BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ALERJEN PROTEİNLERİN OTOMATİK SINIFLANDIRILMASI

ÖYKÜ EREN

YÜKSEK LİSANS TEZİ
2008

ALERJEN PROTEİNLERİN OTOMATİK SINIFLANDIRILMASI

AUTOMATED CLASSİFİCATİON OF ALLERGEN PROTEİNS

ÖYKÜ EREN

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

2008

Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Bu çalışma, jürimiz tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan	:….....
	(Doç. Dr. Nizami GASİLOV)
Üye (Danışman)	:….....
	(Yrd. Doç. Dr. Hasan OĞUL)
Üye	:….....
	(Yrd. Doç. Dr. Özlen KONU)

ONAY

Bu tez 21/05/2008 tarihinde, yukarıdaki jüri üyeleri tarafından kabul edilmiştir.

..../05/2008

Prof.Dr. Emin AKATA

FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRÜ

Anneanneme...

TEŞEKKÜR

Çalışma alanımı seçerken gösterdiği yol, tüm çalışmam boyunca verdiği bilimsel katkılar, eğitimim boyunca bana kazandırdıkları, çalışma sürem boyunca yaptığım işten heyecan duymamı sağladığı, ayrıca değerli zamanını ve tecrübesini benimle paylaşarak, yaptığım çalışmayı ve mesleğimi bana bir kere daha sevdirdiği için çok sevgili hocam ve tez danışmanım Yrd. Doç. Dr. Hasan Oğul'a gönülden teşekkür ederim.

Akademik hayatım boyunca örnek aldığım, eğitimimde, akademik alanda ilerlememde, neredeyse attığım her adımda bana destek olan, bilgisini ve tecrübesini ihtiyacım olan her an benimle paylaşan, desteğini hiç esirgemeyen canım hocam Sayın Prof. Dr. Hayri Sever'e gönülden teşekkür ederim.

Beni büyüten, bugünlere getiren birtanecik anneannem Türkan Yılmaz'a bu çalışmamda da manevi desteğini esirgemediği için çok teşekkür ederim. Varlığımın sebebi, canım annem Neriman Eren ve canım babam A. Abdullah Eren'e her zaman olduğu gibi sevgi ve destekleriyle yanımda oldukları için çok teşekkür ederim. Çalışmam boyunca her türlü kaprisime katlanan, esprileriyle moralimi hep yüksek tutmamı sağlayan çok sevdiğim birtanecik kardeşim Çağlar'a çok teşekkür ederim.

Ne zaman ihtiyacım olsa imdadıma yetişen, doyumsuz sohbetleriyle bana moral veren, çalışmam boyunca yaşadığım zorlukları hafifleten, benim için yeri ve önemi ayrı, çok sevdiğim Dr. İstemi Barış Özsoy'a teşekkür ederim.

İyi günde, kötü günde olduğu gibi yaptığım bu çalışmamda da yanımda olan, maddi manevi desteğini esirgemeyen, çalışmam boyunca bana ilham veren, gece gündüz ayırmadan fikirlerini paylaşan canım arkadaşım ve meslektaşım Ayşe Yaşar'a bana göstermiş olduğu sabır ve ilgi için çok teşekkür ederim.

Başkent Üniversitesi Bilgisayar Mühendisliği kadrosunda yeralan herkese teşekkür ederim. Başka bir ortamda olsam belki de bu kadar severek ve heyecanla çalışamazdım…

ÖZ

ALERJEN PROTEİNLERİN OTOMATİK SINIFLANDIRILMASI

Öykü EREN

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Alerjen proteinlerin tanınması ve sınıflandırılması, özellikle son yıllarda sıkça kullanılan genetik değişikliğe uğramış gıdaların denetlenmesi ve biyo-ilaçların tasarımı açısından büyük önem kazanmıştır. Dünya Sağlık Örgütü ve Gıda ve Tarım Örgütü kurumları bu amaçla alerjen proteinlerin tespiti için bazı rehberler hazırlamıştır. Ancak, bu rehberlerde önerilen yöntemler çoğunlukla yarı-otomatik gerçekleştirilen ve tahmin yeterliliği düşük olan yöntemlerdir. Son birkaç yılda bazı otomatik yöntemler önerilse de bunlar ya istenilen yeterlilik seviyesine ulaşamamış ya da işlem zamanı ve bellek gereksinimi açısından avantajsız olmuşlardır. Bu çalışmada, alerjen proteinlerin sadece dizilim verisi kullanılarak, farklı makine öğrenme yöntemleri bilinen bazı dizilim gösterim yaklaşımları ile denenmiştir. Farklı dizilim gösterim yöntemleri için K-En Yakın Komşu, Bulanık K-En Yakın Komşu ve Destek Vektör Makineleri (DVM) kullanılmış ve sonuçlar karşılaştırmalı olarak verilmiştir.

ANAHTAR SÖZCÜKLER: Protein Dizilimi, Alerjen Protein, Genetik Değişikliğe Uğramış Gıda, Destek Vektör Makineleri, K- En Yakın Komşu Algoritması, Benzerlik.

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

ABSTRACT

AUTOMATED CLASSIFICATION OF ALLERGEN PROTEINS

Öykü EREN

Baskent University Institute of Science

Department of Computer Engineering

The prediction and classification of the allergen proteins have received great importance on the inspection of genetically modified food, which are used especially in the recent years, and the design of bio-pharmaceuticals. World Health Organization (WHO) and Food and Agriculture Organization (FAO) prepared guidelines for the prediction of allergen proteins. However, the methods proposed in these guidelines are mostly semi-automatic and have low prediction accuracy. Although some automated methods have been proposed in the last few years, either they could not reach the required sufficiency level or they were insufficient as for the processing time and memory usage. In this study, various machine learning methods were tried with some known sequence representation approaches by using only the sequence data of the allergen proteins. For various sequence representation approaches, K-Nearest Neighbour, Fuzzy K-Nearest Neighbour and Support Vector Machines (SVM) were used and the results were given with comparison.

KEYWORDS: Protein Sequence, Allergen Protein, Genetically Modified Food, Support Vector Machines, K-Nearest Neigbour Algorithm, Similarity.

Supervisor: Asst. Prof. Dr. Hasan OĞUL, Baskent University, Department of Computer Engineering

İÇİNDEKİLER LİSTESİ

Sayfa No

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

Kısaltma	Açıklama
DNA	Deoksiribonükleik Asit
DVM	Destek Vektör Makinesi
MAX	Maksimum
KNN	K-En Yakın Komşu
WHO	World Health Organization
FAO	Food and Agriculture Organization
SVM	Support Vector Machine
GDO	Genetiği Değiştirilmiş Organizmalar
MEME/MAST	Multiple Expectation Maximization for Motif Elicitation – Motif Alignment and Search Tool
RNA	Ribonükleik Asit
BLAST	Basic Local Aligment Search Tool
ILSI	International Life Sciences Institute/International Food Biotechnology Committee
PAM	Point Accepted Mutations
PPV	Positive Prediction Value
NPV	Negative Prediction Value
MCC	Matthew's Correlation Coefficient

1. GİRİŞ

1.1 Tezin Kapsamı ve Düzeni

Alerjik reaksiyon proteinlerle ilgilidir. Son yıllarda yapılan çalışmalar sonucunda ortaya çıkan gıda alerjisinin, ilaç alerjisi vb. alerjilerin sebebi genetiği ile oynanmış maddelerdir. Alerjiye sebep olan bu tür gıda veya ilaçların tespit edilmesi büyük önem kazanmıştır. Yapılan bu çalışmada alerjen proteinlerin otomatik sınıflandırılması için farklı dizilim yöntemleri ile birlikte K-En Yakın Komşu, Bulanık K-En Yakın Komşu ve Destek Vektör Makineleri kullanılmıştır. Bunlara ek olarak yalnızca dizilim yöntemlerinden benzerlik skorları kullanılırken, en benzerin sınıfına bağlı bir sınıflandırma uygulaması denenmiştir. Bu yöntem en büyük benzerlik değerini bulan ve sınıflandırma işlemini en benzer ile aynı şekilde sınıflayan bir mantıkla çalışmaktadır.

Birinci bölümde, tezin kapsamı ve düzeni hakkında bilgi verilerek, tez içeriğinde sıkça rastlanacak ve tez içeriğini anlaşılır kılacak biyoloji bilimine ait kavram ve öğelere değinilmiştir. Alerjinin ne olduğu, alerjiye nelerin sebep olduğu ve protein yapısı gibi kavramlar açıklanmıştır.

İkinci bölümde, son yıllarda sıkça kullanılan genetik değişikiliğe uğramış gıdaların denetlenmesi ve ilgili biyo-ilaçların tasarımı ile büyük önem kazanan, alerjen proteinlerin tanınması ve sınıflandırılması ile ilgili yapılan literatür taraması ve çalışmalar anlatılmaktadır. Konu ile ilgili denenen yöntemler ve bazı çalışmaların sonuçları verilmiştir.

Üçüncü bölüm, Makine öğrenme yöntemleri hakkında bilgi vermektedir. Bu bölümde K-En Yakın Komşu Yöntemi, Bulanık K-En Yakın Komşu Yöntemi ve Destek Vektör Makineleri ile ilgili bilgi verilmiştir. Protein gösterim yöntemleri anlatılmıştır. Deney düzeneği, yapılan çalışmalar, kullanılan yöntemler anlatılmıştır.

Dördüncü bölümde, yapılan çalışma ile ilgili performans değerlendirmesi sonuçları her yöntem için ayrı ayrı verilmiştir. Ayrıca yöntemlerden elde edilen sonuçlar karşılaştırmalı olarak değerlendirilmiştir.

Beşinci bölümde yapılan çalışma ile ilgili uygulanan yöntemler, hesaplanan değerler, alınan sonuçlar değerlendirilmiş ve yorumlanmıştır.

1.2 Biyoenformatik

Biyoenformatik genel olarak biyolojik problemlerin, özellikle moleküler biyolojideki problemlerin çözümünü bilgisayar teknolojisi ve bununla ilişkili veri işleme aygıtları ile gerçekleştiren bilimsel disiplinin ismidir. Matematik, enformatik ve yaşam bilimlerini birleştirerek gen ve protein işlevlerini anlamaya yönelik bir bilim dalıdır. Bu bilim dalında temel olarak herhangi bir sorunun çözümü için izlenecek yol olan algoritmanın çıkarılması ve veri tabanı işlemesi yapılarak, protein ve gen dizilimleri ile ilgili bilgilerin işlenmesi ve derlenmesi yapılır. Enformatik teknikler kullanılarak çeşitli biyoloji veri bankalarından gelen bilgi anlaşılır ve organize hale getirilir. Bilgisayarların moleküler biyolojide kullanımı üç boyutlu molekülller yapıların grafik temsili, moleküler dizilimler ve üç boyutlu moleküler yapı veritabanları oluşturulmasıyla başlamıştır. Daha sonra kısa süre içerisinde bu alandaki gelişmeler hızla artmıştır. Çok yüksek miktarda veri üretilmesi, endüstri düzeyinde gen ifadesi, protein-protein ilişkisi, biyolojik olarak aktif molekül araştırmaları, bakteri, maya ,hayvan ve insan genom projeleri gibi biyolojik deneylerin doğurduğu taleple bu alana verilen önem artmıştır.

Biyoenformatik araçların kullanıldığı araştırma konularından bazıları şunlardır:

- DNA dizilimleri
- Protein dizilimleri
- Protein-protein ilişkileri
- Karmaşık genetik fonksiyon ya da regülasyon faaliyetlerinin tanımlanması
- İnsan genom projesi
- Genetik faktörlerin,hastalık yatkınlığına olan etkileri
- Etkileşimli genler için bilgi ağları oluşturulması
- Heterojen biyolojik veritabanlarının entegrasyonu
- Bilgisayarlı veri analizleri

• Makromoleküler yapıların üç boyutlu dizilimleri ve üretimi
• Biyolojik bilginin paylaşımının kolaylaştırılması
• Biyolojik olayların simülasyonu
• Metabolik yol izleri ve hücre algılama modellemesi
• Protein familyalarının nasıl evrimleştiği mekanizmasının anlaşılması
• Hücre ve doku proteinlerinin haritalarının çıkarılması
• Protein yapı ve fonksiyonunun belirlenmesi
• Herhangi bir biyolojik fonksiyonu artıran veya engelleyen küçük moleküllerin tasarlanması

Yeni genlerin bulunması, genlerin yapı analizinden fonksiyonlarının tayini ve bir genin yapısındaki değişmenin hastalıklarla ilişkisinin araştırılmasında dizi analizleri kullanılmaktadır. Günümüzde Biyoenformatik insan genomundaki genlerin dizilimlerinin ve haritalarının elde edilmesinde kullanılmakta ve yeni bilgilerin analizlerinin yapılması ile uğraşmaktadır. Yapılan bu çalışmalarla elde edilen bilgiler değişik genetik ve diğer hastalıkların daha iyi anlaşılmasına ve yeni ilaçların belirlenmesine fayda sağlayacaktır. Sonuç olarak Biyoenformatik; ilaç tasarımı, gen terapisi, biyokimyasal işlemler gibi biyoteknoloji alanlarında uygulama bulan bir disiplin olarak kendini gösterir.

Biyolojik Veri tabanları :

Araştırıcıların nükleotidlerle ilgili bilgilere ulaşabilmesi ve yeni veriler girebilmeleri için biyolojik veritabanları oluşturulmuştur. Bu veritabanlarında milyonlarca nükleotidin depolanması ve organizasyonu yapılmaktadır. Biyoenformatikte nükleotid dizi bilgilerinin organizasyonu ve depolanmasını yapan kuruluşlar şunlardır:

1. GenBank (Gen Bankası- Maryland, ABD)
2. EMBL (Avrupa Moleküler Biyoloji Laboratuvarı – Hinxton , İngiltere)
3. DDBJ (DNA Japonya Veritabanı – Mishima, Japonya)

Bütün araştırmacılara açık olan bu 3 kuruluş, nükleotid dizi bilgilerinin toplanması ve dağıtılmasında işbirliği içinde çalışmaktadır.

Protein dizi verileri ile ilgili hizmet sağlayan kuruluşlar ise şunlardır:

- GenBank
- EMBL
- PIR İnternational (Protein Identification Resource)
- Swiss-Prot.

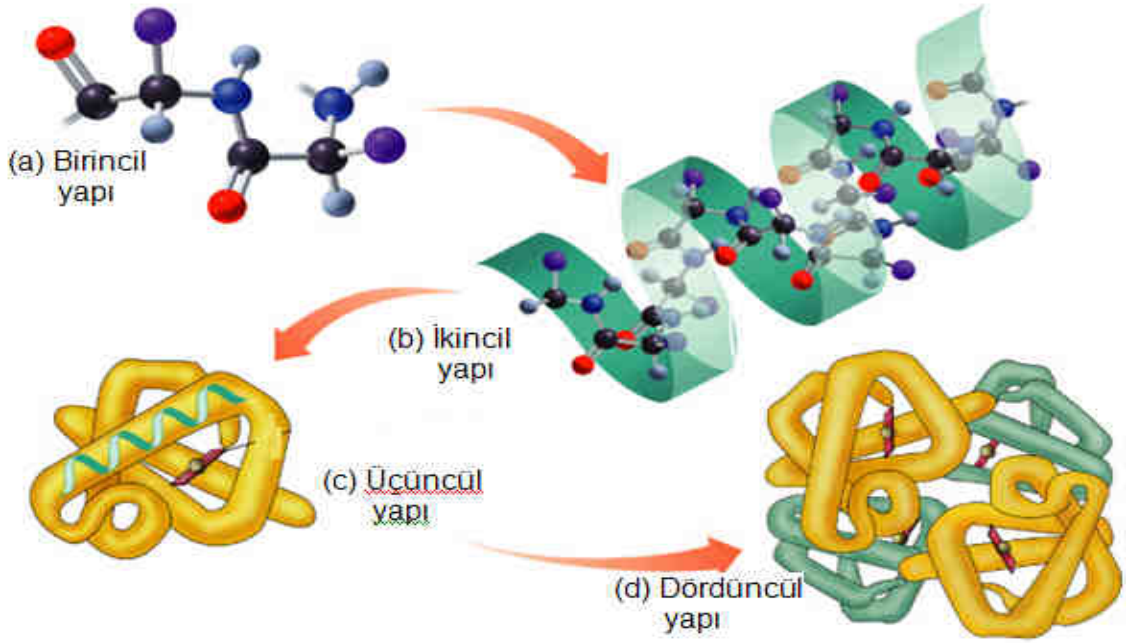
1.3 Gerekli Biyoloji Bilgisi

1.3.1 Proteinler

Proteinler, bütün biyolojik olayların gerçekleşmesindeki en önemli bileşenlerden biridir. Enzimlerin tamamı, hormonların çoğu, bağışıklık sistemimizin büyük bir bölümü, kaslarımızın bütünü ve birçok vücut dokusu (saç, tırnak, kas, diş minesi) proteinden oluşur. Kandaki hemoglobin proteini oksijen taşımakta, antikor denilen proteinler vücudun savunma sisteminin temelini oluşturmakta, insülin hücrelere glikoz alımını sağlamakta, keratin saç ve tırnak yapısını meydana getirmekte, enzim denilen proteinler ise hücre içi kimyasal reaksiyonları yerine getirmektedir. Proteinler, birbirlerine bir zincir şeklinde peptit bağıyla bağlanmış amino asitlerden oluşan çok büyük organik bileşiklerdir. Karbon, hidrojen, oksijen ve azottan oluşurlar.

Hücrelerde protein sentezi sonrasında üretilen amino asitlerin birbirine bağlanarak oluşturdukları düz zincir, daha sonra amino asitler arasındaki kimyasal bağlar neticesi katlanarak proteine nihai bir şekil verir. Proteinlerin bazıları heliks/sarmal yapıda olabileceği gibi küresel veya antikorlar gibi Y şeklinde de olabilirler. Proteinler üç boyutlu yapılarındaki girinti çıkıntılar sayesinde ya başka proteinlere ya da alıcı moleküllere bağlanarak hücre içi faaliyetleri gerçekleştirirler. Anahtar-kilit ilişkisine benzer sistemlerle proteinlerin birbirlerine ya da diğer moleküllere bağlanıp ayrılması, protenlerin üç boyutlu yapılarını çok önemli kılar. Bir proteinin aktif bölgesindeki sadece bir amino asidin bile yerinin değişmesi, proteinin şeklini değiştirip iş görmesini engellemektedir. Bu nedenle protein sentezi sonrası zincir gibi olan amino asit dizisinin katlanarak asli şeklini alması çok önemlidir.

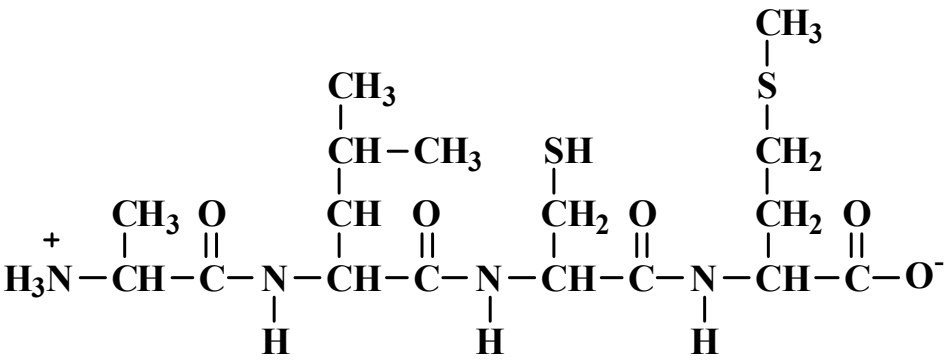
Proteinlerin 4 çeşit yapısı vardır. Bunlar primer (birincil), sekonder (ikincil), tersiyer (üçüncül) ve kuarterner (dördüncül) yapılardır. Bu dört yapının birbiriyle olan ilişkisi Şekil 1.3.1.1'de gösterilmiştir.



Şekil 1.3.1.1 Proteinin (a) birincil (b) ikincil (c) üçüncül (d) dördüncül yapısı (Karen C. Timberlake, "General, Organic, and Biological Chemistry", Benjamin Cummings, 2003)

1-Primer (Birincil) Yapı:

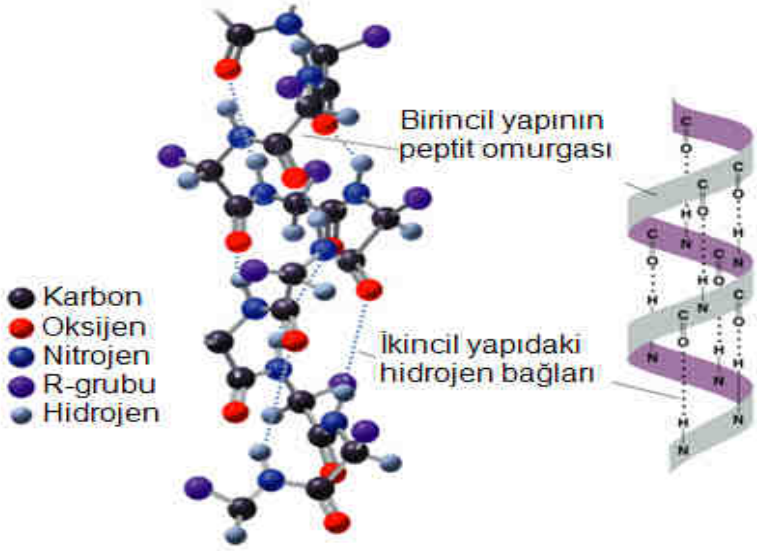
Bir proteindeki amino asitler belirli bir sıraya göre dizilmişlerdir. Bu dizilişe o proteinin birincil yapısı denir. Şekil 1.3.1.2'de proteinin birincil yapısı gösterilmiştir. Bir başka deyişle bir protein için karakteristik ve genetik olarak tespit edilmiş olan amino asit dizilişidir. Belirli türde ve sayıda, belirli diziliş sırasındaki amino asitlerin birbirlerine peptit bağlarıyla bağlanarak oluşturdukları bir polipeptit zinciri biçimindeki yapısıdır. Polipeptitlerde yer alan amino asit bölümlerine de amino asit kalıntılıları (residue) denir. Mutasyona uğramış bir proteinin birincil yapısı bilinmesiyle hastalık teşhisi mümkün olmaktadır.



