pp. 1037-1038, 1925.

14. KOSHLAND, D.E., Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci, 44(2), pp.98-104, 1958.

15. GAO, F., BAILES, E., ROBERTSON, D., CHEN, Y., RODENBURG, C., MICHAEL, S., CUMMINS, L., ARTHUR, L., PEETERS, M., SHAW, G., SHARP, P., HAHN., B., Origin of HIV-1 in the chimpanzee Pantroglodytes troglodytes, Nature, pp. 397:436-41, 1999.

16. Structure Of HIV, National Institute Of Allergy and Infectious Diseases. http://www.niaid.nih.gov/topics/HIVAIDS/Understanding/Biology/Pages/structure.aspx. (Erişim tarihi: Mayıs 2011)

17. SWANSON, C., MALIM, M., SnapShot: HIV-1 proteins. Cell. 133:742, 2008.

18. ANDERSON, J., AKKINA, R., Complete knockdown of CCR5 by lentiviral vector-expressed siRNAs and protection of transgenic macrophages against HIV-1 infection. Gene Ther 14 (17):1287-1297, 2007.

19. KARAÇAY, B., Yüzyılın Salgını Devam Ediyor; HIV/AIDS'in Dünü, Bugünü ve Yarını, TÜBITAK Bilim ve Teknik Dergisi, 519, pp. 59-65, 2011.

20. CRAIGIE, R., HIV integrase, a brief overview from chemistry to therapeutics. J. Biol. Chem. 276: pp. 23213-23216, 2001.

21. DEMIROV, D., FREED, E., Retrovirus budding. Virus Res. pp. 106:87-102, 2004.

22. NICHOLSON L.K., Flexibility and Function in HIV-1 Protease, Nat Struct Biol 2(4): pp. 274-280, 1995.

23. HIV Replication Cycle, National Institute Of Allergy and Infectious Diseases. http://www.niaid.nih.gov/topics/HIVAIDS/Understanding/Biology/Pages/hivReplicationCycle.aspx. (Erişim tarihi: Mayıs 2011)

24. WLODAWER, A., ERICKSON J.W., Structure-based inhibitors of HIV-1 protease. Annu. Rev. Biochem. 62: pp. 543-585, 1993.

25. RÖGNVALDSSON, T., L., YOU. Why neural networks should not be used for HIV-1 protease cleavage site prediction. Bioinformatics 20(11): 1702-1709, 2004.

26. HORNAK, V., OKUR, A., RIZZO, R. C., SIMMERLING, C., HIV-1 protease flaps spontaneously open and reclose in molecular Dynamics simulations, Proceedings of the National Academy of Sciences 103, pp. 915-920, 2006.

27. DE CLERCQ, E., HIV-chemotherapy and -prophylaxis: new drugs, leads and approaches. Int J Biochem Cell Biol, 36: pp. 1800-1822, 2004.

28. GULICK, R.M., MELLORS, J.W., HAVLIR, D., ERON, J.J., GONZALEZ, C., MCMAHON, D., RICHMAN, D.D., VALENTINE, F.T., JONAS, L., MEIBOHM, A., EMINI, E.A., CHODAKEWITZ, J.A., Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. N Engl J Med 337(11): pp. 734-9, 1997.

29. TAMALET, C., PASQUIER, C., YAHI, N., COLSON, P., POIZOT-MARTIN, I., LEPEU, G., GALLAIS, H., MASSIP, P., PUEL, J., IZOPET, J., Prevalence of drug resistant mutants and virological response to combination therapy in patients with primary HIV-1 infection. J. Med. Virol. 61(2): pp. 181-186, 2000.

30. SHOEMAN, R.L., HONER, B., STOLLER, T.J., KESSELMEIER, C., MIEDEL, M.C., TRAUB, P., GRAVES, M.C., Human immunodeficiency virus type 1 protease cleaves the intermediate filament proteins vimentin, desmin, and glial fibrillary acidic protein. Proc Natl Acad Sci, 87: pp. 6336-6340, 1990.

31. DUDA, R.O., HART, P.E., STORK, D.G., Pattern Classification. 2nd edition, John Wiley & Sons Inc, 2001.

32. JAIN, A.K., DUIN, R.P.W., J., MAO, Statistical Pattern Recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22: pp. 4-37, 2000.

33. SCHÖLKOPF, B., SMOLA, A.J., Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press, 2001.

34. WANG, L., Support Vector Machines: Theory and Applications, Springer, 2005.

35. BISHOP C.M., Pattern Recognition and Machine Learning, 2006.

36. NG, A., JORDAN, M., On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In NIPS-14, Cambridge, MA. MIT Press, 2002.