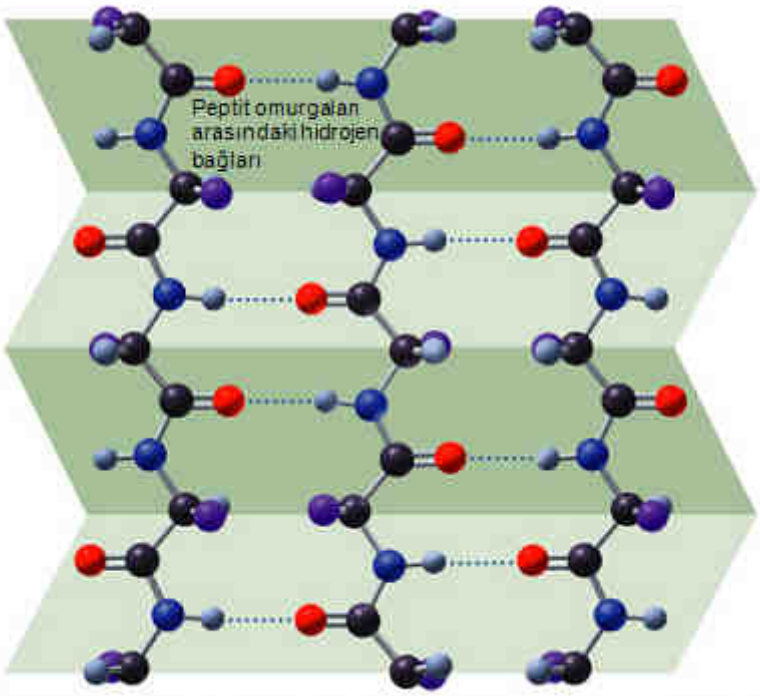
Şekil 1.3.1.2 Proteinin Birincil Yapısı

2-Sekonder (İkincil) Yapı:

Bulunan amino asitlerin doğasına ve düzenine bağlı olarak hidrojen bağları ile kararlı halde olan düzenli tekrarlanan 3-boyutlu yerel yapılardır. Peptit zincirine ait amino asitlerin uzaydaki düzenleniş biçimidir. İkincil yapı Şekil 1.3.1.3'te gösterilen beta konformasyonu (yassı) ya da Şekil 1.3.1.4'te alfa heliks (bobin) gibi yapıların farklı parçalarından oluşur. Bu yapıların yerel olmalarından dolayı bir proteinin içinde farklı ikincil yapılara sahip pek çok bölge olabilir.



Şekil 1.3.1.3 β-konformasyonu (Karen C. Timberlake, "General, Organic, and Biological Chemistry", Benjamin Cummings, 2003)

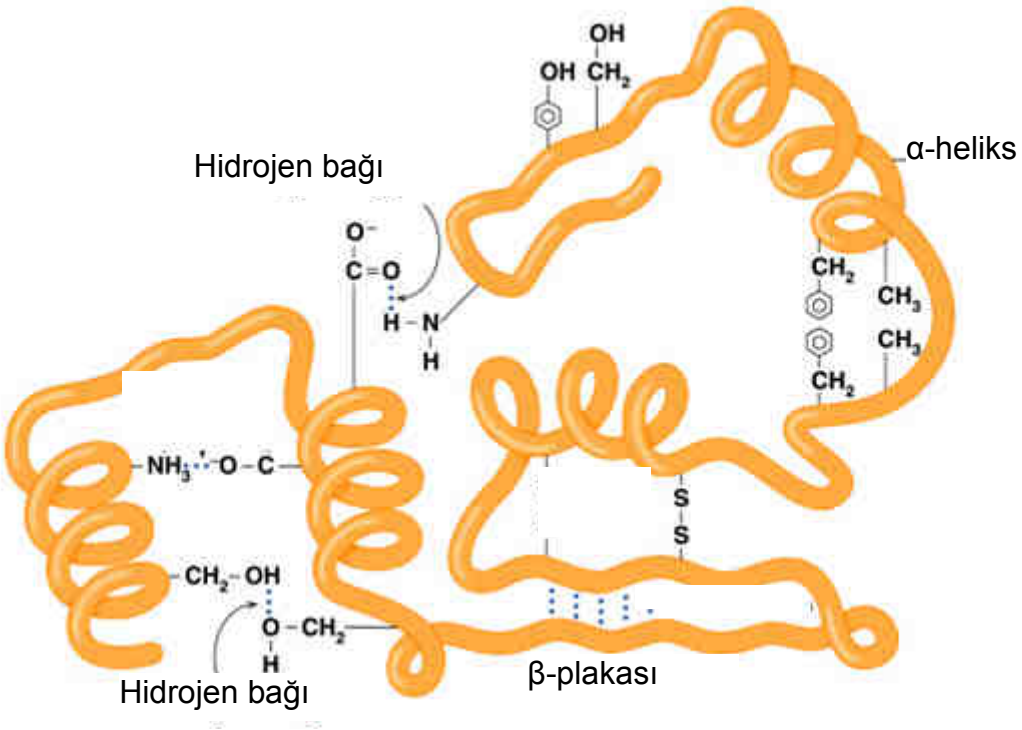


Şekil 1.3.1.4 Alfa Heliks (Karen C. Timberlake, "General, Organic, and Biological Chemistry", Benjamin Cummings, 2003)

3- Üçüncül (Tersiyer) yapı

Polipeptit zincirinin, ikincil yapının oluşmasından sonra, bağ kuvvetlerinin tümüyle uzayda daha ileri seviyede katlanmalar ve düzenlenme sonucu oluşan yapısıdır. Molekül içerisinde daha fazla kıvrım ve tekrar düzenlenme oldukça, daha fazla tabaka oluşur. Oluşan bu yapıya da üçüncül yapı denir. Tek bir proteinin tamamının şekli olan bu yapı, ikincil yapıların birbiriyle olan uzaysal ilişkisidir. Her proteinin içerisinde alfa heliksler, beta plakalılar ve rastgele parçalar bulunur. Üçüncül yapı ile katlama (fold) eş anlamlıdır.

İkincil ve üçüncül yapı ve molekül içindeki çeşitli amino asit yan zincirleri ve etrafını saran su molekülleriyle arasındaki kovalent olmayan etkileşimler sonucunda, yani iyonik bağlar, hidrojen bağlar, hidrofobik etkileşimler sonucunda olur. Şekil 1.3.1.5'te molekül, en kararlı biçimindedir. Proteinin farklı bölgeleri, sıklıkla farklı işlevler ile, yapısal olarak farklı alanlar oluşturabilir. Yapı ile ilgili alanlar, benzer işlevleri yapan farklı proteinlerde bulunurlar.



Şekil 1.3.1.5 Kararlı Yapı (Karen C. Timberlake, "General, Organic, and Biological Chemistry", Benjamin Cummings, 2003)

4-Dördüncül (Kuarterner) Yapı

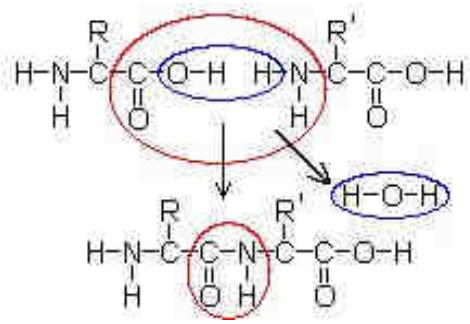
Proteinin açığa vurulan yüzeyi, proteinleri içeren diğer moleküller ile etkileşimlerini de kapsıyor olabilir. Protein-protein etkileşimleri organize olduğu en yüksek seviye dördüncül yapıdır. Enzimlerin alt-birimlerinin arasındaki organizasyon veya yapısal polimerik proteinler buna örnek olarak verilebilir.

1.3.2 Amino asitler

Proteinler çok sayıda amino asit denilen küçük yapı taşlarından oluşur. Doğada 300'den fazla amino asit vardır. Fakat memelilerde bunlardan yalnız 20 tane bulunur. Proteinlerde 20 çeşit amino asit (kimyasal yapısı RCH(NH2)COOH olan) bulunabilir. Bir nitrojen (N) ve iki hidrojen (H) atomu amino grubunu (-NH2) oluşturur. Karboksil grubu (-COOH) ise asit varlığını oluşturur. Amino asidi belirleyen yan zincir ise R- grubudur.

Amino asitler, bir amino asitin karboksil gurubu ile diğer amino asitin amino grubunun reaksiyona girmesi sonucu bir molekül su (H_2O) açığa çıkararak, oluşan peptit bağıyla -C(=O)NH- birbirine bağlanır. Peptitler 2 veya daha fazla amino asitten oluşan bileşiklerdir. Oligopeptitlerde 10 veya daha az amino asit bulunur. Polipeptitler ve proteinler 10 veya daha fazla amino asit zincirlerinden oluşmuşlardır. 50'den fazla amino asitten oluşan peptitler, protein olarak sınıflandırılır.

İki amino asitten bir peptitin oluşumu Şekil 1.3.2.1'de gösterilmiştir.



Şekil 1.3.2.1 İki Amino asitten Peptit Oluşumu

Burada gösterilen R ve R', fonksiyonel gruplardır. Mavi bölge açığa çıkan suyu (H_2O), kırmızı bölge de oluşan peptit bağını (-C(=O)NH-) gösterir. Bu reaksiyonun tersi, yani peptit bağlarının amino asit bileşenlerine parçalanması hidrolizle gerçekleşir. Piyasadaki birçok gıda ürününde lezzet verici olarak hidrolize edilmiş sebze proteinleri kullanılır.

Proteinleri karbonhidrat ve yağlardan ayırt eden özellik azot bulundurmalarıdır. İnsan vücudu protein sentezi için yaklaşık 20 farklı amino aside ihtiyaç duyar. Bunlardan sekizi vücutta sentez edilmemektedir. Gerekli protein ya da amino asit te denilen bu amino asitler izolösin, lösin, lisin, metiyonin, fenilalanin, trionin, triptofan ve valin'dir. Gerekli amino asitler vücut dışındaki kaynaklardan alınır.

Bir protein birçok amino asitin birbirine bağlanması ile oluşmuştur. 20 amino asitten oluşmuş bir çok farklı kombinasyonlar büyük sayılarda protein oluşmasına izin vermektedir. Aynı alfabedeki 29 harfin farklı dizilişleri ile farklı kelime ve cümlelerin yazılabilmesi gibi; 20 amino asit ile de sonsuz sayıda farklı protein

üretmek mümkündür. Proteinlerin 50 kadar amino asit içeren türlerinden, binlerce amino asit içeren türlerine kadar yüzbinlerce çeşidi vardır.

Bir gıdadaki protein gerekli amino asitleri bize sağlıyorsa, bu protein tam protein olarak adlandırılır. Eğer bütün gerekli amino asitleri sağlamıyorsa, eksik protein olarak adlandırılır. Hububat, meyve ve sebzelerdeki proteinler eksik proteinler olarak kabul edilirler. Bitki proteinleri bütün gerekli amino asitleri içermek için ve tam protein oluşturmak için kombine edilebilirler.

Gıdalar vücudumuzda sindirildikten sonra, amino asitlere ayrışırlar ve gerekli proteinlerin sentezini yapmak için vücudumuz tarafından kullanılırlar. Bunlar büyüme ve gelişim için gereklidirler. Protoplazma gibi yeni hücre bileşenlerinin yapılmasında ve aynı zamanda antikorlar, enzimler ve hormonlar vb. yapımı için de gereklidir. Ayrıca proteinler enerji kaynağı olarak ta kullanılırlar. Bütün et ve diğer hayvansal ürünler tam proteinlerin kaynağıdır. Bu ürünler arasında sığır, kuzu, domuz, kümes hayvanları, balık, kabuklu deniz ürünleri, yumurta (tam proteinlerin en iyi kaynağı), süt ve süt ürünleri yer almaktadır.

1.4 Alerji

Alerji, insan vücudunun zararsız bazı maddelere karşı aşırı reaksiyon göstermesidir. Bağışıklık (İmmün) sistemi, vücudu zararlı organizmalara karşı korumak için antikorlar üretirler. Bu vücut savunucuları istilacı olan antijenleri zararsız hale getirirler. Alerjik reaksiyona yol açan antijene alerjen adı verilir.

İyi bir hafızaya sahip olan bağışıklık sistemimiz, yaşamımızın başlamasından itibaren vücudumuzun karşılaştığı yabancı maddeleri tanımayı ve belleğine almayı öğrenir. Adına antijen denilen bu yabancı maddelere karşı antikorlar üreterek tepkisini hazırlar. Vücutta ne zaman aynı antijen görülse hatırlama özelliği nedeniyle daha önceden hazırlanmış yanıt başlar. Bu nedenle saman nezlesi olan bir kişi her yıl polenlerle karşılaşınca bağışıklık sistemindeki bu özellik sebebiyle hemen reaksiyon gösterir.

Alerjik reaksiyon proteinlerle ilgilidir. Gıda proteinlerinin sadece küçük bir kısmı alerjik reaksiyonlara sebep olur. Bağışıklık sistemimiz normal bir alerjik reaksiyonda antijenlere karşı antikorlar oluşturur. Antikor, bir antijeni etkisiz hale getirmek için o antijene özel olarak bağlanan ve vücuttan atan bir proteindir. Antikorlardan imünoglobulin E (IgE) olarak bilinen grup antijenlerle reaksiyona girer ve bir kan hücresi tipi olan bazofillerle doku hücrelerinin (mast hücreleri) içinde olduğu bir reaksiyona sebep olur. Mast hücreleri deri yüzeyinin altında, burunda, solunum yollarında, gözlerde ve bağırsaklarda bulunur. Mast hücrelerinde histamin, lökotrien ve prostoglandinler olarak adlandırılan kimyasal maddeler salgılanır ve bu maddeler alerjik tepkiye neden olurlar. Bu reaksiyonlar aniden gelişir ve genellikle bölgeseldir. Alerjik reaksiyonlar tek tip değildir ve birçok yolla ortaya çıkarlar. Vücudun değişik bölümlerinde meydana gelebilirler ve çeşitli şiddette olabilirler.

Ev tozları, maytlar, küf mantarları, polenler ve bazı gıdalar alerjiye sebep olabilirler. Gıdalar insan vücudunda birçok reaksiyona sebebiyet verebilir, ancak gıdalara bağlı her reaksiyon alerji olarak nitelendirilemez. Gıdalarla oluşan reaksiyonların bir kısmı, o gıdayı alan her insanda oluşabilir. Bunlar gıdalar içindeki zararlı toksik veya mikrobik maddelere bağlıdır.

Gıdalarla oluşan reaksiyonların diğer bir kısmı ise, sadece bazı kişilerde oluşur. Genel olarak 2 gruba ayrılır. Bunlar doğuştan olan bir enzim eksikliği ile görülen alerjik olmayan reaksiyonlar ve alerjik gıda reaksiyonlarıdır. Gıda alerjileri erişkinlerde %1, çocuklarda %2.5 oranında görülmektedir.

Alerjiye en çok neden olan gıdalar çocuklarda süt, yumurta, yerfıstığı, buğday, soya, fındık, ceviz, erişkinlerde ise yerfıstığı, fındık, ceviz, balık ve deniz kabuklularıdır.

Bağışıklık tepkileri maddelerin tamamına değil, epitop adı verilen bazı özel alerji yapıcı bölgelerine karşı gelişir. Gıdalar ve diğer alerjenler arasındaki bu alerji yapıcı bölge benzerliği, çapraz reaksiyonlara neden olabilir. Gıda-gıda, gıda-polen, gıda-mayt ya da gıda-lateks şeklinde çapraz reaksiyonlara rastlanmaktadır. Örneğin, inek sütüne alerjisi olanlar, koyun veya keçi sütüne karşı çapraz alerjik

reaksiyon verebilirler. Buna benzer şekilde, latekse alerjisi olanlar, kestane, avokado veya muza karşı çapraz alerjik reaksiyon verebilirler.

Bir gıda farklı kişilerde farklı semptomlara neden olabilir. Aynı kişide farklı zamanlarda veya farklı dozlarda farklı semptomlara neden olabilir. Bulantı, kusma, karın ağrısı, kramplar, nezle, astım, deride kaşıntı, kızarıklık, egzema, migren, astım, rinit gibi şekillerde görülebilir.

Genetiği değiştirilmiş organizmanın kısaltması olan GDO ifadesi, Avrupa kuralları tarafından tanımlanmaktadır. Bir organizmanın genetik olarak değiştirilmiş olması demek doğal yollarla, geçiş ile veya rekombinasyon yoluyla yapılamayan genomun değişmesi demektir. Bu tanımda, gen rekombinasyonunun doğal yolu ile elde edilen yeni tür amacındaki organizmanın bir türden diğer bir türe transferi sırasındaki işleme, geçiş işlemi denilmektedir. Fakat, bitkileri ele alırsak, genler vektörlerin yardımı ile transfer edilirse, bu doğal bir işlem değildir ve bu nedenle genetik olarak değişmiştir. Ayrıca hücre ergimesinin prosedürleri de genetik modifikasyon olarak adlandırılır. Bunun aksine, örneğin kimyasal ajanlarla zorlamalı mutasyon içeren bitkiler genetik olarak modifiye edilmiş bitkiler olarak adlandırılmazlar.

Bir transgenik (genetik olarak modifiye edilmiş) bitki yapay yolla elde edilmiş bir veya birkaç gen içerir. Hatta birbiriyle hiç alakası olmayan iki tür bu yolla karıştırılabilir. Bu yolla herbisite karşı dayanıklılık veya böceklere karşı direnç gibi istenilen özellikler istenilen üründe elde edilebilir.

ISAAA ‘ ın yönetim kurulu başkanı Dr. Clive James tarafından derlenen Ticari Transgenik (genetik modifiye) Ürünlerin Yıllık Dünya Genelinde Durumuna yönelik derlemeye göre, başlıca yetiştirilen transgenik ürünler soya fasulyesi, mısır, pamuk ve kolza tohumudur.

2. ÖNCEKİ ÇALIŞMALAR

Çeşitli gıdalar, polenler, veya toz zerrecikleri ve bunların içerisinde yaşayan küçük organizmalar alerjiye sebep olabilirler. Genetiği değiştirilmiş organizmaların ve biyo-ilaçların artmasından dolayı alerjenlerin tahmin edilmesi önem kazanmaktadır [28;42;63;65]. Endüstrileşmiş pek çok ulusta atopik alerji ve yüksek hassasiyet %25 lere kadar çıkabilmektedir [50;52]. Avrupa birliğindeki gıda alerjisinin yaygınlığının toplam nüfus içerisindeki oranının %2.5-%3.2 olduğu tahmin edilmektedir [38;39].

Dünya sağlık örgütü (WHO) ve Gıda ve Tarım Örgütü (FAO), proteinlerin potansiyel alerjenliklerinin tespiti icin karar ağacı tabanlı bazı yönergeler önermiştir [21;22]. 2001 yılında yayınlanan yönergenin biyoenformatik bölümünde; bilinen alerjenlerle karşılaştırıldığında, bir proteinde 80 amino asitlik bir pencerede en az %35'lik dizilim benzerliği veya 6 ardışık amino asit özelliği aynı ise, o protein alerjendir, denilmektedir. Bu doğrultuda, pek çok araştırma grubu bu kriterleri sağlayan hesaplamalı yöntemler geliştirmişlerdir. Bu iki yaklaşımdan 6 ardışık amino asit kuralının spesifik olmadığı ve yanlış pozitif sonuçlar ürettiği görülmüştür [10;25;30,43]. Minimum %35 dizilim benzerliğinin ise çok katı olduğu sonucuna varılmıştır [61;64]. Ayrıca, başvurulan genel veritabanlarından başka alerjen reaksiyonlarla ilgili özel veritabanları da ortaya çıkmıştır [13]. 2003 yılında Codex Alimentarius Commission (Codex) bir panel düzenlemiş ve FAO/WHO 2001 önerileri gözden geçirilerek farklı testlerdeki belirsizliklere dikkat çekmiştir. Proteinlerin alerjen davranışlarının incelenmesi konusunda gen kaynağını, bilinen alerjenlerle olan dizilim benzerliğini, protein ve IgE bağlarının stabilitesini kapsayan çok çeşitli testlerle beraber delil ağırlığı yaklaşımı önerilmiştir.

Bu sorunlardan dolayı potansiyel alerjenlerin tespitinde kullanılmak üzere yeni yöntemlere ihtiyaç duyulmuştur. Alerjen proteinlerin tahmininde kullanılan çok çeşitli kriterlere dayanmış metodlar geliştirilmiştir. Bunlardan birisi alerjen proteinlerin tipik bölümlerinin tespiti için geliştirilen, iteratif motif bulma esasına dayanan ve MEME/MAST (Multiple Expectation Maximization for Motif Elicitation – Motif Alignment and Search Tool) metodudur [64]. Motifin anlamı protein dizilimi içerisinde kendisini sürekli tekrar eden dizilim desenidir. MEME'de girdi olarak

protein dizilimleri kullanılır (eğitim kümesi) ve çıkış olarak istenen miktarda motif üretilir. Bu metodda olası her harfin desendeki yeri için olasılık matrisleri oluşturulur. Her motif için dizilim genişliği, tekrar etme sayısı ve her motifin tanımını en iyi şekilde istatistiki yöntemle otomatik olarak seçer. MAST ise bilinen motiflerin bir kısmını barındıran dizilimleri dizilim veritabanlarında tarayan bir araçtır [6].

Protein dizilimi hizalaması yaparak benzerlik bulmaya yarayan bir başka araç ta FASTA'dır [46] .FASTA programları protein veya DNA dizilimleri arasndaki yerel ya da global benzerlikleri bulurlar. Bunu yaparken protein veya DNA veritabanlarını tararlar, ya da dizilimdeki yerel kopyaları belirlerler. FASTA girdi olarak amino asit dizilimini alır ve buna karşılık gelen veritabanını yerel dizilim hizalamayı kullanarak benzerlikler bulur. FASTA paketinde DNA-DNA, protein-protein, sıralı veya sırasız peptit taraması yapan programlar bulunur.

Veritabanlarında araştırma yapabilmek için tasarlanmış bilgisayar programlarından bir tanesi de BLAST (Basic Local Aligment Search Tool) [4] programıdır. Veritabanlarında tarama yapan çok çeşitli BLAST programı bulunmaktadır. Veritabanında homoloji araştırması için öncelikle uygun BLAST programının seçilmesi gerekir. Bunlardan bir tanesi BLASTN'dir. BLASTN bir nükleotid dizisi ile tamamlayıcı diziyi ele alarak nükleotid dizisi veritabanlarıyla yüksek hızda karşılaştırır. Yüksek duyarlılık ihtiyacı olan durumlar için uygun değildir.

FASTA3 algoritmasının k-en yakın komşuluk sınıflandırılmasıyla birleştirilmesiyle bir başka alerjen protein tahmin metodu geliştirilmiştir [73]. Bu metod, Gauss sınıflandırıcıları kullanılarak genişletilmiş ve daha büyük bir protein veritabanı kullanılarak daha kapsamlı hale getirilmiştir [62]. Dalgacık dönüşümlü alerjen protein tahmin etme metodları [43] ve alerjen simgeleyen peptitleri kullanan metodlar da [64] literatürde yer almaktadır. IgE epitoplari, epitop profilleri, yapı profilleri vb. benzerlik taramalarına dayanan metodlar da geliştirilmiştir [36, 40]. Bunlardan başka destek vektör makineleri de (SVM) yeni yöntemlerde kullanılmaktadır [18;57].

Bir proteinin diziliminin karşılaştırılmasının iki amacı vardır; birincisi alerjen protein olup olmadığının anlaşılması için, ikincisi ise bir başka proteinle alerjen çapraz reaksiyona sebep olabilecek bir proteine yeteri kadar benzerlik taşıyıp taşımadığının anlaşılması içindir [3;21;29;51]. Karşılaştırma metodu ve ne kadarlık bir benzeşmenin anlamlı olduğu önemlidir. GDO'ların ortaya çıkmasından sonra tespit edilen alerjen veritabanı sürekli büyümüştür. 1998 yılında ilk internet tabanlı alerjen dizilim veritabanı derlenmiştir [24]. AllergenOnline veritabanının (http://www.AllergenOnline.com) oluşturulmasında da benzer yöntemler kullanılmış, bugün itibariyle (Nisan 2008) 229 türden 1313 alerjen bu veritabanında yer almıştır.

ILSI (International Life Sciences Institute/International Food Biotechnology Committee) değerlendirmesinde, GDO proteini ile bilinen bir alerjen proteinin ardışık 8 veya daha çok amino asit eşleşmesi olması durumu, çapraz reaksiyon potansiyeli açısından ilk düşünülmesi gereken kriter olacağı önerilmiştir [51]. FASTA taramasının yapılması da önerilmiş fakat ilgili eşleşme tanımlanmamıştır.
Bazı çalışmalarda görüldüğü üzere 5 amino asitli IgE bağlanan peptitlerin de olabileceği gibi uzunluğu 8 amino asitten fazla olan peptitlerde IgE bağlanmasının daha olası olduğu sonucuna varılmıştır [7;8;56]. Ayrıca, molekülün bütününde %70 benzerlik gösteren proteinlerin çapraz reaksiyon ihtimali çok fazla iken bu benzerlik %50'nin altındaysa çapraz reaksiyon ihtimali çok düşüktür [1]. Proteinler hakkında daha fazla bilgi ortaya çıktıkça alerjen taraması FASTA ve BLAST algoritmalarının daha fazla kullanımına doğru kaymaktadır [3;23;30] .

6 veya 8 amino asitin tam eşleşmesiyle alerjen tahmini yapan pekçok çalışma vardır. Bu çalışmaların sonucu FASTA veya diğer metodlarla kıyaslanmıştır [3; 27;30;40;64] . FASTA3 algoritmasının BLOSUM50 puanlama matrisiyle kullanıldığı bir çalışmada 6 amino asitin tam eşleşmesi durumunun alerjen çapraz reaktivite tahmini için kullanılamayacağını göstermiştir [30]. Bir başka çalışmada ise Swiss-Prot veritabanı içinde protein dizilim setinin tamamıyla yapılan bir karşılaştırmada 6 amino asit eşleşmesi durumunda proteinlerin %67si alerjen olarak tespit edilmiştir [64]. 12 amino asit eşleşmesi kullanıldığında ise bu oran %7'ye düşmüştür. Bu durum 8 amino asit eşleşmesi durumunda elde edilecek sonuçların tahmin doğruluğunda şüphe uyandırmaktadır.

FASTA veya BLAST taraması sonucu %70 den fazla genel benzerlik gösteren proteinlerin genellikle paylaşımlı IgE reaktivitesi gösterdiğinin klinik bulgularda da görüldüğünden bahseden çalışmayı [1] günümüz sonuçları da desteklemektedir. %40-50 veya daha az benzerlik olduğu durumlarda ise önemli ölçüde IgE veya alerjen çapraz reaktivite olmadığı gösterilmiştir [2;58;59;70].

Bir başka çalışmada FAO/WHO [21] kriteri (6 ardışık amino asit eşleşmesi) öncü bir tarama olarak kullanılmış, IgE bağlanan epitoplarla ilgili literatür taranmış, ve yanlış pozitif oranı azaltılarak antijen bölgelerinin teorik bir değerlendirmesi yapılmış [40], ancak bu birleşik yöntemin tahmin derecesinden bahsedilmemiştir. Antijenlik tahmini yapan algoritmaların antikor bağlanan epitoplar için yüksek tahmin oranları olduğu kanıtlanmamıştır [66]. Alerjen tahmininde motif tabanlı yöntem de önerilmiştir [64]. Kısa peptit dizilimi eşleşmesi yerine dizilim benzerliğine dayanan bu yöntemde bilinen alerjen dizilimleri motiflerle sınıflandırılmıştır. Dizilim benzerliği ve kısa peptit eşleşmesi Swiss-Prot veritabanından rastgele alınan protein dizilimleri kullanılarak karşılaştırılmıştır. 6 ardışık amino asit eşleşmeleri taranmış ve 200 proteinin en az bir alerjenle eşleştiği tespit edilmiştir. Fakat, bu 200 proteinden 199'unun alerjen olduğuna dair yayınlanmış bir kanıt yoktur. Aynı protein veri kümesi motif bulma metoduyla da değerlendirilmiş ve sonuçlar göstermiştir ki proteinlerin %90'ı yanlış bir şekilde alerjen olarak tahmin edilmiştir.

Geliştirilen bir başka motif tabanlı metodda da IgE bağlantı bölgelerinin tahmini, önceden belirlenmiş dizilimler ve yapılara dayanarak gerçekleştirilmiştir [36]. Her iki metod da [36;64] FASTA algoritmasının sonuçlarıyla kıyaslanmamıştır ancak büyük oranlarda dizilim benzerliğine dayandıkları için FASTA ve BLAST algorimalarının sonuçlarına yakın değerler bulacakları düşünülmektedir.

FASTA3 taramasını iyileştirmek için yapılan çalışmada ise en yakın komşuluk metodu kullanılmıştır [73]. Bu metodu geliştirmek için yapılan bir başka çalışmada doğru ve yanlış pozitif tespitlerinin istatistiki değerlendirmesi için BLOSUM50 ve BLOSUM80 puanlama matrisleri test edilmiştir [62]. Hemen hemen bütün alerjenlerin bulunduğu kapsamlı bir veritabanında yapılan FASTA taramasının,

uzunluklarının büyük kısımlarında %50 benzerlik olan çapraz reaktif proteinlerin tanınmasında çok etkili olduğuna dikkat çekilmektedir [1]. Bu yaklaşım hem basittir, hem de sonuçları kolayca irdelenebilir.

80 veya daha fazla amino asit üzerinde %35'lik bir eşleşme kriterinin [21] fazlasıyla yanlış pozitif veya yanlış negatif sonuçlar üretip üretmediği şüphelidir. Hiçbir tahmin aracı, bilgisayar programı, taraması veya algoritması %100 doğrulukla bir proteinin alerjen veya çapraz reaktif olup olamayacağını tahmin edemeyeceği göz önünde bulundurulmalıdır. Amaç her zaman için proteinlerin alerjen veya çapraz reaktif olabilecekleri düşünülerek tahminlerin yanında klinik ortamda IgE testleri de yapılmalıdır.

Her yaklaşımın kendi sınırlaması vardır. Örneğin, epitop tabanlı yaklaşımlar, bilinen epitopların sınırlı sayıda olmasından ve varolan epitop tahmin metodlarının düşük doğrulukta olmalarından dolayı başarısızlardır.

Bu çalışmada, K-En Yakın Komşu, Bulanık K-En Yakın Komşu ve Destek Vektör Makinesi (DVM) kullanılmıştır. Protein dizilimlerinin gösterilmesi için, amino asit bileşimi, dipeptit bileşimi, tripeptit bileşimi ve benzerlik skorları kullanılmıştır. Sonuçlar karşılaştırmalı olarak sunulmuştur. Üçüncü bölümde çalışmada yararlanılan veri kümesi verilmiştir. Veri kümesi bölümleme yöntemi anlatıldıktan sonra protein dizilimlerinin gösteriminde kullanılan yöntemler belirtilmiştir. Yapılan çalışmada kullanılan yöntemlerle ilgili bilgi verildikten sonra deney düzeneği başlığı altında performans ölçümlerinin değerlendirme kriterleri anlatılarak, çalışmada uygulanan yöntemler açıklanmıştır. Dördüncü bölümde yapılan çalışmaların sonuçları karşılaştırmalı olarak verilmiştir. Son olarak sonuçların ve yapılan çalışmanın değerlendirilmesi beşinci bölümde verilmiştir.

3. MATERYALLER ve YÖNTEMLER

3.1 Çalışmada Yararlanılan Veri Kümesi

Yapılan bu çalışmada yararlanılan veri kümesi (http://bioinformatics.uams.edu /mirror/algpred/algo.html) adresinden elde edilmiştir. Veri kümesi, 578 alerjen ve 700 alerjen olmayan (gıdadan türemiş) proteinlerden oluşan bir kümedir.

3.2 Veri Kümesi Bölümleme

Tahmin metodlarının geliştirilmesinde karşılaşılan problemlerden bir tanesi test için kullanılan proteinlerle eğitim için kullanılan proteinler arasındaki benzerliği en aza indirgemektir. Tekrarlılık yaratan veriyi kaldırmak eğitim için kullanılan protein sayısını azaltmaktadır ve bu da bir öğrenme metodu için istenilen birşey değildir. Bu çalışmada kullanılan veri kümesi için, toplam protein sayısını azaltmadan eğitim ve test için kullanılan proteinlerin benzerliklerini en aza indirgemek için farklı bir metod kullanılmıştır [72]. İlk olarak proteinler BLAST E-değeri 8E-4 (bir dizilim çifti için %26 eşleşme) kullanılarak gruplanmıştır. Bu gruplar beşerli kümeler şeklinde ayrılmıştır. Her kümede yaklaşık olarak aynı sayıda dizilim vardır ve verilen bir gruptaki bütün proteinler bir kümededir. Bir gruptaki dizilimler, diğer gruptakiler ile benzerlik göstermemektedir. Böylece bir küme ile diğer kümedeki dizilimler benzerlik göstermezler.

3.3 K-En Yakın Komşu

K-en yakın komşu algoritması eğiticili ve örnek tabanlı bir sınıflandırma algoritmasıdır. Uygulanabilirliği diğer yöntemlere göre daha kolaydır. Bu tip algoritmalarda eğitime ihtiyaç yoktur. İşleyişi, 'birbirine yakın olan nesneler muhtemelen aynı kategoriye aittir' diyen sezgisel fikir üzerine kuruludur. K-en yakın komşu algoritması, veri madenciliği, bilgi güvenliğinin sağlanmasında saldırı tespit sistemlerinde, genetik ve biyoenformatiğin birçok alanında, örüntü tanıma sistemleri gibi birçok benzeri sistemde kullanılmaktadır.

K-en yakın komşu algoritmasında bir vektörün sınıflandırılması, sınıfı bilinen vektörler kullanılarak yapılmaktadır. Test edilecek örnek, eğitim kümesindeki her

bir örnek ile tek tek işleme alınır. Test edilecek örneğin sınıfını belirlemek için eğitim kümesindeki o örneğe en yakın K adet örnek seçilir. Seçilen örneklerden oluşan küme içerisinde hangi sınıfa ait en çok örnek varsa test edilecek olan örnek bu sınıfa aittir denilir. Örnekler arası uzaklıklar 3.1 eşitliğinde verilen öklit (Euclidean) uzaklığı ile bulunur. Bu formülde x_i ve x_j vektörleri arasındaki uzaklık verilen iki vektörün karşılıklı koordinatlarının farklarının, karesinin toplamının karekökü alınarak bulunur.

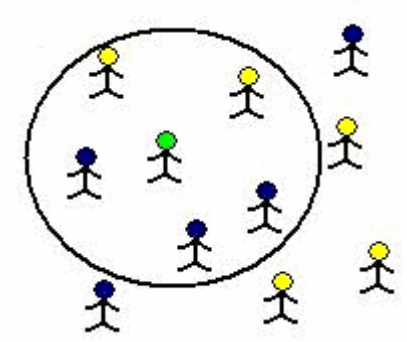
$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (dizi_r(x_i) - dizi_r(x_j))^2} \tag{3.1}$$

Öklit uzaklığı kullanılarak hesaplanan tüm uzaklık değerleri sıralanır. Sıralı değerler arasından K sayısına bağlı olarak en küçük K tanesi belirlenir. Test edilecek örneğe en yakın K tane komşu örnek belirlenmiş olur. Test edilecek örneğin sınıflandırılması için bulunan K tane komşunun sınıf etiketleri kullanılır. Sınıf etiketlerinin "+" ve "-" olarak belirlendiği varsayılırsa; test edilecek örnek ile eğitim örneklerinin arasındaki uzaklık değerleri hesaplandıktan sonra, K tane en yakın örneğin sınıf etiketlerine bakılır. Sınıf etiketi "+" olanlar "-" olanlardan fazla ise test örneğinin sınıfı "+"'dır, tersi durumda da "-" olarak sınıflandırılır. Test örneğinin sınıfına karar verilmesi aşamasında K değerinin seçilmesine bağlı olarak iki durum yaşanabilir. Birinci durumda K değeri tek sayı seçilerek "+" ve "-" örneklerin sayısının eşit değerde çıkması önlenir. K değeri çift sayı seçilirse de K tane örnek için her bir sınıfa ait örnekler kendi aralarında toplanır ve ortalamaları bulunur. En küçük ortalamaya sahip olan sınıf, test edilecek örneğe daha yakın olacağı için test örneğinin sınıfı en küçük ortalamaya sahip olan sınıf olacaktır. Bu algoritma için sınıf sayısında bir kısıtlama yoktur. İstenilen sayıda sınıf belirlenerek (en az bir sınıf olacak şekilde) sınıflandırma işlemi yapılabilir. Algoritma aşağıda verilen şekildedir.

```
Başla
     Test edilecek örnek vektör y'yi gir.
     K değerini 1<=K<=n olacak şekilde seç
     i=1 olarak başlat
```

```
    Tekrarla (k-en yakın komşular bulunana kadar)
           y'den x_i'ye olan uzaklığı hesapla
           Eğer (i<=k)
               x_i'yi k-en yakın komşular kümesine dahil et
           Eğer Değilse (x_i y'ye daha önceki en yakın
               komşudan daha yakınsa)
               k-en yakın komşular kümesinden en uzağını sil
               x_i'yi k-en yakın komşular kümesine dahil et
           i'yi artır
    i>n ise, döngüden çık.
    k-en yakın komşu kümesinde temsil edilen çoğunluk
    sınıfını belirle
    Eğer(sınıf sayıları eşit ise)
         her sınıftaki komşuların uzaklıklarının toplamını
         hesapla
         Eğer (sınıflar için toplam değerleri eşit değilse)
              y'yi minimum toplam içinde sınıfla
         Değilse
              y'yi son bulunan minimum sınıf içinde sınıfla
    Değilse
         y'yi çoğunluk sınıfında sınıfla
Bitir
```

K-en yakın komşu algoritması herkes tarafından bilinen "bana arkadaşını söyle sana kim olduğunu söyleyeyim" sözü ile örneklenebilir. Şekil 3.3.1'de yeşil çöp adamın test edilecek örnek olduğu varsayılmıştır ve K değeri 5 olarak seçilmiştir. Yeşil adama en yakın 5 arkadaşı yeşil adamın, mavi sınıfa ya da sarı sınıfa ait olduğunu belirleyecektir. Yeşil adamın en yakın beş arkadaşı işaretlenen daire ile belirlenmiştir. Bu dairenin içerisinde kalan beş arkadaşının üç tanesi mavi sınıftan iki tanesi ise sarı sınıftandır. Bu sonuçlara göre yeşil adam mavi sınıftandır sınıflandırılması yapılır.



Şekil 3.3.1 Yeşil Çöp Adam

3.4 Bulanık K- En Yakın Komşu

Bulanık mantık kavramı ilk kez 1965 yılında Prof. Lotfi Asker Zadeh tarafından ortaya atılmıştır. Bu mantık insanların günlük hayatta kullandıkları ifadelerin, bilgisayar ortamında matematiksel olarak modellenmesi prensibine dayalı bir yaklaşımdır. Klasik yöntemlerde 1 ya da 0'larla gösterilen kararların, keskin geçişleri, bulanık mantık ile yumuşatılmış ve ara değerlerle ifade edilebilir hale gelmiştir. İnsan hayatına bakıldığında ılık, yarı açık, yarı dolu gibi çok sık kullanılan kavramların bulanık mantık ile bilgisayar ortamında da ifade edilebilmesiyle problemlerin çözümü gerçekleştirilmiştir.

Bulanık k-en yakın komşuluk yöntemi de K-En Yakın Komşu yöntemi gibi bir sınıflandırma algoritması olmasına rağmen sonuçlarının ifadesi itibariyle K-En Yakın Komşu yönteminden ayrılır. Bulanık k en yakın komşuluk algoritması, vektörü belirli bir sınıfa atamak yerine, sınıf üyeliğini bir örnek vektöre atar. Bir elemanın bir kümeye veya bir sınıfa ait olması klasik küme kavramında ya aittir (üyelik=1) veya ait değildir (üyelik=0) şeklinde tanımlanmaktadır. Gerçekte bir eleman bir kümeye ne tam aittir ne de değildir. Yani bu elemanın o küme veya sınıf için bir aitlik derecesi (üyelik değeri) olmalıdır. Bu üyelik değeri 0 ile 1 arasında sonsuz değer alabilmektedir. Bulanık algoritmalarda, test edilecek örnek sınıflanırken örneğin sınıfını belirlemenin yanında o sınıfa ne kadar ait olduğuna dair bir bilgi de verilmektedir. Bu bilgi, örneğin o sınıfa olan üyelik değeri olmaktadır. Bulanık K-En Yakın Komşu yönteminin, K-En Yakın Komşu yöntemine göre avantajı Bulanık K-En Yakın Komşu yönteminin daha fazla bilgi içermesidir.

Bulanık K-En Yakın Komşu yönteminde test edilecek örnek için belirlenen üyelik değerleri sonuçta oluşan sınıflandırma için bir güvence seviyesi sunar. Örnek olarak, elimizde iki adet sınıf olduğunu düşünürsek, test edilecek örnek için üyelik değerlerinin bir sınıfa 0.89 üyelik değeri ile ve diğer sınıfa 0.11 üyelik değeri ile üyelik değerlerinin hesaplandığı varsayılırsa; 0.89 üyelik değerinin belirlendiği sınıfın test edilecek örneğin sınıfı olduğu kararına üyelik değerlerinin sayılarına bakarak kolayca karar verebiliriz. Farklı bir örnek için; elimizde üç adet sınıfın olduğu varsayılırsa, eğer test edilecek örneğin birinci sınıfa üyeliği 0.55 üyelik değeri, ikinci sınıfa üyeliği 0.44 üyelik değeri ve üçüncü sınıfa üyeliği 0.01 üyelik değeri olarak hesaplanmış ise test edilecek örneğin sınıflandırılmasında kesin bir karara varmak mümkün olmayabilir. Fakat, üçüncü sınıfa ait olmadığı konusunda da emin olabiliriz. Böyle bir durumda, sınıfının anlaşılabilmesi için test edilecek örnek için farklı yöntemler de denenerek incelenebilir, çünkü test edilecek örnek her iki sınıfta da yüksek üyelik derecesi göstermektedir.

Bulanık K-En Yakın Komşu yönteminin temeli, test edilecek örneğin k-en yakın komşularına olan uzaklığı ve bu komşuların olası sınıflara üyelikleri cinsinden bir fonksiyon olarak üyelik atamasıdır. Bulanık K-En Yakın Komşu yöntemi, K-En Yakın Komşu yöntemine şu yönde benzer; Bulanık K-En Yakın Komşu yöntemi aynı zamanda K-En Yakın Komşu yönteminde olduğu gibi, k-en yakın komşuları bulmak zorundadır. Bu K örneklerini elde etmenin ötesinde bu yöntemler sınıflandırma işlemi için önemli farklılıklar gösterirler.