37. ÇEVIKALP H., Feature Extraction Techniques in High-dimensional Spaces: Linear and non-linear Approaches, Ph.D., Nashville, Tennessee, 2005.

38. AKAY M., Wiley Encyclopedia Of Biomedical Engineering, John Wiley & Sons Inc., Hoboken, New Jersey, 2006.

39. BELHUMEUR, P.N., HESPANHA, J.P., KRIEGMAN, D.J., Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions On Pattern Analysis and Machine Intelligence, 19(7): pp. 711-720, 1997.

40. KITTLER, J., DEVIJVER, P.A., Statistical Properties of Error Estimators in Performance Assessment of Recognition Systems. Pattern Analysis and Machine Intelligence, PAMI-4, pp. 215-220, 1982.

41. AĞIR K., Null Space Approach Of Fisher Discriminant Analysis For Face Recognition, M.Sc., Istanbul Technical University, 2006.

42. AKDEMİR B., Tahmin uygulamalarında performans geliştirmek için kullanılan normalizasyon metotlarına yeni bir yaklaşım. Doktora Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, 2009.

43. JAIN A., NANDAKUMAR K., Score normalization in multimodal biometric systems. Pattern Recognition, 2005.

44. KUNCHEVA, L.I., Combining pattern classifiers, John Wiley & Sons Inc., New Jersey, 2004.

45. GÖK M., ÖZCERİT A.T., Linear Support Vector Machines for HIV-1 Protease Site Detection, ISSD'09, Sarajevo, Bosnia Herzogovia, pp. 381-384, 2009.

46. MAETSCHKE S., TOWSEY M., BODEN M., Blomap: An encoding of amino acids which improves signal peptide cleavage prediction. Proceedings of the 3rd Asia-Pacific Bioinformatics Conference, pp.141-150, 2005.

47. NARAYANAN, A., WU, X., YANG, Z.R., Mining viral protease data to extract cleavage knowledge. Bioinformatics, 18: Suppl 1, pp. 5-13, 2002.

48. ORSENIGO, C., VERCELLIS, C., Predicting HIV protease-cleavable peptides by discrete support vector machines, in EvoBIO, 2007.

49. PEVSNER J., Bioinformatics and Functional Genomics, John Wiley & Sons Inc, 2009.

50. BARNES M., GRAY I., Bioinformatics for Genetics. John Wiley & Sons Inc, 2003.

51. NANNI L., LUMINI A., A reliable method for HIV-1 protease cleavage site prediction methods. Neuro Computing 69: pp. 838-841, 2006.

52. WU, C., WHITSON, G., Protein Classification Artificial Neural System. Protein Science 1(5): pp. 667-677, 1992.

53. GUO, J., LIN., A., Novel Method for Protein Subcellular Localization: Combining Residue-Couple Model and SVM. Proceedings of the 3rd Asia-Pacific Bioinformatics Conference, pp. 117-129, 2005.

54. GÖK M., ÖZCERİT, A.T., OĞUL H., Prediciting HIV-1 Protease Cleavage Site Using Support Vector Machines with Physicochemical Properties. INISTA 2010, Kayseri, Turkey, pp. 260-263, 2010.

55. FAWCETT, T., ROC graphs: Notes and practical considerations for researchers. Technical Report, HP Laboratories. California, USA, 2004.

56. ELMALI F., Altın Standartlı ve Altın Standartsız Durumlarda, Yarı Parametrik ve Parametrik Olmayan ROC eğrisi Yöntemlerinin Karşılaştırılması. Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı, Doktora Tezi, 2009.

57. JUNSHUI, M.A., Y.I., ZHAO., OSU SVM Toolbox for MATLAB, 2002. (http://sourceforge.net/projects/svm/)

58. DUIN, R.P.W., JUSZCZAK, P., PACLIK, P., PEKALSKA, E., DE RIDDER, D., TAX, M.J., VERZAKOV, S., PRTools5.1. A Matlab Toolbox for Pattern Recognition. Delft University of Technology, 2007.

59. SCHILLING, O., OVERALL, C.M., Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nat Biotechnol 26(6): pp. 685-694, 2008.

60. ROGNVALDSSON, T., ETCHELLS, T.A., How to find simple and accurate rules for viral protease cleavage specificities. BMC Bioinformatics 10: 149, 2009.

61. NANNI L., LUMINI A., MppS: An ensemble of support vector machine based on multiple physicochemical properties of amino acids. Neurocomputing 69 (13-15): pp. 1688-1690, 2006.

62. NANNI L., Comparison among feature extraction methods for HIV-1 protease cleavage site prediction. Pattern Recognition 39(4): pp. 711-713, 2006.

63. CHOU. K.C., Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Order Effect. Biochem. biophys. res. commun. 278: pp. 477-483, 2001.

KİŞİSEL YAYINLAR ve ESERLER

[1] GÖK M., ÖZCERİT A.T., Linear Support Vector Machines for HIV-1 Protease Site Detection, ISSD'09, Sarajevo, Bosnia Herzogovia, pp. 381-384, 2009.

[2] GÖK M., ÖZCERİT, A.T., OĞUL H., Prediciting HIV-1 Protease Cleavage Site Using Support Vector Machines with Physicochemical Properties. INISTA 2010, Kayseri, Turkey, pp. 260-263, 2010.

[3] GÖK M., ÖZCERİT, A.T., A New Feature Encoding Scheme for HIV-1 Protease Cleavage Site Prediction, Neural Computing and Applications, 2011 (gönderildi).

[4] GÖK M., ÖZCERİT, A.T., OETMAP: A New Feature Encoding Scheme for MHC Class I Binding Prediction, Mol. and Cell. Biochemistry, 2011 (gönderildi).

EKLER

Ek A. PR-1625 ve PR-3261 Veri Setlerine Göre Seçilen En İyi 50 Fizikokimyasal Özellik

Tablo A.1. PR-1625 veri setine bağlı olarak doğrusal DVM sınıflandırıcısı, sınıf doğruluğu değerlerine göre seçilen en iyi 50 fizikokimyasal özellik

Sıra	AA ind.	Fizikokimyasal Özellik	Sıklık
1	320	Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)	10
2	148	Side chain interaction parameter (Krigbaum-Komoriya, 1979)	9
3	302	Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982)	8
4	242	Average gain in surrounding hydrophobicity (Ponnuswamy et al., 1980)	8
5	519	Hydrophobicity-related index (Kidera et al., 1985)	7
6	66	Membrane preference for cytochrome b: MPH89 (Degli Esposti et al., 1990)	7
7	535	ALTLS index (Cornette et al., 1987)	7
8	529	Hydrophobicity scale from native protein structures (Casari-Sippl, 1992)	7
9	130	Ratio of buried and accessible molar fractions (Janin, 1979)	7
10	237	Normalized frequency of turn in alpha+beta class (Palau et al., 1981)	6
11	448	Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)	5
12	530	NNEIG index (Cornette et al., 1987)	4
13	73	Melting point (Fasman, 1976)	4
14	505	Linker propensity from small dataset (linker length is less than six residues) (George-Heringa, 2003)	4
15	520	Apparent partition energies calculated from Wertz-Scheraga index (Guy, 1985)	4

Tablo A.1.(Devam) PR-1625 veri setine bağlı olarak doğrusal DVM sınıflandırıcısı, sınıf doğruluğu değerlerine göre seçilen en iyi 50 fizikokimyasal özellik

16	298	Average reduced distance for side chain (Rackovsky-Scheraga, 1977)	4
17	540	Optimized relative partition energies - method B (Miyazawa-Jernigan, 1999)	4
18	88	Positive charge (Fauchere et al., 1988)	3
19	468	Interior composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001)	3
20	534	ALTFT index (Cornette et al., 1987)	3
21	341	Information measure for middle helix (Robson-Suzuki, 1976)	3
22	417	Normalized positional residue frequency at helix termini C1 (Aurora-Rose, 1998)	3
23	196	Normalized composition from fungi and plant (Nakashima et al., 1990)	3
24	243	Average gain ratio in surrounding hydrophobicity (Ponnuswamy et al., 1980)	3
25	128	Percentage of buried residues (Janin et al., 1978)	3
26	241	Surrounding hydrophobicity in folded form (Ponnuswamy et al., 1980)	3
27	441	Distribution of amino acid residues in the 18 non-redundant families of mesophilc proteins (Kumar et al., 2000)	3
28	41	Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)	2
29	221	Optimized average non-bonded energy per atom (Oobatake et al., 1985)	2
30	356	Side chain hydropathy, corrected for solvation (Roseman, 1988)	2
31	172	Normalized frequency of extended structure (Maxfield-Scheraga, 1976)	2
32	428	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al., 1994)	2
33	526	Weights from the IFH scale (Jacobs-White, 1989)	2
34	131	Transfer free energy (Janin, 1979)	2
35	304	Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga, 1982)	2
36	522	Apparent partition energies calculated from Janin index (Guy, 1985)	2
37	536	TOTFT index (Cornette et al., 1987)	2
38	541	Optimized relative partition energies - method C (Miyazawa-Jernigan, 1999)	2