$W=\{x_1, x_2,\ldots, x_n\}$ n adet eğitim sınıfında yeralan sınıfları belirlenmiş örneğin kümesi olsun. Aynı zamanda, $u_i(x)$; x vektörünün üyeliği (hesaplanacak olan), ve u_{ij} de sınıfı belirlenmiş örnek kümesinin j'ninci vektörünün i'ninci sınıf içindeki üyeliği olsun. Algoritma şu şekildedir:

```
Başla
     Test edilecek örnek vektor x'i gir.
     K değerini 1<=K<=n olacak şekilde seç
     i=1 olarak başlat
     Tekrarla (x'in k-en yakın komşuları bulunana kadar)
          x'den xi'ye olan uzaklığı hesapla
```

```
          Eğer (i<=k)
               xi'yi k-en yakın komşular kümesine dahil et
          Eğer Değilse (xi x'e daha önceki en yakın komşudan
          daha yakınsa)
               k-en yakın komşulardan en uzağını sil
               xi'yi k-en yakın komşular kümesine dahil et
     i'yi artır
     i>n ise, döngüden çık.
i'ye bir değerini ata
     Tekrarla (x'in her sınıftaki üyeliği atanana kadar)
          (3.2)'yi kullanarak ui'yi hesapla
          i'yi artır
     döngüden çık
Bitir
```

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)}{\sum_{j=1}^{k}\left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)} \qquad (3.2)$$

Eşitlik 3.2'de görüldüğü gibi x'in atanmış üyelikleri en yakın komşulara olan uzaklığın tersiyle ve onların sınıf üyeliklerinden etkilenmektedir. Ters uzaklık bir vektörün üyeliğine, vektörden olan uzaklık azsa daha fazla ağırlık verir, uzaklık çoksa daha az ağırlık verir. Sınıfı bilinen örnekler için çok çeşitli şekilde üyelikler atanabilir. Bilinen sınıflarında tam üyelik verilip diğer bütün sınıflarda hiç üyelik verilmeyebilir. Bu yöntemin yerine bulanık temelli alternatif metodların da kullanılması mümkündür.

m değişkeni her bir komşunun üyelik değerine katkısını hesaplarken mesafenin ne kadar ağırlıkta verildiğini belirler. m değeri 2 ise her komşu noktanın katkısı, sınıflanan noktadan olan uzaklığın karşıtıyla ağırlıklandırılır. m arttıkça komşular daha eşit bir şekilde ağırlıklandırılır, ve sınıflandırılan noktadan olan göreceli uzaklıklarının etkileri daha az olur. m değeri 1'e yaklaştıkça yakın olan komşular uzaktaki komşulardan çok daha fazla ağırlıklandırılırlar ve bunun etkisi sınıflanan

noktanın üyelik değerine katkıda bulunan nokta sayısını azaltan şekilde olmaktadır.

Sonuç olarak; Bulanık K-en yakın komşu algoritmasının, K-en yakın komşu algortimasından en büyük farkı bilinmeyen bir test örneğini sınıflamak yerine, test örneğinin belirli sınıfa ne kadar ait olduğu sorusuna yönelik verdiği cevaptır. Bulanık K-en yakın komşu algoritmasında ise örnek için bir sınıfa ait olma ya da olmama bilgisine ek olarak o sınıfa ne kadar ait olduğuna ilişkin değeri de hesaplanır. Bu değer kullanılarak örneğin sınıflandırılması yapılır.

3.5 Destek Vektör Makineleri (DVM)

Destek vektör makinelerinin temelleri İstatistiksel Öğrenme Kuramına göre V. Vapnik tarafından atılmıştır [60]. DVM'ler iki sınıflı bir sınıflandırma ve uyumlama (regresyon) metodu olup sağlam ve etkin bir yöntem olarak kullanılmaktadır [17;68]. El yazısı tanıma, ses ve yüz tanıma,, biyoenformatik-gen ve protein sınıflandırması, kanser hücrelerinin tanınması, ve uzaysal veri analizi gibi birçok alanda kullanılmaktadır [11;12;14;15;26;32;34;35;41;44;68].

Pozitif ve negatif örneklerin ayırt edilmesinde DVM'ler kullanılır. DVM'ler, kullanılabilir örneklerden (eğitim aşamaları), yeni nesneleri doğru bir şekilde sınıflandırma (test aşamaları) işlemini gerçekleştirir.. ilk olarak DVM'ler verinin daha iyi ayırt edilebileceği şekilde, yüksek boyutlu girdi uzayı doğrusal olmayan bir şekilde daha yüksek boyutlu öznitelik uzayına eşlenir.

Başlangıçta sınıflandırma için geliştirilen DVM'ler, sonraları uyumlama için sınıflandırmaya benzer olarak genişletilmiştir. Yapısal risk minimizasyonu prensibine dayanır, yani beklenen riskin üst sınırı küçük tutulmaya çalışılır [69]. DVM'ler deneysel ve yapısal risklerin her ikisini de en az olacak şekilde eğitilirler. DVM'lerin tasarımında genelleme hatası için verilen bir üst sınır minimuma indirgenir. Doğrusal olmayan örnek uzayının, örneklerin doğrusal olarak ayrılabileceği bir yüksek boyuta aktarılmasıyla, örnekler arasındaki en büyük sınır bulunur.

DVM uygulamaları diger geleneksel metodlardan daha iyi sonuçlar vermektedir. Çok kullanışlı olan bu öğrenme yöntemi basit fikirler üzerine kurulmuş olması ve yüksek performans isteyen uygulamalarda kullanabildiklerinden dolayı çok avantajlıdırlar.

Pratikte karşılaşılan uygulamalar karmaşıktır ve teorik olarak çözülmesi zordur. Verilerin bir bölümü doğrusal olarak ayrılabilen bir yapıdayken bir bölümü de doğrusal olarak ayrılamayabilir. DVM yöntemi bu zorlukları ortadan kaldırarak oldukça karmaşık olan problemlere çözüm getirir.

DVM ile bulunan fonksiyon, veriye yakınlık ve çözümün karmaşıklığı arasındaki bir geçiştir. Sadece iki sınıfın bulunduğu bir sınıflandırma probleminde DVM iki sınıf arasındaki sınırı maksimize eden optimal ayırt etme yüzeyini belirler, yani eğitim kümesi ile ayırt etme yüzeyine en yakın noktaların arasındaki mesafeyi maksimize eder.

DVM'lerde dönüşüm düşük boyutlu bir giriş uzayından alınan vektörler yüksek boyutlu bir diğer uzaya doğrusal olmayan bir biçimde taşınarak yapılır. Bu dönüşümü belirleyen bir çekirdek (kernel) ile dönüşümü uygulayan sistem, makine veya ağ, tanımlanır. Sınıflama yapılırken yüksek boyutlu uzaya taşınan vektörler doğrusal olarak ayrılabilir duruma gelir. Ayrıştıran düzlemler içerisinde sınıflara uzaklığı en çok olan en uygun doğrusal ayrıştırıcı olarak belirlenir. Yüzeye en yakın vektörler belirlenerek en yakın uzaklık tespit edilir. Destek vektörler olarak adlandırılan bu vektörler ayrıştırıcı düzlemi belirlerler.

Sürekli geliştirilen DVM'lerin yaygın bir şekilde pek çok alanda kullanılmalarına rağmen bazı eksik yanları bulunmaktadır. DVM'ler öncelikle veriyi iki sınıfa ayrıştırmak için tasarlanmıştır [16;67]. Bu yüzden de çok sınıflı ayrıştırmalarda etkili olmamaktadır ve bu konuda çok sayıda çalışma yapılmaktadır [32;49;54]. Çok sınıflı problemlere doğrudan çözüm öneren formülasyonların başarıları genelde iyi değildir [71]. DVM'lerin gürültü ve aykırı verilere olan hassasiyeti bir başka eksik taraftır [33]. Ayrıca, hesaplama ve bellek gereksinimi çok fazla olduğu için çözüm çok yavaştır [16, 47]. Veri kümeleri büyüdükçe DVM'lerin uygulanması da sınırlı olmaktadır. Bunlardan başka çekirdek ve parametre seçiminde bazı

problemler ortaya çıkmaktadır [20;47;72]. Uygun çekirdek ve parametresi seçilmezse, boyutu yüksek olan uzaydaki uzaklık sırası korunmaz veya uzaklıklar arası farklar küçülür ve sınıflama hatalı olur. Bu hatayı gidermek üzere yapılan çalışmalar da bulunmaktadır [5]. Bunlardan başka bazı tasarım yöntemlerinde kullanılan penaltı katsayısının sonucu çok etkilediği saptanmış ve bir başka problem olarak tespit edilmiştir [60].

Doğrusal ve doğrusal olmayan destek vektör makineleri olmak üzere iki tip DVM vardir. Bundan sonraki bu DVM tiplerinden bahsedilmiştir.

3.5.1 Doğrusal destek vektör makineleri

Doğrusal DVM'ler de kendi içinde verilerin doğrusal olarak ayrılabilme ya da ayrılamama durumlarına göre ayırılırlar.

Doğrusal ayrılabilme durumu

Doğrusal ayrılabilir durumlarda sınıfları birbirinden ayıran pek çok karar düzlemi bulunabilir. DVM bu düzlemlerden iki sınıf arasındaki mesafeyi en büyük yapanını tespit eder. Bu düzleme en yakın vektörlere de destek vektörleri denir.
Farzedelim ki eğitim için kullanılacak N elemanlı veri aşağıdaki şekilde olsun:

$$X = \{xi,\ yi\}, \quad i = 1,\ 2,\ ....,\ N$$

Burada $y_i \in \{-1,1\}$ sınıf etiket değerleri ve $x_i \in R^d$ de özellikler vektörüdür (x_i giriş vektörü, y_i çıkış vektörü). Doğrusal olarak ayrılabilir durumlarda bu veriler ayırıcı hiperdüzlem denilen bir düzlemle doğrudan ayrılabilirler. DVM'nin amacı bu hiperdüzlemin iki gruba da eşit uzaklıkta olmasını sağlamaktır. DVM ilk işlem olarak doğrusal olarak ayrılamayan verileri, verinin özelliklerinin boyutundan daha büyük derecede ki yüksek boyutlu bir hale dönüştürür. Bu işlem verilerin bir hiperdüzlem ile ayrılmasını sağlamak üzere yapılır.
Hiperdüzlem üzerindeki herhangi bir x noktası,

$$w \cdot x + b = 0 \tag{3.3}$$

Eşitlik 3.3'te verilen koşulu sağlar. Burada w hiperdüzlemin normal vektörü ve |b|/||w|| hiperuzayın orijine dik uzaklığıdır. DVM metodu doğrusal olarak ayrılabilen durumlarda, y_i = +1 ve y_i = -1 şeklinde etiketlenmiş örneklere eşit uzaklıkta olan optimum ayırıcı hiperdüzlemin bulunmasını sağlar. Bunun için eğitim seti eşitsizlik 3.4 ve 3.5'i sağlamalıdır [68]:

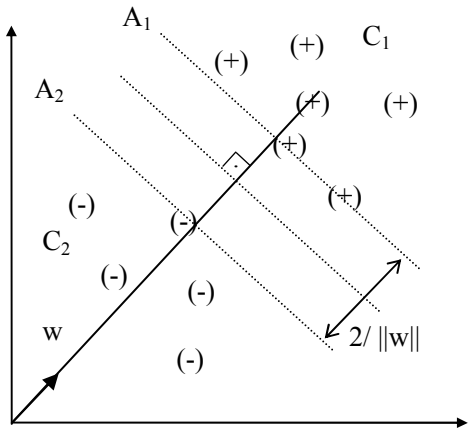
$$y_i = +1 \text{ için, } w\,x_i + b \geq +1 \tag{3.4}$$

$$y_i = -1 \text{ için, } w\,x_i + b \leq -1 \tag{3.5}$$

Bu eşitsizlikler bir arada ifade edilirse, $i = 1,\ 2,\ .....,\ N$ için

$$y_i(w\,x_i + b) \geq +1 \tag{3.6}$$

olur. Eşitsizlik 3.6'yı sağlayan hiperdüzlemin iki tarafındaki en yakın örneklere olan dik uzaklıkları toplamı sınır olarak adlandırılır. Sınırı maksimum yapan hiperdüzlem optimum ayırıcı hiperdüzlemdir (Şekil 3.5.1.1). Optimum ayırıcı hiperdüzlemi bulmak için uygun w ve b değerleri hesaplanır.



Şekil 3.5.1.1 Doğrusal Ayrılabilme Durumunda Optimum Ayırıcı Hiperdüzlem

Şekil 3.5.1.1'de C_1ve C_2 sınıflarını ayıran birbirine paralel A_1 ve A_2 hiperdüzlemleri gösterilmiştir. C_1 sınıfını ayıran A_1 hiperdüzlemini oluşturan eşitsizlik (3.4) eşitsizliği ile, C_2 sınıfını ayıran A_2 hiperdüzlemini oluşturan eşitsizlik ise (3.5) eşitsizliği ile tanımlanmıştır. Bu durumda A_1 hiperdüzleminin orijine olan uzaklığı

|1-b|/||w|| ve A_2 hiperdüzleminin orijine olan uzaklığı |-1-b|/||w|| olmaktadır. Bu iki hiperdüzlemin optimal hiperdüzleme uzaklıkları ise 1/||w|| kadardır, yani iki örnek kümesi arasındaki uzaklık 2/||w|| kadardır. Bu iki hiperdüzlem arasındaki maksimum uzaklık ise en küçük ||w|| değerinin tespitiyle bulunabilir. DVM yöntemiyle yapılmaya çalışılan bu iki hiperdüzlemin arasındaki uzaklığın (sınırın) maksimum olmasını sağlamaktır.

Burada A_1 ve A_2 hiperdüzlemleri arasında eğitim verilerine ait hiçbir örneğin olmadığına da dikkat çekilir.

Maksimum sınırın bulunması için $\frac{1}{2}\|w\|^2$ ifadesinin en küçük değeri şu koşulla beraber bulunmalıdır:

$$y_i(w\,x_i + b) \geq +1, \quad \forall i \tag{3.7}$$

Bu problem ikinci dereceden optimizasyon problemidir ve çözümü için problemin Lagrange formülasyonu yapılır. Bunun yapılması iki yönden kolaylık sağlar. Birincisi Lagrange formülasyonu yapılarak Lagrange çarpanlarının hesaplanması daha kolaydır. İkincisi ise doğrusal ayrılamayan durumlar için de genelleştirilmesi bakımından daha uygundur [14]. Problemin Lagrange formülasyonu eşitlik 3.8'de verilmiştir.

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

Eşitlik 3.8'de verilen α_i değerleri pozitif Lagrange çarpanlarıdır. Ancak 3.8'de ifade edilen formülasyonun çözülmesi oldukça karmaşıktır ve Karush-Kuhn-Tucker (KKT) koşulları kullanılarak dual problemine dönüştürülerek çözülebilir.Çözüm için gerekli KKT koşullar eşitlik 3.9 ve 3.10'da verilmiştir.

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

$$\frac{\partial L_p}{\partial b} = 0 \quad \Rightarrow \quad \sum_i \alpha_i y_i = 0 \tag{3.10}$$

Bu koşullar 3.8'de yerlerine yazılırsa eşitlik 3.11 ve 3.12 elde edilir.

$$L_d = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{3.11}$$

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

Görüldüğü gibi her eğitim örneği için bir tane Lagrange çarpanı vardır. Çözümde elde edilen Lagrange çarpanlarının büyük bir kısmının değeri sıfır olacak ve geriye kalan pozitif α_i değerli x_i vektörleri destek vektörleridir. Bu vektörler A_1 veya A_2 hiperdüzlemlerinin üzerindedirler. Lagrange çarpanı sıfır olan örnekler ise A_1 veya A_2 hiperdüzlemlerinin gerisinde kalan örneklerdir.

Doğrusal ayrılamama durumu

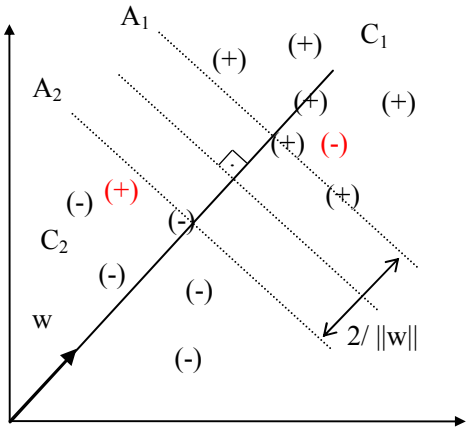
Bir önceki bölümde belirtilen işlemler eğitim örneklerinin ayrılabilir olması durumunda kullanılabilmektedir. Örneklerin doğrusal olarak ayrılabilir durumda olmadığı durumlarda problemin çözümü için eğitim hatasının sapması olan $\xi_i, \ i = 1, \ 2, \ ....., \ N$ kullanılır [17]. Eşitsizlik 3.4 ve 3.5'teki koşulları bu sapmalar ile yeniden tanımlayacak olursak 3.13, 3.14 ve 3.15 eşitsizlikleri elde edilir.

$$y_i = +1 \ \text{için}, \ w x_i + b \geq +1 - \xi_i \tag{3.13}$$

$$y_i = -1 \ \text{için}, \ w x_i + b \leq -1 + \xi_i \tag{3.14}$$

$$\xi_i \geq 0, \ \forall i \tag{3.15}$$

x_i örneğinin yanlış sınıflandırılmış olması için $\xi_i \geq 1$ olmalıdır. Diğer durumlarda doğru sınıflandırılmış fakat $0 < \xi_i < 1$ olması durumunda Şekil 3.5.1.2'deki A_1 ve A_2 hiperdüzlemleri arasında yer alıyor demektir.



Şekil 3.5.1.2 Doğrusal Ayrılamama Durumunda Optimum Ayırıcı Hiperdüzlem

Doğrusal olarak ayrılamama durumunda sisteme *C* üst sınırı eklenir. Lagrange çarpanlarının alabilecekleri maksimum değeri olan bu üst sınır, bu çarpanların $0 \le \alpha_i \le C$ aralığında kalmasını sağlar. Bu durumda Lagrange formülasyonu olarak eşitlik 3.16 ifadesi elde edilir.

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_i \xi_i - \sum_i \alpha_i \{y_i (w\, x_i + b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i \qquad (3.16)$$

Burada μ_i değerleri, ξ_i'nin pozitif olmasını sağlamak için kullanılmış olan Lagrange parametreleridir. Bu formülasyonun da çözülmesi zordur ve doğrusal ayrılabilir problemlere benzer şekilde dönüşümler yapılarak çözülür. Benzer şekilde bu probleme de KKT şartları uygulanarak;

$$\frac{\partial L_p}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \qquad (3.17)$$

$$\frac{\partial L_p}{\partial b} = -\sum_i \alpha_i y_i = 0 \qquad (3.18)$$

$$\frac{\partial L_p}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \qquad (3.19)$$

eşitsizlik 3.17, 3.18 ve 3.19'da verilen ifadeleri elde edilir. Bunlar eşitlik 3.16'ya uygulanarak;

$$L_d = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{3.20}$$

$$0 \le \alpha_i \le C, \quad \forall i \tag{3.21}$$

Eşitlik 3.20 ve 3.21'deki ifadeler elde edilir. $0 < \alpha_i < C$ aralığında yer alan Lagrange çarpanlarına karşılık gelen x_i vektörleri destek vektörleridir.

3.5.2 Doğrusal olmayan destek vektör makineleri

Doğrusal olmayan problemlerde çekirdek fonksiyonları kullanılarak örnekler daha yüksek boyutlu ve doğrusal olarak ayrılabilecekleri bir uzaya taşınır ve çözüm bu yeni uzayda yapılır. Farzedelim ki örnekleri farklı bir H Öklit uzayına taşıyan Φfonksiyonu eşitsizlik 3.22'de verilmiştir.

$$\Phi : R^d \mapsto H \tag{3.22}$$

H uzayındaki verilerin $\Phi(x_i) \cdot \Phi(x_j)$ iç çarpımını K ile ifade edersek;

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_i) \tag{3.23}$$

Eşitlik 3.23'teki fonksiyonunu elde ederiz. Bu durumda DVM'nin eğitim algoritması sadece H uzayındaki verilerin bu iç çarpımına bağlı olur.
Eşitsizlik 3.23'te tanımlanan *K* fonksiyonu, çekirdek fonksiyonu olarak adlandırılmaktadır. Test aşamasında ise sistemin test örneğinin alacağı değer f(x) fonksiyonu (3.24) ile belirlenir:

$$f(x) = \sum_{i=1}^{N_v} \alpha_i y_i \Phi(v_i) \cdot \Phi(x) + b = \sum_{i=1}^{N_v} \alpha_i y_i K(v_i, x) + b \tag{3.24}$$

Buradaki N_v ve v_i destek vektörlerin sayısı ve destek vektörlerin kendileridir.

3.5.3 Çok sınıflı destek vektör makineleri

Örnekler her zaman yukarıda anlatılan DVM'lerde olduğu gibi iki sınıflı değildir. İkiden fazla sınıf olduğu durumlarda DVM tekniğinde temel olarak kullanılan iki yaklaşım vardır. Bunlardan birincisi Lagrange fonksiyonunun kullanımıdır. Bu fonksiyon çok sınıfla işlem yapacak hale getirilerek problem çözülür. Ancak sınıf sayısı arttıkça hatalar da buna paralel olarak çok arttığı için bu yöntem fazla tercih edilmemektedir.

İkinci yaklaşımda da DVM, sınıf sayısına göre ikili sınıflandırmalar yapacak şekilde çalıştırılır. Bu amaçla kullanılan yöntemlerden ikisi, bire karşı bir ve bire karşı hepsidir.

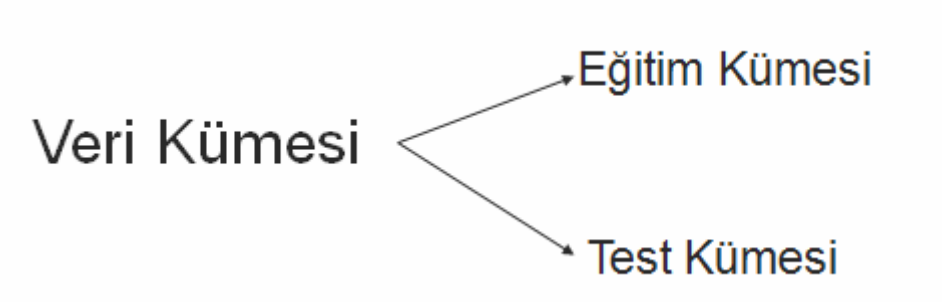
Bire karşı bir yönteminde her örnek kümesi diğer örnek kümeleriyle ayrı ayrı eğitilir. Yani n sınıf varsa n(n-1)/2 tane eğitim işlemi yapılır. Daha sonra eğitim işlemlerinin sonucunda bulunan destek vektörlerle test aşamasında gelen örnek kıyaslanır ve sınıfı bulunur.

Bire karşı hepsi yönteminde ise her örnek kümesinin eğitimi, diğer bütün örneklerin aynı kümeye ait olduğu varsayımı ile yapılır. Yani n farklı sınıf var ise n tane eğitim işlemi yapılır. Test aşamasında da örneğin ait olduğu sınıf bu eğitim verilerinde elde edilen destek vektörler ile kıyaslanmasıyla tespit edilir.

3.6 Deney Düzeneği ve Yapılan Çalışmalar

3.6.1 Veri kümesi

Bu çalışmada kullanılan veri kümesi protein dizilimlerinden oluşmaktadır. Veri kümesi içerisinde Şekil 3.6.1.1'de verildiği şekilde eğitim kümesi ve test kümesi vardır.



Şekil 3.6.1.1- Veri Kümesi

Şekil 3.6.1.2'de veri kümesine ait bir bölüm gösterilmiştir. Şekilde görüldüğü gibi her bir dizilim bir satır ile ifade edilmektedir. Proteinin alerjen olup olmadığı satır başında belirtilen ifade ile gösterilmektedir(>Allergenx).

```
>Allergen1
MDSPCLIAVLVFLCAIVSCYSDNPIDSCWRGDSNWDQNRMKLADCAVGFGSSTMGGKGGDIYTVTSAEDNPVNPTPGTLRYGATREKALWIIFSQNMNIKLKMPLYVAGHKTIDGRGADVHLGNGGPCLFMRKVSHVILHGLHIHGCNTS
SESIGVEPVHAQDGDAITMRNVTNAWIDHNSLSDCSDGLIDVTLSSTGITISNNHFFNHHKVMLLGHDDTYDDDKSMKVTVAFNQFGPNAGQRMPRARYGLVHVANNNYDQWNIYAIGGSSNPTILSEGNSFAAPNENYKKEVTKRIGCV
WVWRSTRDAFSNGAYFVSSGKTEETNIYNSNEAFKVENGNLAPQLTKNAGVVA
>Allergen2
MDCIRILWSVAVGLLLVSWRPTMFAASPTCDTVQNILAPCAGFLTGQEPSKACCTGVNNLNNSRKTKADRVAVCNCIKELTKSIAYDPKRMPLLSTKCGVKPDFPAVDKNLDCSKLPV
>Allergen7
MSSSQLEQAITDLINLFHKYSGSDDTIEKEDLLRLMKDNFPNFLGACEKRGRDYLSNIFEKQDKNKDRKIDFSEFLSLLADIATDYHNHSHGAQLCSGGNQ
>Allergen11
VDYCKIKCSSGIHTVCQYGESTKPSKNCADKVIKSVGPTEEEKKLIVNEHNRFRQKVAQGLETRGNPGPQPAASDMNNLVWNDELAHIAQVWASQCQILVHDKCRNTAKYQVGQNIAYAGGSKLPDVVSLIKLWNEVKDFNYNKGITKC
YTQMIWAKTKEIGCGSLKYMKNNMQHHYLICNYGPAGNYLGQLPYTKK
>Allergen13
LSVCFLILFHGCLASRQEWQQQDECQIDRLDALEPDNRVEYEAGTVEAWDPNHEQFRCAGVALVRHTIQPNGLLLPQYSNAPQLIYVVQGEGMTGISYPGCPETYQAPQQGRQQGQSGRFQDRHQKIRRFRRGDIIAIPAGVAHWCYNEG
TLLDVSNSQNQLDRTPRKFHLAGNPKDVFQQQQQHQSRGRNLFSGFDTELLAEAFQVDERLIKQLKSEDNRGGIVKVKDDELRVIRPSRSQSERGSESEEESEDEKRRWGQRDNGIEETICTMRLKENINDPARADIYTPEVGRLTTLNS
WLQLSVEKGVLYKNALVLPHWNLNSHSIIYGCKGKGQVQVVDNFGNRVFDGEVREGQMLVVPQNFAVVKRAREERFEWISFKTNDRAMTSPLAGRTSVLGGMPEEVLANAFQISREDARKIKFNNQQTTLTSGESSHHMRDDA
>Allergen14
MAGLMKLACLVLACMIVAGPITSKAALSCGTVNTNVAACIGYLTVGALPRACCTGVSKLNSIARTTPDRQQACRCLKTAASALGSGLNAGRAAGLPKACGVNVPFPISLLTRCINCNSVK
>Allergen15
MADAEHERIFKKFDTDGDGKISAAELEEALKKLGSVTPDDVTRMMAKIDTDGDGNISFQEFTEFASANPGLMKDVAKVF
>Allergen18
MRGRVSPLMLLLGILVLASVSATQAKSPYRKTENPCAQRCLQSCQQEPDDLKQKACESRCTKLEYDPRCVYDTGATNQRHPPGERTRGRQPGDYDDDRRQPRREEGGRWGPAEPRERERREEDWRQPREDWRRPSHQQPRKIRPEGREGEC
EVREETSRNNPFYFPSRRFSTRYGNQNGRIRVLQRFDQRSKQFQNLQNHRIVQIEARPNTLVLPKHADADNILVIQQGQATVTVANGNNRKSFNLDEGHALRIPSGFISYILNRHDNQNLRVAKISMPVNTPGQFEDFFPASSRDQSSYL
LEAAFNAEFNEIRRVLLEENAGGEQEERGQRRRSTRSSDNEGVIVKVSKEHVQELTKHAKSVSKKGSEEEDITNPINLRDGEPDLSNNFGRLFEVKPDKKNPQLQDLDMMLTCVEIKEGALMLPHFNSKAMVIVVVNKGTGNLELVAVRK
REQEWEEEEEDEEEEGSNREVRRYTARLKEGDVFIMPAAHPVAINASSELHLLGFGINAENNHRIFLAGDKDNVIDQIEKQAKDLAFPGSGEQVEKLIKNQRESHFVSARPQSQSPSSPEKEDQEEENQGGKGPLLSILKAFN
>Allergen22
KVNYCKIKCLKGGVHTACKYGTSTKPNCGKMVVKAYGLTEAEKQEILKVHNDFRQKVAKGLETRGNPGPQPPAKNMNNLVWNDELANIAQVWASQCNYGHDTCKDTEKYPVGQNIAKRSTTAALFDSPGKLVKMWENEVKDFNPNIEWSK
HYTQMVWAKTKEIGCGSVKYVKDEWYTHYLVCNYGPSGNFRNEKLYEKK
>Allergen27
MDSPCLIAVLVFLCAIVSCYSDNPIDSCWRGDSNWDQNRMKLADCAVGFGSSTMGGKGGDIYTATSAEDNPVNPTPGTLRYGATREKALWIIFSQNMNIKLKMPLYVAGHKTIDGRGADVHLGNGGPCLFMRKVSHVILHGLHIHGCNTS
SESIGVEPVHAQDGDAITMRNVTNAWIDHNSLSDCSDGLIDVTLSSTGITISNNHFFNHHKVMLLGHDDTYDDDKSMKVTVAFNQFGPNAGQRMPRARYGLVHVANNNYDQWNIYAIGGSSNPTILSEGNSFAAPNENYKKEVTKRIGCE
WVWRSTRDAFSNGAYFVSSGKTEETNIYNSNEAFKVENGNLAPQLTKNAGVVA
```

Şekil 3.6.1.2- Veri Kümesi İçeriği

Benzer şekilde hazırlanmış beş test kümesi ve beş eğitim kümesi mevcuttur. Eğitim 1 kümesinde 1020 adet protein diziliminin 460 tanesini alerjen proteinler, 560 tanesi ise alerjen olmayan proteinler oluşturmaktadır. Test 1 kümesinde 254 adet protein dizilimi vardır. Bunlardan 114 tanesi alerjen 140 tanesi alerjen olmayan proteinlerdir. Eğitim 2 kümesinde 1019 adet protein diziliminin 459 tanesini alerjen proteinler, 560 tanesi ise alerjen olmayan proteinler oluşturmaktadır. Test 2 kümesinde 255 adet protein dizilimi vardır. Bunlardan 115 tanesi alerjen 140 tanesi alerjen olmayan proteinlerdir. Eğitim 3 kümesinde 1019 adet protein diziliminin 459 tanesini alerjen proteinler, 560 tanesi ise alerjen olmayan proteinler oluşturmaktadır. Test 3 kümesinde 255 adet protein dizilimi vardır. Bunlardan 115 tanesi alerjen 140 tanesi alerjen olmayan proteinlerdir.

Eğitim 4 kümesinde 1019 adet protein diziliminin 459 tanesini alerjen proteinler, 560 tanesi ise alerjen olmayan proteinler oluşturmaktadır. Test 4 kümesinde 255 adet protein dizilimi vardır. Bunlardan 115 tanesi alerjen 140 tanesi alerjen olmayan proteinlerdir. Eğitim 5 kümesinde 1019 adet protein diziliminin 459 tanesini alerjen proteinler, 560 tanesi ise alerjen olmayan proteinler oluşturmaktadır. Test 5 kümesinde 255 adet protein dizilimi vardır. Bunlardan 115 tanesi alerjen 140 tanesi alerjen olmayan proteinlerdir.

3.6.2 Protein dizilimlerinin gösterilmesi

Sınıflandırma teorisine göre, sınıflandırıcıya verilecek girdi, sabit uzunlukta nitelik vektörlerinin bir koleksiyonu şeklinde olmalıdır. Proteinler, amino asitlerin zincir halinde birbirlerine bağlanmasından oluşan büyük organik bileşikler olduğu için, dizilim bilgisinin sınıflandırıcıya doğrudan verilmesi mümkün değildir. Bu nedenle, proteinlerin sabit uzunlukta nitelik vektörleri şeklinde gösterilmesine ihtiyaç duyulmaktadır. Bilinen 20 çeşit amino asit vardır. Çizelge 3.6.1.1'de bu amino asitler gösterilmiştir. Bu nedenle giriş vektörleri bilinen 20 çeşit amino asit üzerinden hazırlanmıştır.

Çizelge 3.6.1.1 Amino asitler ve Kısaltmaları

Aminoasit	Üç Harfli Kodu	Tek Harfli Kodu
Glycine	Gly	G
Alanine	Ala	A
Valine	Val	V
Leucine	Leu	L
Isoleucine	Ile	I
Methionine	Met	M
Phenylalanine	Phe	F
Trytophan	Trp	W
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Cysteine	Cys	C
Tyrosine	Tyr	Y
Asparagine	Asn	N
Glutamine	Gln	Q
Aspartic	Asp	D
Glutamic	Glu	E
Lysine	Lys	K
Arginine	Arg	R
Histidine	His	H

Protein dizilerinin gösterimi için birçok yöntem önerilmiştir. Protein dizilim gösterimlerinden bir tanesi de n-peptit bileşim yöntemidir [53]. Bu çalışmada protein dizilimlerinin gösterilmesi için farklı beş yöntem uygulanmıştır. Amino asit bileşim yöntemi (20 boyutlu vektörel gösterim), dipeptit bileşim yöntemi (400 boyutlu vektörel gösterim), amino asit ve dipeptit bileşimin bir arada kullanılması (420 boyutlu vektörel gösterim), amino asit ve tripeptit bileşimin birlikte kullanılması (8020 boyutlu vektörel gösterim) ve benzerlik skorları ile proteinlerin ifade edilmesi yöntemleri uygulanmıştır.

Protein dizilerinin gösteriminde kullanılan yöntemlerden biri amino asit bileşim yöntemidir. Bu yöntemle, tüm proteinler, bilinen 20 çeşit amino asit olması gözönünde bulundurularak 20 boyutlu nitelik vektörleri ile gösterilmişlerdir. Her boyut ilgili amino asidin dizilim içerisinde bulunma sıklığıdır. Bu sıklık veri kümesi içinde bulunan her bir protein için ayrı ayrı hesaplanmıştır. Verilen dizilim bilgisi içerisinde ilgili amino asitin yüzdesi bulunmuştur ve sınıflandırma için bu değer kullanılmıştır. Amino asit bileşim gösterimi Şekil 3.6.1.3'te verilmiştir.

G A V L I M F W P S T C Y N Q D E K R H

Şekil 3.6.1.3 Amino asit Bileşim Gösterimi

GG	GA	GV	GL	GI	GM	GF	GW	GP	GS	GT	GC	GY	GN	GQ	GD	GE	GK	GR	GH
AG	AA	AV	AL	AI	AM	AF	AW	AP	AS	AT	AC	AY	AN	AQ	AD	AE	AK	AR	AH
VG	VA	VV	VL	VI	VM	VF	VW	VP	VS	VT	VC	VY	VN	VQ	VD	VE	VK	VR	VH
LG	LA	LV	LL	LI	LM	LF	LW	LP	LS	LT	LC	LY	LN	LQ	LD	LE	LK	LR	LH
IG	IA	IV	IL	II	IM	IF	IW	IP	IS	IT	IC	IY	IN	IQ	ID	IE	IK	IR	IH
MG	MA	MV	ML	MI	MM	MF	MW	MP	MS	MT	MC	MY	MN	MQ	MD	ME	MK	MR	MH
FG	FA	FV	FL	FI	FM	FF	FW	FP	FS	FT	FC	FY	FN	FQ	FD	FE	FK	FR	FH
WG	WA	WV	WL	WI	WM	WF	WW	WP	WS	WT	WC	WY	WN	WQ	WD	WE	WK	WR	WH
PG	PA	PV	PL	PI	PM	PF	PW	PP	PS	PT	PC	PY	PN	PQ	PD	PE	PK	PR	PH
SG	SA	SV	SL	SI	SM	SF	SW	SP	SS	ST	SC	SY	SN	SQ	SD	SE	SK	SR	SH
TG	TA	TV	TL	TI	TM	TF	TW	TP	TS	TT	TC	TY	TN	TQ	TD	TE	TK	TR	TH
CG	CA	CV	CL	CI	CM	CF	CW	CP	CS	CT	CC	CY	CN	CQ	CD	CE	CK	CR	CH
YG	YA	YV	YL	YI	YM	YF	YW	YP	YS	YT	YC	YY	YN	YQ	YD	YE	YK	YR	YH
NG	NA	NV	NL	NI	NM	NF	NW	NP	NS	NT	NC	NY	NN	NQ	ND	NE	NK	NR	NH
QG	QA	QV	QL	QI	QM	QF	QW	QP	QS	QT	QC	QY	QN	QQ	QD	QE	QK	QR	QH
DG	DA	DV	DL	DI	DM	DF	DW	DP	DS	DT	DC	DY	DN	DQ	DD	DE	DK	DR	DH
EG	EA	EV	EL	EI	EM	EF	EW	EP	ES	ET	EC	EY	EN	EQ	ED	EE	EK	ER	EH
KG	KA	KV	KL	KI	KM	KF	KW	KP	KS	KT	KC	KY	KN	KQ	KD	KE	KK	KR	KH
RG	RA	RV	RL	RI	RM	RF	RW	RP	RS	RT	RC	RY	RN	RQ	RD	RE	RK	RR	RH
HG	HA	HV	HL	HI	HM	HF	HW	HP	HS	HT	HC	HY	HN	HQ	HD	HE	HK	HR	HH

Şekil 3.6.1.4 Dipeptit Bileşim Gösterimi

Dipeptit bileşim yöntemi dizilim gösterilmesi için kullanılan bir yöntemdir. Bu yöntemde amino asit bileşimi yerine dipeptitler kullanılarak 400 boyutlu bir vektör yaratılmıştır. Bu vektörün her bir boyutu 20 amino asitten iki tanesinin birlikte bir boyut olarak davranması ile elde edilen 400 boyutlu vektörlerden oluşur. Şekil 3.6.1.4'te mevcuttur. Değerler 2-uzunluklu amino asit zincirinin, protein dizisi içerisindeki sıklık değeri bulunarak hesaplanmıştır.

Amino asit bileşim yöntemi ve dipeptit bileşim yöntemi ile bulunan vektörler birlikte kullanılarak 420 boyutlu vektör elde edilmiştir. Bu vektör için her bir protein diziliminin sıklık değerleri hesaplanmıştır. Bu değerler her bir amino asit sıklığı ve dipeptit sıklığı bulunarak yapılmıştır.

Tripeptit bileşim yöntemi 20 bilinen amino asitin üçerli gruplanması ile elde edilen 8000 (20*20*20) boyutlu vektörel gösterimdir. Amino asit ve tripeptit bileşim yöntemleri bir arada kullanılarak 8000+20=8020 boyutlu vektörel gösterim için her bir protein diziliminin frekans değerleri hesaplanmıştır.

G	A	V	L	I	M	F	W	P	S	T	C	Y	N	Q	D	E	K	R	H
GGG	GGA	GGV	GGL	GGI	GGM	GGF	GGW	GGP	GGS	GGT	GGC	GGY	GGN	GGQ	GGD	GGE	GGK	GGR	GGH
GAG	GAA	GAV	GAL	GAI	GAM	GAF	GAW	GAP	GAS	GAT	GAC	GAY	GAN	GAQ	GAD	GAE	GAK	GAR	GAH
GVG	GVA	GVV	GVL	GVI	GVM	GVF	GVW	GVP	GVS	GVT	GVC	GVY	GVN	GVQ	GVD	GVE	GVK	GVR	GVH
GLG	GLA	GLV	GLL	GLI	GLM	GLF	GLW	GLP	GLS	GLT	GLC	GLY	GLN	GLQ	GLD	GLE	GLK	GLR	GLH
GIG	GIA	GIV	GIL	GII	GIM	GIF	GIW	GIP	GIS	GIT	GIC	GIY	GIN	GIQ	GID	GIE	GIK	GIR	GIH
GMG	GMA	GMV	GML	GMI	GMM	GMF	GMW	GMP	GMS	GMT	GMC	GMY	GMN	GMQ	GMD	GME	GMK	GMR	GMH
GFG	GFA	GFV	GFL	GFI	GFM	GFF	GFW	GFP	GFS	GFT	GFC	GFY	GFN	GFQ	GFD	GFE	GFK	GFR	GFH
GWG	GWA	GWV	GWL	GWI	GWM	GWF	GWW	GWP	GWS	GWT	GWC	GWY	GWN	GWQ	GWD	GWE	GWK	GWR	GWH
GPG	GPA	GPV	GPL	GPI	GPM	GPF	GPW	GPP	GPS	GPT	GPC	GPY	GPN	GPQ	GPD	GPE	GPK	GPR	GPH
GSG	GSA	GSV	GSL	GSI	GSM	GSF	GSW	GSP	GSS	GST	GSC	GSY	GSN	GSQ	GSD	GSE	GSK	GSR	GSH
GTG	GTA	GTV	GTL	GTI	GTM	GTF	GTW	GTP	GTS	GTT	GTC	GTY	GTN	GTQ	GTD	GTE	GTK	GTR	GTH
GCG	GCA	GCV	GCL	GCI	GCM	GCF	GCW	GCP	GCS	GCT	GCC	GCY	GCN	GCQ	GCD	GCE	GCK	GCR	GCH
GYG	GYA	GYV	GYL	GYI	GYM	GYF	GYW	GYP	GYS	GYT	GYC	GYY	GYN	GYQ	GYD	GYE	GYK	GYR	GYH
GNG	GNA	GNV	GNL	GNI	GNM	GNF	GNW	GNP	GNS	GNT	GNC	GNY	GNN	GNQ	GND	GNE	GNK	GNR	GNH
GQG	GQA	GQV	GQL	GQI	GQM	GQF	GQW	GQP	GQS	GQT	GQC	GQY	GQN	GQQ	GQD	GQE	GQK	GQR	GQH

Şekil 3.6.1.5- Tripeptit Bileşim Gösterimi

DVM için kullanılan eğitim ve test sınıflarındaki protein dizilimleri ilk olarak amino asit bileşim yöntemi kullanılarak gerçekleştirilmiştir. Amino asit bileşim yöntemi, bir proteindeki her bir amino asitin oranıdır. 20 amino asitin hepsinin oranı eşitlik 3.25'te verilen denklemle hesaplanmıştır.

$$R(i) = \frac{S(i)}{T} \tag{3.25}$$

R: i. amino asitinin oranı

i: herhangi bir amino asit

S: Toplam amino asit sayısı

T: Proteindeki toplam amino asit sayısı

DVM için ikinci olarak, dipeptit bileşim yöntemi kullanılarak, her bir protein dizilimiyle ilgili uzunluğu 400 (20x20) olan desen elde edilmiştir. Amino asit dizilim bilgisi, amino asit bileşim yöntemi boyunca korunmaktadır. Her bir peptidin oranı eşitlik 3.26'da verilen denklem ile hesaplanmıştır.

$$R_D(i) = \frac{S_D(i)}{T_D} \tag{3.26}$$

i: 400 dipeptitten bir tanesi

R_D: i. dipeptitin oranı

S_D: Toplam dipeptit sayısı

T_D : Bütün dipeptitlerin toplam sayısı

DVM için son olarak, tripeptit bileşim yöntemi kullanılarak, her bir protein dizilimiyle ilgili uzunluğu 8000 (20x20x20) olan desen elde edilmiştir. Her bir peptidin oranı eşitlik 3.27'de verilen denklem ile hesaplanmıştır.

$$R_T(i) = \frac{S_T(i)}{T_T} \tag{3.27}$$

i: 8000 tripeptitten bir tanesi

R_T: i. tirpeptitin oranı

S_T: Toplam tripeptit sayısı

T_T: Bütün tripeptitlerin toplam sayısı

Amino asit bileşim ve dipeptit bileşim yöntemlerini biraraya getirerek 420 boyutlu yeni vektör için oranlar hesaplanmıştır ve sınıflandırma için kullanılmıştır. Aynı şekilde tripeptit bileşim ve amino asit bileşim yöntemlerini birarada kullanılarak 8020 boyutlu vektörler oluşturulmuştur.