Tablo A.1.(Devam) PR-1625 veri setine bağlı olarak doğrusal DVM sınıflandırıcısı, sınıf doğruluğu değerlerine göre seçilen en iyi 50 fizikokimyasal özellik

39	425	Normalized flexibility parameters (B-values), average (Vihinen et al., 1994)	2
40	308	Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982)	1
41	179	Retention coefficient in HPLC, pH2.1 (Meek, 1980)	1
42	186	Normalized frequency of alpha-helix (Nagano, 1973)	1
43	275	Weights for beta-sheet at the window position of -2 (Qian-Sejnowski, 1988)	1
44	124	Normalized relative frequency of helix end (Isogai et al., 1980)	1
45	449	Hydropathy scale based on self-information values in the two-state model (20% accessibility) (Naderi-Manesh et al., 2001)	1
46	307	Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga, 1982)	1
47	170	Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)	1
48	151	Hydropathy index (Kyte-Doolittle, 1982)	1
49	469	Entire chain composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)	1
50	318	Transfer free energy from vap to oct (Radzicka-Wolfenden, 1988)	1

Tablo A.2. PR-3261 veri setine göre doğrusal DVM sınıflandırıcısının sınıf doğruluğu değerlerine göre seçilen en iyi 50 fizikokimyasal özellik

Sıra	AA ind.	Fizikokimyasal Özellik	Sıklık
1	129	Percentage of exposed residues (Janin et al., 1978)	10
2	36	Proportion of residues 100% buried (Chothia, 1976)	9
3	128	Percentage of buried residues (Janin et al., 1978)	9
4	318	Transfer free energy from vap to oct (Radzicka-Wolfenden, 1988)	8
5	35	Proportion of residues 95% buried (Chothia, 1976)	8
6	275	Weights for beta-sheet at the window position of -2 (Qian-Sejnowski, 1988)	8
7	522	Apparent partition energies calculated from Janin index (Guy, 1985)	7
8	310	Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)	7
9	308	Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982)	6
10	400	Polarity (Zimmerman et al., 1968)	6
11	131	Transfer free energy (Janin, 1979)	5
12	257	Beta-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)	5
13	199	Transmembrane regions of non-mt-proteins (Nakashima et al., 1990)	5
14	30	A parameter of charge transfer capability (Charton-Charton, 1983)	4
15	276	Weights for beta-sheet at the window position of -1 (Qian-Sejnowski, 1988)	4
16	127	Average accessible surface area (Janin et al., 1978)	4
17	201	Ratio of average and computed composition (Nakashima et al., 1990)	4
18	92	Helix initiation parameter at posision i-1 (Finkelstein et al., 1991)	4
19	344	Information measure for pleated-sheet (Robson-Suzuki, 1976)	4
20	526	Weights from the IFH scale (Jacobs-White, 1989)	4
21	130	Ratio of buried and accessible molar fractions (Janin, 1979)	3
22	277	Weights for beta-sheet at the window position of 0 (Qian-Sejnowski, 1988)	3
23	518	Average internal preferences (Olsen, 1980)	3
24	213	Average non-bonded energy per atom (Oobatake-Ooi, 1977)	3
25	172	Normalized frequency of extended structure (Maxfield-	3

		Scheraga, 1976)	

Tablo A.2. (Devam) PR-3261 veri setine göre doğrusal DVM sınıflandırıcısının sınıf doğruluğu değerlerine göre seçilen en iyi 50 fizikokimyasal özellik

26	388	Polar requirement (Woese, 1973)	3
27	426	Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbours (Vihinen et al., 1994)	3
28	334	Relative preference value at C2 (Richardson-Richardson, 1988)	2
29	101	Beta-strand indices (Geisow-Roberts, 1980)	2
30	75	pK-N (Fasman, 1976)	2
31	438	p-Values of mesophilic proteins based on the distributions of B values (Parthasarathy-Murthy, 2000)	2
32	343	Information measure for extended (Robson-Suzuki, 1976)	2
33	382	Average interactions per side chain atom (Warme-Morgan, 1978)	2
34	27	The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)	2
35	365	Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)	2
36	166	Frequency of occurrence in beta-bends (Lewis et al., 1971)	2
37	348	Information measure for middle turn (Robson-Suzuki, 1976)	2
38	531	SWEIG index (Cornette et al., 1987)	2
39	41	Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)	2
40	487	Hydrophobicity scales (Ponnuswamy, 1993)	2
41	167	Conformational preference for all beta-strands (Lifson-Sander, 1979)	2
42	237	Normalized frequency of turn in alpha+beta class (Palau et al., 1981)	1
43	356	Side chain hydropathy, corrected for solvation (Roseman, 1988)	1
44	478	Slopes tripeptide FDPB VFF all (Avbelj, 2000)	1
45	279	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)	1
46	242	Average gain in surrounding hydrophobicity (Ponnuswamy et al., 1980)	1
47	532	PRIFT index (Cornette et al., 1987)	1
48	212	Transfer energy, organic solvent/water (Nozaki-Tanford, 1971)	1
49	110	Composition (Grantham, 1974)	1