Diğer bir gösterim yöntemi ise test edilen protein ile diğer proteinler arasındaki benzerlik skorlarıdır. Biyolojik olarak dizi hizalamanın iki farklı biçimi mevcuttur. Bunlardan biri yerel hizalama diğeri ise global hizalamadır. Global hizalama, benzerlik (ya da uzaklık) için her iki dizinin tüm uzunluğu üzerinden puanı en uygun şekle getirmektedir.

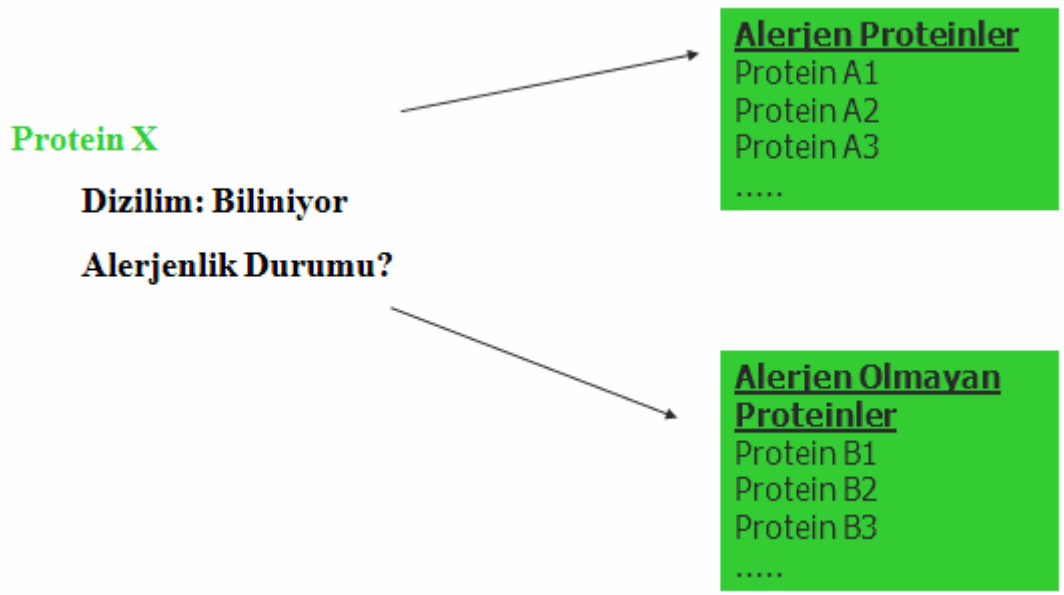
Bu çalışmada PAM (point accepted mutations) matrisleri kullanılmıştır. PAM, amino asitin %1'lik kısmının değişime uğraması için gereken zaman uzunluğudur [5]. PAM bir milyar yıl olarak tahmin edilmektedir. Örneğin, bir PAM70 matrisi, 70 PAM sürede meydana gelen değişimler hakkındaki puanlama bilgisini içerir. Şekil 3.6.1.5'te bu çalışmada kullanılan PAM70 matrisi gösterilmiştir. Matriste de görüldüğü gibi T ile S'nin değişimi T ile P'nin değişiminden daha uygundur, T-S puanı 2, T-P puanı -2'den daha yüksektir.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-4	-2	-1	-4	-2	-1	0	-4	-2	-4	-4	-3	-6	0	1	1	-9	-5	-1
R	-4	8	-3	-6	-5	0	-5	-6	0	-3	-6	2	-2	-7	-2	-1	-4	0	-7	-5
N	-2	-3	6	3	-7	-1	0	-1	1	-3	-5	0	-5	-6	-3	1	0	-6	-3	-5
D	-1	-6	3	6	-9	0	3	-1	-1	-5	-8	-2	-7	-10	-4	-1	-2	-10	-7	-5
C	-4	-5	-7	-9	9	-9	-9	-6	-5	-4	-10	-9	-9	-8	-5	-1	-5	-11	-2	-4
Q	-2	0	-1	0	-9	7	2	-4	2	-5	-3	-1	-2	-9	-1	-3	-3	-8	-8	-4
E	-1	-5	0	3	-9	2	6	-2	-2	-4	-6	-2	-4	-9	-3	-2	-3	-11	-6	-4
G	0	-6	-1	-1	-6	-4	-2	6	-6	-6	-7	-5	-6	-7	-3	0	-3	-10	-9	-3
H	-4	0	1	-1	-5	2	-2	-6	8	-6	-4	-3	-6	-4	-2	-3	-4	-5	-1	-4
I	-2	-3	-3	-5	-4	-5	-4	-6	-6	7	1	-4	1	0	-5	-4	-1	-9	-4	3
L	-4	-6	-5	-8	-10	-3	-6	-7	-4	1	6	-5	2	-1	-5	-6	-4	-4	-4	0
K	-4	2	0	-2	-9	-1	-2	-5	-3	-4	-5	6	0	-9	-4	-2	-1	-7	-7	-6
M	-3	-2	-5	-7	-9	-2	-4	-6	-6	1	2	0	10	-2	-5	-3	-2	-8	-7	0
F	-6	-7	-6	-10	-8	-9	-9	-7	-4	0	-1	-9	-2	8	-7	-4	-6	-2	4	-5
P	0	-2	-3	-4	-5	-1	-3	-3	-2	-5	-5	-4	-5	-7	7	0	-2	-9	-9	-3
S	1	-1	1	-1	-1	-3	-2	0	-3	-4	-6	-2	-3	-4	0	5	2	-3	-5	-3
T	1	-4	0	-2	-5	-3	-3	-3	-4	-1	-4	-1	-2	-6	-2	2	6	-8	-4	-1
W	-9	0	-6	-10	-11	-8	-11	-10	-5	-9	-4	-7	-8	-2	-9	-3	-8	13	-3	-10
Y	-5	-7	-3	-7	-2	-8	-6	-9	-1	-4	-4	-7	-7	4	-9	-5	-4	-3	9	-5
V	-1	-5	-5	-5	-4	-4	-4	-3	-4	3	0	-6	0	-5	-3	-3	-1	-10	-5	6

Şekil 3.6.1.6 PAM70 matrisi.

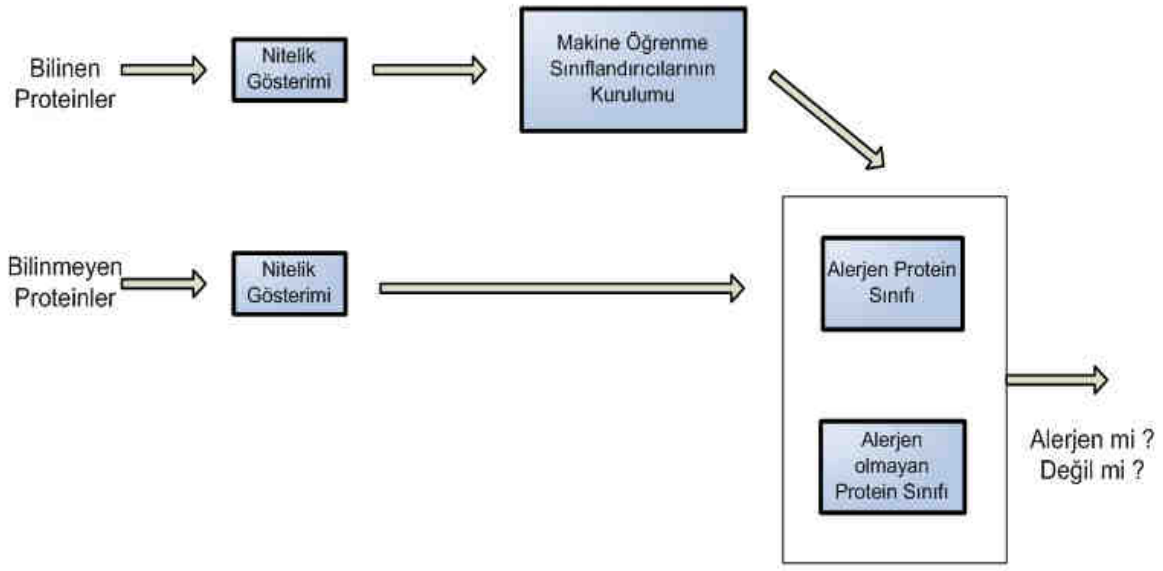
3.6.3 Uygulanan yöntemler

Yapılan çalışmada protein sınıflandırılması için çeşitli yöntemler uygulanmıştır. Proteinlerin sınıflandırılmasında Şekil 3.6.3.1'de gösterildiği gibi alerjen olan proteinler ve alerjen olmayan proteinler olmasından dolayı iki adet sınıfımız vardır.



Şekil 3.6.3.1 Sınıflandırma

Sınıflandırma işlemi gerçekleştirilirken sınıflandırma işlemi için elimizde iki farklı küme vardır. Birinci küme sınıfı bilinen proteinlerden oluşan eğitim kümemiz diğer küme ise sınıfı bilinmeyen ve sınıfı bulunmak istenen proteinlerden oluşan test kümemizdir. Sınıflandırma işlemi alt yapısında her iki küme için de mevcut proteinlerin nitelik gösterim şeklinde çeşitli yöntemlerle belirlenmesi gerçekleştirilir. Bu aşamadan sonra farklı makine öğrenme yöntemleri kullanılarak sınıfı bilinen proteinlerden sınıfı bilinmeyen proteinlerin sınıflandırma işlemi gerçekleştirilmiştir. Şekil 3.6.3.2'de sınıflandırma alt yapısını açıklayan gösterim mevcuttur.



Şekil 3.6.3.2 Sınıflandırma Altyapısı

Bu çalışmada Perl dilinin standart kütüphaneleri kullanılarak farklı protein dizilimleri için üç ayrı sınıflandırma yöntemi uygulanmıştır. Bu yöntemlerin gerçekleştirilmesi ile elde edilen sonuçların performans ölçümleri karşılaştırılmıştır. Kullanılan ilk yöntem K-En Yakın Komşu algoritmasıdır. Bulanık K-En Yakın Komşu algoritması ikinci olarak gerçekleştirilen algoritmadır. Son olarak da Destek Vektör Makinesi (DVM) kullanılmıştır.

1. K-En Yakın Komşu yöntemi : 20 boyutlu amino asit bileşimi ve 400 boyutlu dipeptit bileşimi için uygulanmıştır. Amino asit ve dipeptit bileşimleri bir arada kullanılarak 420 boyutlu vektör için K-En Yakın Komşu yöntemi gerçekleştirilmiştir. K-En Yakın Komşu yöntemi için son olarak benzerlik skorları kullanılmıştır.

2. Bulanık K-En Yakın Komşu Yöntemi : Bulanık K-En Yakın Komşu yöntemi için sırası ile 20 boyutlu amino asit bileşimi, 400 boyutlu dipeptit bileşimi ve son olarak amino asit ve dipeptit bileşimin bir araya getirilmesi ile oluşan 420 boyutlu vektör kullanılarak uygulamalar gerçekleştirilmiştir.

3. Destek Vektör Makineleri : Destek vektör makineleri pozitif ve negatif örnekleri birbirinden ayırmak için amino asit bileşimi, dipeptit bileşimi, amino

asit+dipeptit bileşimi kullanılmıştır. Bu çalışmaya ek olarak tripeptit bileşim ve amino asit bileşim bir arada kullanılarak 8020 boyutlu vektör ile uygulama gerçekleştirilmiştir. DVM ile benzerlik skorları kullanılarak veri kümesindeki protein dizilimleri için tüm dizilim verisi kullanılarak ve dizilimin ilk 20 elemanı kullanılarak iki ayrı uygulama gerçekleştirilmiştir.

K-en yakın komşu uygulaması

Protein dizilimleri kullanılarak, her bir proteini 20 standart amino asitin bileşimi ile ifade edilen vektörler oluşturulmuştur. Veri kümesinde yeralan beş adet eğitim seti ve beş adet test seti için de mevcut tüm dizilimler 20 boyutlu vektörler ile ifade edilecektir. K-En Yakın Komşu yönteminde eğitim ve test setinde yeralan vektörler arasındaki uzaklıklar öklit teoremi kullanılarak hesaplanmıştır. K değerleri sırası ile 5 ve 10 olarak seçilmiştir. K=5 değeri için, 5 sayısı tek sayı olduğundan ve sınıflandırma işlemi için iki sınıfımız bulunduğundan dolayı herhangi bir problem olmadan sınıflandırma işlemi gerçekleştirilmiştir. Ancak K=10 değeri için sınıflandırmada alerjen protein sınıfı ile alerjen olmayan protein sınıfına ait 10 değer 5'erli şekilde 2 sınıf için eşit sayıda çıkabilmektedir. Bu şekilde oluşan bir problem için uzaklık değerleri toplamı esas alınarak sorun çözümlenmiştir.

K-En Yakın Komşu yönteminde protein dizilimleri öncelikle sabit uzunlukta amino asit bileşim yöntemi ile 20 boyutlu vektörler şeklinde ifade edildikten sonra öklit uzaklığı teoremi kullanılarak proteinler arasındaki uzaklık değerleri hesaplanmıştır. K=5 değeri için her bir test proteinine en yakın 5 protein bulunmuştur. Bulunan proteinlerin bilinen sınıflarına göre (alerjen olanlar ve olmayanlar) hangi sınıftan olan protein daha fazla ise test proteinin sınıfı bu sınıf olarak belirlenmiştir. Bu işlem tüm test proteinleri sınıflanana kadar devam eder. K=10 değeri için gerçekleşirilen uygulamada, sınıflandırma işleminin gerçekleştirilmesi sırasında varolan iki sınıf değeri için alerjen sınıfa ait olan en yakın komşular ve alerjen olmayan sınıfa ait olan komşular eşit sayıda çıkabileceği varsayılarak, gerçekleştirilen uygulamada bulunan komşular için hesaplanan uzaklık değerleri toplamı esas alınarak, minimum uzaklık değeri toplamı hangi sınıfa ait ise test edilecek proteinin de sınıfı aynı sınıf olarak sınıflandırılmıştır.

Dipeptit bileşimi için amino asitlerin ikililer şeklinde gösterilmesi ile elde edilen 400 boyutlu vektörler oluşturulmuştur. Eğitim setinde ve test setinde yer alan tüm dizilimler için oluşturulan bu vektörler arasındaki uzaklıklar öklit uzaklığı teoremi ile bulunmuştur. İlk olarak, K=5 değeri için uzaklıklar sıralanmıştır. Uzaklık değerleri en az olan beş örnek alınmıştır ve seçilen örnekler içerisinde en çok örneği bulunan sınıfa göre test örneğinin sınıfı belirlenmiştir. Aynı işlem daha sonra K=10 değeri için tekrarlanmıştır. Tüm test proteinleri sınıflanana kadar uygulamaya devam edilmiştir.

Amino asit ve dipeptit bileşim yöntemlerinin birlikte kullanılması ile elde edilen 420 boyutlu vektörler oluşturulmuştur. Vektörler arasındaki uzaklıklar öklit uzaklığı teoremi ile bulunmuştur. İlk olarak, K=5 değeri için uzaklıklar sıralanmıştır. K-En Yakın Komşu yöntemine göre en yakın beş komşu bulunmuş ve beş komşunun bilinen sınıflarına göre sınıflandırma işlemi gerçekleştirilmiştir. Aynı işlem daha sonra K=10 değeri için tekrarlanmıştır. Tüm test proteinleri sınıflanana kadar uygulamaya devam edilmiştir.

K-En Yakın komşu yöntemi için son olarak protein dizilimleri için benzerlik skorları kullanılmıştır. Yapılan çalışmada ilk olarak dizilim verisinin tamamı için benzerlik skorları hesaplanmıştır. Hesaplanan bu değerlerden K=5 için en yüksek beş değer alınarak sınıflandırma işlemi bu beş değerin ait olduğu proteinlerin bilinen sınıflarına göre yapılmıştır. Benzerlik skorları uygulandığında peptit bileşim yöntemlerinden farklı olarak hesaplanan en büyük değerler alınmıştır. Çünkü en benzer olan sınıflar bulunmak istenilmektedir.

Protein dizilim verisi için tüm dizilim verisi yerine, dizilimin ilk 10 amino asiti ve ilk 20 amino asiti kullanılarak benzerlik skorları hesaplanmıştır. Benzerlik skorlarına uygulanan K-En Yakın komşu yöntemi için K değerleri 5, 10 ve 20 olarak alınmış performanslar karşılaştırmalı olarak bulunmuştur.

K-En Yakın Komşu yöntemi uygulamasına ek olarak benzerlik skorları kullanılarak yapılan çalışmada en yakın K komşu kullanılarak sınıflandırma yapmak yerine en benzer dizilimin sınıfı belirlenerek test edilen örneğin sınıfı en benzer proteinin sınıfı ile aynıdır yaklaşımı izlenerek de sınıflandırma işlemi yapılmıştır. Bu

uygulama için tüm dizilim, ilk 10 amino asit ve ilk 20 amino asit için benzerlik değerleri hesaplanmıştır. Test edilen proteine en benzer protein bulunmuştur. Test proteini en benzer proteinin sınıfındandır varsayımı ile sınıflanırma işlemi gerçekleştirilmiştir.

Bulanık K- en yakın komşu uygulaması

Bulanık K-en yakın komşu algoritması, proteinlerin sınıflandırılması için ilk kez yapılan bu çalışma ile kullanılmıştır.

Bulanık K-en yakın komşu algoritmasında, protein dizilimlerinde amino asit bileşim yöntemi, dipeptit bileşim yöntemi ve amino asit+dipeptit bileşim yöntemleri birlikte kullanılarak 420 boyutlu vektörler oluşturulmuştur. Vektörler arası uzaklık değerlerinin hesaplanması Bulanık K-en Yakın Komşu yönteminde de K-En Yakın Komşu yönteminde olduğu gibi öklit uzaklığı teoremi kullanılarak hesaplanmıştır.

Bulanık K-En Yakın Komşu yöntemi için ilk olarak amino asit bileşimleri kullanılarak 20 boyutlu vektörler oluşturulmuştur. Tüm eğitim ve test örneklerinden oluşan beş farklı küme için uygulama gerçekleştirilmiştir. Test edilecek protein vektörünün sınıflandırılması için öncelikli olarak öklit uzaklığı teoreminden yaralanılarak uzaklıklar bulunmuştur. K=5 değeri için uzaklık değerleri sıralanmış ve beş en yakın komşu en az uzaklık mesafesine göre belirlenmiştir. Bu aşamaya kadar K-En Yakın Komşu yöntemine benzer işlemler yapılmıştır. Bulanık K-En Yakın Komşu yöntemi ile bu aşamadan sonra gerçekleştirilecek olan sınıflandırma bölümü için eşitlik 3.28'de verilen formül uygulanmıştır.

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)}{\sum_{j=1}^{k}\left(1/\left\|x - x_j\right\|^{2/(m-1)}\right)} \qquad (3.28)$$

Burada m ile ifade edilen, bulanık algoritma parametremizdir. m değeri gerçekleştirilen uygulama için 2 olarak alınmıştır. K ile gösterilen parametre seçilen en yakın komşu değeridir. $u_i(x)$, x vektörünün i. sınıfa olan üyelik

değeridir. Bu değer 0 ile 1 arasında yapılan hesaplama sonucu elde eldilmiş olan değerdir. $||x-x_j||$ x vektörünün x_j vektörüne yani j. komşuya olan ve öklit uzaklık teoremi ile hesaplanan uzaklık değerini vermektedir. U_{ij} değeri, j. komşunun i. sınıfa olan üyelik değeridir. Gerçekleştirilen uygulamada U_{ij} ile belirlenen uzaklık değerleri için crispet yöntemi adı verilen yöntem kullanılmıştır. Buna göre x_i vektörü sınıfa ait ise 1, ait değilse 0 değerini alacaktır. U_{ij} değerinin hesaplanması için bulanık temelli alternatif yollar da kullanılabilmektedir. Verilen eşitlik ile test proteini için alerjen olan sınıf ve alerjen olmayan sınıf için üyelik değerleri hesaplandıktan sonra hesaplanan en yüksek üyelik değeri doğrultusunda test proteinin sınıfı belirlenmiştir.