50	315	Transfer free energy from oct to wat (Radzicka-Wolfenden, 1988)	1

Ek B. PR-3261 Veri Seti Üzerinde Seçilen En İyi 50 Fizikokimyasal Özelliğe Ait $\{\vec{y}\}_i^2$ Kod Tablosu

Tablo B.1. PR-3261 veri seti üzerinde $\{\vec{y}\}_i^2$ için belirlenen kod tablosu

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	0	0	1	0
2	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	1
3	1	0	0	0	1	0	0	1	1	1	1	0	1	1	0	0	0	1	0	1
4	1	0	0	1	1	0	0	1	0	1	1	0	1	1	1	1	0	0	0	1
5	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	1
6	0	1	0	0	0	1	0	1	0	1	1	0	1	1	0	1	1	1	1	1
7	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	0	1	0
8	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	0	1	0	1	1
9	0	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1	0	1	0
10	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
11	1	0	0	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	0	1
12	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0	0	1	1	1	1
13	1	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1
14	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
15	0	1	0	0	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	1
16	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	0	0	0	1	0
17	1	0	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	1	0	0
18	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
19	0	1	0	0	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	1
20	1	0	0	0	1	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1
21	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	1	0	1
22	0	1	0	0	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1
23	1	0	0	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	1
24	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	0	0	1	0
25	0	0	0	0	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1	1
26	0	1	1	1	0	1	1	1	1	0	0	1	0	0	0	1	0	0	0	0
27	1	1	1	1	0	1	1	1	0	0	0	1	0	0	1	1	1	0	0	0
28	1	1	1	0	1	1	1	0	1	0	0	1	1	0	0	1	0	0	0	0
29	0	1	0	0	1	1	0	0	0	1	1	0	0	1	0	1	1	1	1	1
30	1	0	0	1	0	0	1	1	0	1	1	0	0	0	1	0	0	1	0	1
31	1	1	1	1	0	1	1	1	0	0	0	1	0	0	1	1	1	0	0	0
32	0	1	0	0	1	1	0	0	0	1	1	0	1	1	0	0	1	0	1	1

Tablo B.1.(Devam) PR-3261 veri seti üzerinde $\{\vec{y}\}_i^2$ için belirlenen kod tablosu

0	0	1	0	0	1	1	1	1	0	1	1	1	0	0	1	0	1
1	1	0	1	1	0	1	0	1	1	1	1	0	0	0	1	1	0
0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
1	1	0	0	0	1	0	0	0	1	1	0	1	1	1	1	1	0
1	1	1	0	0	1	0	0	0	1	0	0	1	1	1	1	1	0
0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
0	1	0	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0
0	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
0	0	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1
1	0	0	1	0	1	1	0	0	1	0	0	1	1	0	1	1	0
1	1	0	1	1	1	1	0	0	1	0	0	0	1	0	0	1	0
1	1	1	0	1	0	0	1	0	0	0	1	1	0	1	1	1	1
0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	1	1	1
0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1
0	0	1	0	0	0	1	1	1	0	1	1	0	0	0	0	1	1
0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1	1	1
1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1

ÖZGEÇMİŞ

Murat Gök, 1976 yılında Kırşehir'de doğdu. Ortaöğretimini Ankara'da 1993 yılında tamamladı. 2000 yılında Marmara Üniversitesi, Teknik Eğitim Fakültesi, Elektronik-Bilgisayar Eğitiminden mezun oldu. 2006 yılında Muğla Üniversitesi Fen Bilimleri Enstitüsü Elektronik-Bilgisayar Anabilim Dalında yüksek lisansını tamamladı. Aynı yıl Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Elektronik-Bilgisayar Eğitimi, Anabilim Dalında doktoraya başladı. Başlıca akademik çalışma alanları örüntü tanıma, makine öğrenmesi algoritmaları, öznitelik çıkarımı ve seçimi, protein sınıflandırma ve karar destek sistemleridir.