Bulanık K-En Yakın Komşu yöntemi için dipeptit bileşimleri kullanılarak 400 (20*20) boyutlu vektörler oluşturulmuştur. Vektörler arası uzaklık değerleri öklit teoremi ile bulunduktan sonra uzaklık değerleri sıralanarak en küçük uzaklık değerlerine sahip K=5 değeri için en yakın beş komşu bulunmuştur. Bu aşamadan sonra sınıfı aranan vektörler için vektörlerin alerjen sınıf ve alerjen olmayan sınıf için üyelik değerleri hesaplamıştır. Bu hesaplamadan sonra üyelik değeri büyük olan sınıf hangi sınıf ise vektör üyelik değeri büyük olan sınıfa aittir şeklinde sınıflandırma işlemi tamamlanmıştır. Veri kümemizde mevcut beş küme için aynı işlem K=5 değeri için tekrarlanmıştır ve performans ölçümleri hesaplanmıştır. K=10 değeri için uzaklık değerleri hesaplandıktan sonra sınıf sayılarının eşit çıkması sorununu engellemek için uzaklık, alerjen sınıfa ait vektörler ve alerjen olmayan sınıfa ait vektörler için toplamlar bulunarak gerçekleştirilmiştir. K=10 değeri için uygulmanın diğer bölümleri K=5 değeri için yapılan uygulama ile benzer şekilde gerçekleştirilmiştir.

Bulanık K-En Yakın Komşu yöntemi için son olarak amino asit bileşim ve dipeptit bileşimleri kullanılarak 420 boyutlu vektörler oluşturulmuştur. Vektörler arasındaki uzaklıklar öklit uzaklığı teoremi ile hesaplanmıştır. İlk olarak, K=5 değeri için uzaklıklar sıralanmıştır. K-En Yakın Komşu yöntemine göre en yakın beş komşu bulunmuş ve alerjen sınıf ve alerjen olmayan sınıf için üyelik değerleri hesaplanarak sınıflandırma işlemi büyük olan üyelik değerinin ait olduğu sınıf test vektörünün ait olduğu sınıftır şeklinde gerçekleştirilmiştir. Aynı işlem daha sonra K=10 değeri için en yakın komşular uzaklık değerleri toplamları gözönünde

bulundurularak hesaplanmış ve tekrarlanmıştır. K değerine bağlı bulunan üyelik değerlerine göre, değer hangi sınıfa aitse, test örneği bu değere bağlı olarak belirlenmiştir. Tüm test proteinleri sınıflanana kadar uygulamaya devam edilmiştir.

Destek vektör makineleri (DVM) uygulaması

DVM (www.bioinformatics.ubc.ca) *UBC Bioinformatics* sayfasında mevcut olan SVM-Gist isimli açık kaynak yazılımı kullanılarak uygulanmıştır. Bu yazılım kullanıcıya birçok parametre seçme imkanı sağlar. Çekirdek (kernel) fonksiyonu girdi vektör çiftleri arasındaki benzerlik skoru olarak davranır. Temel çekirdek, her vektörün öznitelik uzayındaki uzunluğunun 1 olması için eşitlik 3.29'daki formül ile normalize edilir.

$$K(X,Y)=\frac{X\cdot Y}{\sqrt{(X\cdot X)(Y\cdot Y)}} \tag{3.29}$$

Verilen eşitlikte, X ve Y girdi vektörleri, K(.,.) çekirdek fonksiyonu, ve "." nokta çarpımı (vektörel çarpım) simgelemektedir. Bu çekirdek daha sonra radyal tabanlı K'(X,Y) çekirdeğine eşitlik 3.30'da verilen formül ile dönüştürülür.

$$K'(X,Y)=1+e^{-\frac{K(X,X)-2K(X,Y)+K(Y;Y)}{2\sigma^2}} \tag{3.30}$$

Burada σ genişliği, herhangi bir pozitif eğitim örneğinin en yakın negatif örneğe olan medyan öklit uzaklığıdır. DVM'nin ayırıcı hiperdüzleminin orijinden geçmesi gerektiği için 1 sabiti çekirdeğe eklenir. Böylece veri orijinden uzaklaştırılır. Bir asimetrik değişebilir marjin, çekirdek matrisin köşegenine 0.02*ρ eklenerek uygulanır, burada ρ daha önceki DVM sınıflandırma metodlarında olduğu gibi o andaki protein ile aynı etikete sahip eğitim setindeki proteinlerin oranıdır. DVM çıktısı test setindeki her protein için olan diskriminant skorlarının listesidir.

SVM-Gist isimli açık kaynak yazılım için verilerin belirli formatlarda hazırlanması gerekmektedir. SVM-Gist için veri formatı ve çalıştırılması gereken komutlar aşağıda açıklanmıştır :

- İlk olarak DVM için makine öğrenmesinin yani eğitimin uygulanması gerekmektedir. Bunun için SVM-Gist yazılımında aşağıda verilen parametrelerle ilgili komut satırının çalıştırılması gerekmektedir.

  gist-train-svm [özellikler] -train <eğitim dosyası > -class <etiketler>

  Girdi dosyaları:

  <eğitim dosyası > : Eğitim dosyası için gereken format “tab” karakteri ile ayrılmış eğitim proteinlerinin yeraldığı dosyadır. Dosya içinde ilk sütun eğitim proteinlerinin kimliğinden, kalan sütunlar ise özellik gösterimi ile elde edilen frekanslardan oluşmaktadır. Şekil 3.6.3.3’te amino asit bileşim yöntemi ile oluşturulmuş eğitim dosyası örneği verilmiştir.

corner	G	A	V	L	I	M	F	...
>Allergen1	0.0923913	0.0733695	0.0788043	0.0625	0.0570652	0.0271739	0.032608	...
>Allergen2	0.042372	0.08474576	0.0847457	0.101694	0.0423728	0.02542372	0.0254237	...
>Allergen7	0.059405	0.04950495	0	0.1188118	0.05940594	0.0198019	0.0594059	...
>Allergen11	0.082524	0.05825242	0.0679611	0.05825242	0.058252	0.01941747	0.0145631	...
>nonAllergen13	0.074235	0.05458515	0.0698689	0.0829694	0.0480349	0.01310043	0.0371179	...

Şekil 3.6.3.3 Eğitim Dosyası (amino asit bileşim)

<etiketler> : etiketler dosyası çift sütunlu bir dosyadır. Ayırıcı karakter “tab” karakteridir. Etiket dosyası, Eğitim dosyasında yeralan her bir protein için aynı sıra ile kimlik bilgisini ve bilginin karşılığında yer alan yeralan “-1” veya “1” değeri ile nitelenmesinin oluşturduğu dosyadır. Alerjen sınıf için “1” , alerjen olmayan sınıf için “-1” değerleri verilmiştir. Şekil 3.6.3.4’te amino asit bileşim yöntemi ile oluşturulmuş etiket dosyası örneği verilmiştir.

corner	label
>Allergen1	1
>Allergen2	1
>Allergen7	1
>Allergen11	1
>nonallergen13	-1

Şekil 3.6.3.4 Etiket Dosyası (amino asit bileşim)

Çıktı :

Çıktı olarak oluşturulan dosya beş sütundan oluşmaktadır. İlk iki sütun giriş olarak sağlanan sınıflandırma dosyası ile aynıdır. Üçüncü sütun her biri karşılığı olan etiket değeri ile çarpılmış DVM için öğrenilmiş ağırlık değerlerini belirtir. Sütun dört ve beş tahmin edilen sınıflandırma ve buna karşılık gelen diskriminant değerini belirtir. Bu çıktı dosyası sınıflandırma işlemi için kullanılacaktır. *Çıktı dosyasının kullanıldığı yerlerdeki dosya ismi "cikti" olarak belirtilecektir.* Şekil 3.6.3.5'te dosya formatı örneği verilmiştir.

- Makine öğrenmesi aşaması tamamlandıktan sonra, eğitilmiş destek vektör makinesi ile test proteinlerinin sınıflandırılması işleminin gerçekleştirilmesi için aşağıdaki komut satırı kullanılmıştır.

  gist-classify [özellikler] -train < eğitim dosyası > -learned <cikti> -test <test dosyası>

  Girdiler :

  <eğitim dosyası > : Eğitim dosyası için gereken format "tab" karakteri ile ayrılmış eğitim proteinlerinin yeraldığı dosyadır. Dosya içinde ilk sütun eğitim proteinlerinin kimliğinden, kalan sütunlar ise özellik gösterimi ile elde edilen frekanslardan oluşmaktadır.

  <cikti> : Destek vektör makinesinin eğitim sonucunda oluşturduğu öğrenilmiş ağırlık değerlerinin bulunduğu dosyadır. Bu dosyanın başlığında kernel parametreleri yeralmaktadır.

<test dosyası> : sınıflandırılacak test proteinlerinin yeraldığı dosyadır.

Çıktı :

Çıktı dosyası üç sütunlu bir dosyadır. Sütunlar "tab" karakteri ile ayrılmıştır. İlk sütun test proteininin kimliğini, ikinci sütun sınıflandırılmış değeri (1, -1), üçüncü sütunda hesaplanan diskriminant değerlerini ifade etmektedir.

```
# Generated by compute-weights
# Gist, version 2.1
# For more information, go to http://svm.sdsc.edu
#
# train_file=/home/pavlidis/gist/website/userfiles/193.140.161.84__20433/193.140.161.84__20433_train
# class_file=/home/pavlidis/gist/website/userfiles/193.140.161.84__20433/193.140.161.84__20433_class
# matrix_from_file=false zero_mean=false variance_one=false  normalize=true
# constant=10 coefficient=1 power=1 bias=0
# radial=false width_factor=1 two_squared_width=0 add_diag=0
# feature_select=none thresh_type=percent fthreshold=0
# sum_of_weights=0.166434
# positive_constraint=0 negative_constraint=0 constrain_weights=false
# positive_diagonal=0.495486 negative_diagonal=0.604514
# convergence_threshold=1e-06 seed=1193131072
# objective=0.869883 iterations=139
# time=14.94 s host=Rocks-136.sdsc.edu date=Tue Oct 23 02:18:07 PDT 2007
corner  class   weight  train_classification    train_discriminant
>Allergen1          1   1.213       1  0.3975
>Allergen2          1   2.142      -1  -0.06326
>Allergen7          1   1.178       1  0.4145
>Allergen11         1       0       1   1.377
>Allergen13         1   1.893       1  0.05993
>Allergen14         1   1.319       1  0.3476
>Allergen15         1   1.008       1  0.5014
>Allergen18         1   2.399      -1  -0.1916
>Allergen22         1       0       1   1.135
>Allergen27         1   1.259       1  0.3762
>Allergen32         1   1.046       1  0.4814
>Allergen34         1   1.298       1  0.3559
>Allergen40         1   1.398       1   0.306
>Allergen49         1       0       1    1.21
```

Şekil 3.6.3.5 Çıktı Dosyası (amino asit bileşim)

Destek vektör makineleri kullanılarak gerçekleştirilen uygulama için üç dosya oluşturulmuştur. İlk dosyada eğitim kümeleri için farklı dizilim yöntemleri kullanılarak sınıflandırma işlemi için girdiler belirlenmiştir. İkinci dosyaya eğitim kümelerindeki proteinlerin etiketleri yani alerjen olanlar için (+) alerjen olmayanlar için (-) değerleri yazdırılmıştır. Üçüncü dosyada da, ilk dosya için belirlenen veriler kullanılarak test kümesindeki proteinler için frekans analizleri yapılıp yazdırılmıştır.

İlk iki dosya makineye öğretmek için oluşturulmuştur. Son dosya ise öğrenilmiş değerler sonucunda test edilecek protein örneğinin alerjen olup olmadığını belirlemek için oluşturulmuştur.

Destek Vektör Makineleri için ilk olarak amino asit bileşim yöntemi ile sınıflandırma işlemi yapılmıştır. Öncelikle 20 boyutlu vektörler oluşturularak beş eğitim kümesi için tüm kümelerdeki protein dizilimleri vektörel olarak ifade edilmiş ve DVM için oluşturulacak dosya formatında (eğitim dosyası) 20 amino asit için dizilimdeki frekans değerleri hesaplanarak dosyaya yazdırılmıştır. Eğitim kümesindeki sınıfı bilinen proteinler için etiket değerleri tüm kümeler için her protein karşılığı sınıfı ifade edecek şekilde belirlenerek dosyaya gerekli formatta (etiket dosyası) yazdırılmıştır. Alerjen olan proteinler için *1* alerjen olmayan proteiner için *-1* değerleri verilmiştir. Son olarak test kümesindeki proteinler vektörel şekilde amino asit bileşim yöntemi ile ifade edilmiş ve dosyaya (test dosyası) istenilen formatta veriler yazdırılmıştır. Destek Vektör Makineleri ile veri kümemizde bulunan beş küme için uygulama gerçekleştirilmiş ve farklı eşik değerleri kullanılarak sınıflandırma işlemi yapılmıştır.

Destek Vektör Makineleri ile gerçekleştirilen bir diğer uygulama dipeptit bileşim yöntemi için yapılmıştır. Protein dizilimleri dipeptit bileşim yöntemi ile 400 boyutlu vektörler şeklinde ifade edilmiştir. Dipeptit bileşim yöntemi ile oluşturulan eğitim dosyası için belirli bir kesit alınarak Şekil 3.6.3.6'da gösterilmektedir. Oluşturulan üç dosya sonucunda DVM çıktısı örneğinin bir bölümü şekil 3.6.3.7'de gösterilmiştir. Şekilden görüldüğü gibi oluşturulan ve DVM'ye verilen dosya isimleri sırası ile sim_veri20set1.txt, deney21.weights ve sim_testveri20set1.txt dir. DVM çıktısı test kümesindeki her protein için olan diskriminant skorlarının listesi şeklinde gözlemlenmektedir.

corner	GG	GA...	LM...	PC...	HR	HH
>Allergen3	0	0.00925925...	0.0185185...	0.00925925...	0	0
>Allergen17	0.0092592	0 ...	0 ...	0 ...	0	0
>Allergen19	0	0 ...	0.027777 ...	0 ...	0	0
>Allergen19	0.0080808	0.0040404 ...	0.002020 ...	0.0040404 ...	0.00404040	0

Şekil 3.6.3.6 Eğitim Dosyası (dipeptit bileşim)

```
# Generated by classify
# Gist, version 2.3
# For more information, go to http://svm.sdsc.edu
#
# If you use this software in your research, please cite:
#   Paul Pavlidis, Ilan Wapinski and William Stafford Noble.
#   "Support vector machine classification on the web."
#   Bioinformatics. 20(4):586-587, 2004.
#
# train_file=sim_veri20set1.txt
# learned_file=deney21.weights
# test_file=sim_testveri20set1.txt
# matrix_from_file=false zero_mean=false variance_one=false
normalize=true
# constant=1 coefficient=1 power=1
# radial=true width_factor=1 two_squared_width=0.308437
# host=localhost date=Wed Mar 19 21:24:48 EET 2008
corner      classification    discriminant
>Allergen8          1    0.883768
>Allergen9          1     0.26322
>Allergen10         1    0.998249
>Allergen25         1    0.826534
>Allergen26        -1    -0.682599
>Allergen29        -1    -0.240334
>Allergen35         1     1.24135
>Allergen37         1    0.464548
>Allergen48         1    0.552743
>Allergen53        -1    -0.262958
>Allergen57         1    0.243221
>Allergen62        -1    -0.240334
>Allergen65        -1    -0.240334
>Allergen69         1    0.748255
>Allergen72        -1    -0.218156
>Allergen74        -1    -1.37233
>Allergen75         1     1.20155
```

Şekil 3.6.3.7 DVM Çıktısı Örneği

Amino asit bileşim ve dipeptit bileşim yönteminin birlikte kullanılması ile uygulanan yöntemde protein dizilimleri 420 boyutlu vektörler şeklinde ifade edilmiştir. Şekil 3.6.3.8'de 420 boyutlu nitelik vektörleri için oluşturulmuş eğitim dosyası gösterilmiştir. Eğitim dosyası, etiketlenmiş proteinlerin dosyası ve son olarak da test edilecek dosya hazırlandıktan sonra DVM ile farklı eşik değerleri için sınıflandırma işlemi gerçekleştirilmiştir.

```
corner          G           A          ....  GG            GA          ...
>Allergen5      0.462585    0.3673469...     0.001350 ...  0.0149659...
>Allergen11     0.0092592   0          ...   0          ...  0          ...
>Allergen22     0           0          ...   0.027777 ...  0          ...
...
>nonallergen419 0.178923    0.4523891 ...    0.0028968 ...  0.00309850...
```

Şekil 3.6.3.8 Eğitim Dosyası (amino asit + dipeptit)

Protein dizilimleri amino asit ve tripeptit bileşim yöntemi birlikte kullanılarak 8020 boyutlu nitelik vektörleri şeklinde ifade edilmişlerdir. Protein dizilimi içersinde 8020 boyutlu vektörün her boyutu için dizilimde geçen frekans değerleri hesaplanmıştır. Şekil 3.6.3.9'da gösterilmektedir. Her protein için etiketleme işlemi gerçekleştirilmiştir. Son olarak DVM ile uygulama gerçekleştirilerek test kümesindeki proteinler için diskriminant skorlar hesaplanmıştır. Bu işlem veri kümemizde yeralan beş farklı eğitim ve test kümesi için tekrarlanmıştır.

corner	G....	P	....	GGP	PAA	...
>Allergen1	0.678945	0.017869...		0.002471 ...	0.0091184...	
>Allergen7	0.005214	0	...	0	... 0	...
>Allergen11	0	0	...	0.0303030 ...	0	...
...						
>nonallergen406	0.136894	0.45	...	0.0098733...	0.00609950...	

Şekil 3.6.3.9 Eğitim Dosyası (amino asit + tripeptit)

DVM ile son olarak benzerlik skorları kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Öncelikle veri kümemizde yer alan protein diziliminin tamamı kullanılarak elde edilen skorlar kullanılmıştır. Bir sonraki uygulama için, dizilim verisi küçültülerek dizilimin ilk 20 verisi kullanılarak elde edilen sonuçlar ile uygulama gerçekleştirilmiştir. Tüm eğitim proteinlerinin birbirleri ile olan benzerlik skorları hesaplanarak 1020*1020 boyutlu matrisler oluşturulmuştur. Eğitim dosyasına hesaplanan değerler yazdırılmıştır. Şekil 3.6.3.10'da benzerlik skorları ile gösterilen nitelik vektörleri için örnek gösterilmiştir. Eğitim kümesindeki proteinler için etiket değerleri alejen olanlar için 1 alerjen olmayanlar için -1 değerleri ile eşleştirilmiştir ve istenilen formatta dosyaya kaydedilmiştir. Test edilecek olan dosya için tüm test proteinlerinin tüm eğitim proteinleri ile hesaplanan benzerlik değerleri kaydedilmiş ve DVM çıktısı, test proteinlerin diskriminant skorları elde edilmiştir.

corner	>Allergen1	>Allergen2	>Allergen7	>Allergen11	>Allergen13...
>Allergen1	137	-17	-13	-49	-45 ...
>Allergen2	-17	144	-32	-46	-21 ...
>Allergen7	-13	-32	131	-56	-57 ...
>Allergen11	-49	-46	-56	138	-42 ...
>Allergen13	-45	-21	-57	-42	141 ...
...					

Şekil 3.6.3.10 Benzerlik Skorları

Performans ölçümleri

Bu çalışmada kullanılan çeşitli metodlar için eşitliklerde verilen denklemler esas alınarak performans ölçümleri gerçekleştirilmiştir.

- Eşitlik 3.31'de duyarlılık *(sensitivity)* değeri hesaplanmıştır. Bu değer doğru tahmin edilen alerjenlerin yüzdesidir.
- Eşitlik 3.32'de belirlilik (specificity) değeri hesaplanmıştır. Bu değer doğru tahmin edilen alerjen olmayanların (nonalergen) yüzdesidir.
- Eşitlik 3.33'te doğruluk (accuracy) değeri hesaplanmıştır. Bu değer doğru olarak tahmin edilen proteinlerin oranıdır.
- Eşitlik 3.34'de PPV (pozitif tahmin değeri), doğru pozitif tahmin olasılığı hesaplanmıştır.
- Eşitlik 3.35'de NPV (negatif tahmin değeri), doğru negatif tahmin olasılığı hesaplanmıştır.
- Eşitlik 3.36'da MCC (Matthew's Correlation Coefficient) değeri hesaplanmıştır.

Denklemlerde yer alan parametrelerin Şekil 3.6.3.11'de değerlendirme yönteminde kullanılmaları gösterilmiştir.

1. TN doğru negatif,
2. FN yanlış negatif,
3. TP doğru pozitif,
4. FP yanlış pozitiftir.

Sistem / Gerçek	Alerjen(+)	Alerjen Olmayan(-)
Alerjen (+)	Doğru -Pozitif	Yanlış - Negatif
Alerjen Olmayan (-)	Yanlış - Pozitif	Doğru - Negatif

Şekil 3.6.3.11 Değerlendirme Yöntemi

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \times 100\% \tag{3.31}$$

$$\text{Belirlilik} = \frac{TN}{TN+FP} \times 100\% \tag{3.32}$$

$$\text{Doğruluk} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \tag{3.33}$$

$$PPV = \frac{TP}{TP+FP} \tag{3.34}$$

$$NPV = \frac{TN}{TN+FN} \tag{3.35}$$

$$MCC = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.36}$$

4. SONUÇLAR

Yapılan çalışmada, K-En Yakın Komşu, Bulanık K-En Yakın Komşu ve Destek Vektör Makineleri, farklı dizilim yöntemleri ile gerçekleştirilmiştir. Bulanık K-En Yakın Komşu yöntemi, protein sınıflandırması için ilk kez yapılan çalışma ile kullanılmıştır. K-En Yakın Komşu Yönteminin dizilim benzerlik skorları ile uygulanması ilk defa bu çalışma ile gerçekleştirilmiştir. Ayrıca, Destek Vektör Makineleri için yapılan çalışma ile ilk kez tripeptit bileşim yöntemi, aminoasit bileşim yöntemi ile birlikte denenmiştir. Benzerlik skorları kullanılırken, dizilim verisinin tamamı ve dizilim verisinin ilk 20 amino asiti kullanılmıştır. İlk 20 amino asit seçilmesinde en önemli neden, alerjenliğin, protein diziliminin ilk 20 amino asitlik bölümünde (baş tarafında) yer alıp almadığı sorusuna yönelik almak istediğimiz cevaptır. Yöntemler gerçekleştirilerek sonuçlar karşılaştırmalı olarak incelenmiştir.

4.1 K-En Yakın Komşu ve Bulanık K-En Yakın Komşu

K-En Yakın Komşu ve Bulanık K-En Yakın Komşu yöntemleri için girdi olarak oluşturulan nitelik vektörlerinin her bir boyutu sıklık yüzdeleri ile gösterilmiştir. Vektör boyutları amino asit bileşimi kullanılarak 20 boyutlu, dipeptit bileşimi kullanılarak 400 boyutlu ve dipeptit ve amino asit bileşimi biraraya getirlerek 420 boyutlu nitelik vektörleri olarak belirlenmiştir.

K-En Yakın Komşu yöntemi için farklı dizilim yöntemleri uygulanarak elde edilen performans ölçüm sonuçları Çizelge 4.1.1'de karşılaştırmalı olarak verilmiştir. KNN sütunu uygulanan yöntem bilgisini göstermektedir. KNN*X*(*Harf*) şeklinde olan gösterimler de K=X'dir. Yani X adet en yakın komşu olduğu bilgisini verir. *Harf* değeri ise bize farklı dizilim gösterimlerinden hangisinin kullanıldığını ifade eder. Örneğin; KNN5(A), K-En Yakın Komşu yöntemi için K=5 seçildiğini ve amino asit bileşim yöntemi kullanılarak protein diziliminin 20 boyutlu vektör olarak ifade edildiğini göstermektedir. KNN kullanılarak, %69,94 doğruluk değerine ulaşılmıştır. Bu değerde, alerjenlerin %70,34'ü %78,57 belirlilik ile doğru tahmin edilmiştir. KNN yöntemi için en iyi sonuçlar dipeptit bileşim yöntemi kullanılarak elde edilmiştir.

KNN yöntemi için amino asit bileşim yöntemi dipeptit bileşim yöntemine göre düşük doğruluk değerine sahiptir. KNN yönteminde K değerinin 2 farklı değeri için performans sonuçları hesaplanmıştır. Çizelgeden de görüleceği gibi komşu sayısı arttıkça aynı nitelik vektörü kullanılarak gerçekleştirilen uygulamalar için doğruluk değerlerinde artış gözlemlenmişitr. Çizelgede kalın harflerle işaretlenmiş satır uygulanan yöntemler içerisinde bulunan en iyi sonucu ifade etmektedir.

Bulanık K-En Yakın Komşu yönteminde K-En Yakın Komşu yöntemi için gerçekleştirilen uygulamada kullanılan farklı dizilim gösterimlerinin aynıları kullanılmıştır. K-En Yakın Komşu ile Bulanık K-En Yakın Komşu yöntemlerini protein sınıflandırılması için karşılaştırmak amaçlanmaktadır. Bulanık K-En Yakın Komşu yöntemi için 20 boyutlu amino asit, 400 boyutlu dipeptit bileşim yöntemleri denenmiş ve bunlara ek olarak her iki yöntemin birlikte kullanıldığı 420 boyutlu vektörler kullanılarak performans değerlendirmesi karşılaştırmalı olarak Çizelge 4.1.2'de verilmiştir. Bulanık KNN sütununda kullanılan yöntemler belirtilmiştir. BKNN*X*(*Harf*) şeklinde olan gösterimler de K=X'dir. Yani X adet en yakın komşu olduğu bilgisini verir. *Harf* değeri ise bize farklı dizilim gösterimlerinden hangisinin kullanıldığını ifade eder. BKNN yönteminde en iyi sonuç %74,33 doğruluk değeri ile elde edilmiştir. Bu değerde 5 en yakın komşu için alerjenlerin %77,80'ini, %85,14 belirlilik değeri ile doğru tahmin edilmiştir. Amino asit bileşim yöntemi sonucunda elde edilen performans değerleri dipeptit bileşim ile elde edilenlerden daha iyi olduğu sonuçlardan gözlemlenmektedir. Çizelgede kalın harflerle işaretlenmiş satır uygulanan yöntemler içerisinde bulunan en iyi sonucu ifade etmektedir.

K-en yakın komşu ve Bulanık K-en yakın komşu yöntemleri için performans ölçümleri Çizelge 4.1.3'te karşılaştırmalı olarak gösterilmiştir. Sonuçlara bakılarak, K-En yakın komşu ve Bulanık K-en yakın komşu yöntemleri için Bulanık K-en yakın komşu yönteminin sınıflandırma için daha iyi sonuçlar verdiği gözlemlenmiştir. K değeri 5 seçildiğinde ve amino asit dizilimi kullanıldığında Bulanık K-en yakın komşu yönteminin %74,33 doğruluk değeri ile sınıflandırma yaptığı gözlemlenmiştir. Benzer şekilde K değeri 5 seçildiğinde ve amino asit dizilimi kullanıldığında K-En Yakın Komşu Yöntemi %64,06 doğruluk değeri ile

sınıflandırma yapmaktadır. İki yöntem arasında %10 oranında doğruluk değerinde artış gözlemlenmiştir.

Bulanık K-En Yakın komşu yönteminde, amino asit ve dipeptit bileşimlerinin birlikte kullanılması dipeptit bileşimin tek başına kullanılmasından daha iyi sonuçlar verdiği gözlemlenmektedir.

Çizelge 4.1.1 K-En Yakın Komşu performans değerlendirmesi

KNN	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
Knn5(A)	49,84	75,71	64,06	63,07	64,95	26,72
Knn10(AD)	48,45	78,57	65,00	64,73	65,14	28,40
Knn5(AD)	47,05	81,29	65,86	67,85	65,16	30,57
Knn5(D)	45,72	85,44	67,11	70,13	68,87	34,30
Knn10(D)	**59,37**	**78,57**	**69,94**	**70,34**	**70,80**	**39,42**

*A- amino asit bileşimin kullanıldığını ifade etmektedir.

*D- dipeptit bileşimin kullanıldığını ifade etmektedir.

*AD- amino asit ve dipeptit bileşimlerin birlikte kullanıldığını ifade etmektedir.

Çizelge 4.1.2 Bulanık K-En Yakın Komşu performans değerlendirmesi

Bulanık KNN	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
BKNN5(D)	40,21	95,43	70,56	85,91	66,63	42,84
BKNN10(D)	45,96	93,14	71,89	83,55	68,29	44,76
BKNN10(AD)	57,64	85,43	72,92	77,66	71,29	45,85
BKNN5(AD)	55,20	89,29	73,94	81,76	71,32	48,46
BKNN5(A)	**61,13**	**85,14**	**74,33**	**77,80**	**72,97**	48,43

*A- amino asit bileşimin kullanıldığını ifade etmektedir.

*D- dipeptit bileşimin kullanıldığını ifade etmektedir.

*AD- amino asit ve dipeptit bileşimlerin birlikte kullanıldığını ifade etmektedir.

Çizelge 4.1.3 K-En Yakın Komşu ve Bulanık K-En Yakın Komşu Performans Değerlendirmesi.

Yöntem	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
Knn5(A)	49,84	75,71	64,06	63,07	64,95	26,72
Knn10(AD)	48,45	78,57	65	64,73	65,14	28,4
Knn5(AD)	47,05	81,29	65,86	67,85	65,16	30,57
Knn5(D)	45,72	85,44	67,11	70,13	68,87	34,3
Knn10(D)	59,37	78,57	69,94	70,34	70,8	39,42
BKnn5(D)	40,21	95,43	70,56	85,91	66,63	42,84
BKnn10(D)	45,96	93,14	71,89	83,55	68,29	44,76
BKnn10(AD)	57,64	85,43	72,92	77,66	71,29	45,85
BKnn5(AD)	55,2	89,29	73,94	81,76	71,32	48,46
BKnn5(A)	**61,13**	**85,14**	**74,33**	**77,8**	**72,97**	**48,43**

*A- amino asit bileşimin kullanıldığını ifade etmektedir.

*D- dipeptit bileşimin kullanıldığını ifade etmektedir.

*AD- amino asit ve dipeptit bileşimlerin birlikte kullanıldığını ifade etmektedir.

Benzerlik skorları kullanılarak yapılan K-En yakın komşu metodu ile maksimum değeri seçme metodları karşılaştırmalı olarak Çizelge 4.1.4'te verilmiştir. Yapılan uygulama metodları deney sütununda belirtilmiştir. H harfi benzerlik için tüm dizilim verisinin kullanıldığını temsil etmektedir. K5-H, 5 en yakın komşu için tüm dizilim kullanarak hesaplanan benzerlik değerlerine göre gerçekleştirilen uygulamayı ifade etmektedir. KX-Y, K=X en yakın komşu değerini ifade etmektedir. Y değişkeni ise protein diziliminin en baştan kaç amino asitinin kullanıldığını göstermektedir. Örneğin; K5-10, 5 en yakın komşu için dizilimin ilk 10 değeri kullanılarak hesaplanan benzerlik bilgisine göre gerçekleştirilen uygulamayı göstermektedir.

Max ile ifade edilmek istenen K-En yakın komşu yerine bulunan en yüksek benzerlik değerine göre, test edilen örneğin de aynı sınıftan olduğunun kabulünü yapan metod için kullanılmıştır.

Benzerlik skorları (PAM70) kullanılarak tüm dizilim için yapılan benzerlik analiz sonuçlarından yaklaşık olarak %60 doğruluk elde edilirken, protein diziliminin ilk 10 amino asiti ile hesaplanan benzerlik değerlerinde yaklaşık olarak %80 doğruluk elde edilmiştir. Çizelgeden de görüleceği gibi en iyi sonuç K değeri 20 alınarak ve dizilimin ilk 10 amino asiti kullanılarak gerçekleştirilen uygulama sonucunda %83,83 doğruluk olarak bulunmuştur. Alerjen protein tahmin etme oranları incelenecek olursa, tüm dizilim verisi ile gerçekleştirilen KNN uygulaması sonucunda alerjen proteinlerin %44,59'u %56,14 belirlilik ve %45,07 doğru olma olasılığı ile doğru tahmin edilirken, dizilim verisinin ilk 10 amino asit kullanılarak gerçekleştirilen KNN için alerjen proteinlerin %71,26'sı, %92,26 belirlilik ve %88,22 doğru olma olasılığı ile doğru tahmin edilmiştir. Yaklaşık olarak %43 oranla daha doğru tahmin yapılmaktadır.

Maksimum benzerlik değerini bularak, sınıflandırma işlemini bu benzerlik değerinin ait olduğu protein etiketine göre gerçekleştiren yöntem kendi içerisinde incelendiğinde; dizilim verisinin tümü kullanılarak yapılan uygulama sonucunda %62,64 doğruluk elde edilirken, dizilim verisinin ilk 20 amino asiti ile doğruluk değeri %77,47 olarak gözlemlenmektedir. Protein diziliminin ilk 10 verisi kullanılarak gerçekleştirilen uygulama da ise %81,4 doğruluk değeri saptanmıştır.

K-En Yakın Komşu yönteminde, seçilen 20 komşu ve ilk 10 dizilim bilgisi kullanılarak, benzerlik skorları ile gerçekleştirilen uygulamalar sonucunda alerjenlerin %73,71'ini, %92,14'lük belirlilik ve %89,62 doğru olma olasılığı ile tahmin edilmiştir.

Çizelge 4.1.4 PAM70 ile elde edilen benzerlik sonuçları üzerinde yapılan uygulamaların performansı

Deney	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
K5-H	44,59	56,14	50,94	45,07	55,5	0,65
Max-H	50	73	62,64	60,52	64,07	23,78
K5-20	54,71	72,71	64,6	61,29	67,13	27,91
Max-20	65,85	87	77,47	80,42	76,14	54,63
Max-10	71,78	89,29	81,4	84,25	80,19	62,68
K10-20	75,09	87,43	81,87	83,67	81,54	63,82
K5-10	71,26	92,29	82,81	88,22	80,31	65,92
K10-10	73,88	91,71	83,68	88,37	81,85	67,77
K20-10	**73,71**	**92,14**	**83,83**	**89,62**	**81,65**	**68,38**

*K5 K-En Yakın komşu için K değerinin 5 seçilmesini ifade etmektedir.

*K10 K-En Yakın komşu için K değerinin 10 seçilmesini ifade etmektedir.

* K20 K-En Yakın komşu için K değerinin 20 seçilmesini ifade etmektedir.

4.2 DVM Yöntemi

İlk olarak bütün proteinlerin amino asit bileşim yöntemi ile oranları hesaplanmıştır. Daha sonra bu oranlar DVM'nin eğitim ve test uygulamaları için 20 boyutlu girdi vektörü olarak kullanılmıştır. DVM tabanlı metodların performansı, yüksek doğrulukla beraber hemen hemen eşit duyarlılık ve belirliliğe ulaşmak için DVM parametrelerinin ayarlanmasıyla optimize edilmiştir.

Çizelge 4.2.1'de DVM kullanılarak gerçekleştirilen, amino asit bileşim yöntemi performans sonuçları gösterilmiştir. T sütunu seçilen eşik değerlerini göstermektedir. Bu değerler için duyarlılık, belirlilik, doğruluk, PPV pozitif tahmin değeri, NPV negatif tahmin değeri ve MCC korelasyon katsayısı değerleri hesaplanmıştır. Çizelge 4.2.1'de görüldüğü gibi bu yöntemi kullanarak %84,54 doğruluk elde edilmiştir. Bu değerde alerjenlerin %82,22'si, %87,14 belirlilik ve %84,84 doğru olma olasılığı tahmin edilmiştir. Bu yöntem alerjenlerin %94,19'unu

%58,88 belirlilik ile doğru tahmin etmiştir. Aynı zamanda alerjenlerin %56,98'ini de %94,82 belirlilik ile ve %97,38 doğru olma olasılığıyla doğru tahmin etmiştir.

Benzer bir şekilde dipeptid bileşim yöntemini kullanarak DVM tabanlı bir metod geliştirilmiştir. Çizelge 4.2.2'de DVM kullanılarak gerçekleştirilen, amino asit bileşim yöntemi performans sonuçları gösterilmiştir. Bu yöntem kullanılarak %82,19 doğruluk değeri elde edilmiştir. Bu değerde alerjenlerin %78,22'si, %86,86 belirlilik ve %85,72 doğru olma olasılığı ile tahmin edilmiştir. Ayrıca dipeptit bileşim yöntemi kullanılarak, alerjenlerin %95,52'si, %59,47 belirlilik ve %17,24 doğru olma olasılığı aynı zamanda, alerjenlerin %57,23'ü %97,29 belirlilik ve %98,78 doğru olma olasılığı ile tahmin edilmiştir.

Çizelge 4.2.1 ve 4.2.2'den alınan sonuçlara göre amino asit bileşim yöntemi kullanılarak gerçekleştirilen DVM yöntemi, dipeptid bileşim yöntemi ile gerçekleştirilenden daha doğru sınıflandırma yapmaktadır denilebilir.

Ayrı ayrı incelenen amino asit ve dipeptit bileşim yöntemi birlikte kullanılarak elde edilen sonuçlar çizelge 4.2.3'te gösterilmiştir. İki yöntemin birlikte kullanılması ile alerjen proteinlerin doğru sınıflandırılmasında artış saptanması beklenmektedir. Ancak çizelgeden görüleceği gibi sonuçlarda, diğer yöntemler ile kıyaslandığında beklenen oranlarda artışlar gözlemlenmemiştir. Bu yöntem kullanıldığında alerjen proteinlerin %94,89'u, %58,83 belirlilik ile doğru tahmin edilmiştir. Aynı zamanda alerjenlerin %53,70'ini %97,45 belirlilik ile doğru tahmin edilmiştir. Bu yöntem ile %84,07'lik doğruluk değeri elde edilmiştir.

Son olarak, tripeptit bileşim yöntemi ile amino asit bileşim yöntemleri birlikte kullanılarak boyutu 8020 olan vektör elde edilerek her bir boyut için oranlar hesaplanmıştır. DVM yöntemi ile gerçekleştirilen uygulama sonuçları Çizelge 4.2.4'te verilmiştir. T sütunu seçilen eşik değerlerini ifade etmektedir. -0,2 eşik değeri için en iyi sonuç elde edilmiştir. Bu yöntemle, %86,11 doğruluk değerine ulaşılmıştır. Bu değerde, alerjenlerin %86,20'si, %87,28 belirlilik ve %84,31 doğru olma olasılığı ile tahmin edilmiştir. Aynı zamanda bu yöntem uygulanarak, alerjenlerin %96,47'si %57,82 belirlilik ile doğru tahmin edilmiştir. Ayrıca

alerjenlerin %62,45'i de %96,29 belirlilik ve %98,08 doğru olma olasılığı ile doğru tahmin edilmiştir.

Protein dizilim verisinin bileşim yöntemleri kullanılarak nitelik vektörleri şeklinde ifade edilerek DVM ile gerçekleştirilen uygulamaları sonucunda, amino asit bileşim yönteminin, dipeptit bileşim yönteminden daha doğru sınıflandırma yaptığı sonucuna varılmıştır. DVM ile gerçekleştirilen uygulamalar ile performans değerlendirmeleri sonucunda en iyi sınıflandırma, amino asit ve tripeptit bileşim yöntemleri birlikte kullanıldığında elde edilmiştir ve doğruluk değeri %86,11'dir.

Benzerlik skorları kullanılarak oluşturulan nitelik vektörleri ile DVM uygulaması gerçekleştirilmiştir. Protein dizilim verisinin tümü kullanılarak hesaplanan benzerlik skorları DVM ile uygulanmıştır. Çizelge 4.2.5'te DVM ile gerçekleştirilen uygulama için performans değerlendirilmesi sonuçları verilmiştir. Bu yöntemle, %69,62 doğruluk değeri elde edilmiştir. Alerjenlerin %66,98'i, %54,90 belirlilik ile %5,91 doğru olma olasılığı ile doğru tahmin edilmiştir. Aynı zamanda alerjenlerin %50,80'i, %60,82 belirlilik değeri ile %88,16 doğru olma olasılığı ile doğru tahmin edilmiştir.

Çizelge 4.2.6'da, benzerlik skorları kullanılarak protein dizilim verisinin ilk 20 amino asiti kullanılarak oluşturulan nitelik vektörleri kullanılarak gerçekleştirilen DVM yöntemi için performans sonuçları verilmiştir, %74,89'luk doğruluk değeri elde edilmiştir. Bu yöntemle alerjenlerin %89,94'ü, %57,83 belirlilik ve %12,88 doğru olma olasılığı ile doğru tahmin edilmiştir. Aynı zamanda alerjenlerin %57,15'i, %80,23 belirlilik değeri ve %82,76 doğru olma olasılığı ile doğru tahmin edilmiştir.

Benzerlik skorları ile oluşturulan nitelik vektörleri ile gerçekleştirilen DVM için dizilim verisinin tümünün alerjen proteinleri sınıflandırmada etkili olmadığı sonucuna varılabilir. 20 amino asit kullanılarak gerçekleştirilen yöntem ile doğru tahmin özelliği arttırılmış ve daha yüksek doğruluk değeri elde edilmiştir.

Çizelge 4.2.1 DVM ile gerçekleştirilen amino asit bileşim yöntemi performans değerlendirmesi

T	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
1	94.19	58.88	61.38	15.52	99.00	27.28
0,8	93.75	62.07	65.94	26.66	98.14	36.90
0,6	91.20	67.98	72.76	43.91	96.43	48.67
0,4	90.12	74.29	78.65	59.41	94.43	58.81
0,3	88.91	78.02	81.56	67.77	92.86	63.68
0,2	86.06	80.63	82.42	73.35	89.86	64.90
0,1	84.44	84.14	83.99	79.79	87.43	67.89
0	**82.22**	**87.14**	**84.54**	**84.84**	**84.29**	**69.24**
-0,1	78.65	89.09	83.44	88.50	79.29	67.76
-0,2	75.37	90.59	82.03	90.94	74.71	65.80
-0,3	71.49	91.15	79.44	92.33	68.86	61.90
-0,4	67.53	92.42	76.61	94.08	62.29	58.11
-0,6	61.25	93.54	70.49	95.99	49.57	49.89
-0,8	56.98	94.82	65.46	97.38	39.29	43.45
-1	53.45	96.84	60.52	98.78	29.14	37.25

Çizelge 4.2.2 DVM ile gerçekleştirilen dipeptit bileşim yöntemi performans değerlendirmesi

T	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
1	95.52	59.47	62.40	17.24	99.43	30.07
0,8	92.54	61.27	64.83	23.86	98.43	34.36
0,6	89.99	64.95	69.15	35.54	96.71	41.78
0,4	87.93	69.54	73.71	49.13	93.86	49.53
0,3	86.83	72.48	76.30	56.62	92.43	53.85
0,2	84.94	75.06	78.03	63.42	90.00	56.58
0,1	83.44	77.76	79.52	69.52	87.71	59.17
0	82.10	80.12	80.54	74.57	85.43	61.09
-0,1	79.65	83.11	81.09	79.98	82.00	62.36
-0,2	**78.22**	**86.86**	**82.19**	**85.72**	**79.29**	**65.04**
-0,3	76.01	89.37	81.95	89.38	75.86	65.30
-0,4	73.54	90.99	80.93	91.82	72.00	64.17
-0,6	68.83	95.06	78.26	96.34	63.43	61.76
-0,8	63.14	95.78	72.84	97.74	52.43	54.29
-1	57.23	97.29	65.94	98.78	39.00	45.23

Çizelge 4.2.3 DVM ile gerçekleştirilen amino asit +dipeptit bileşim yöntemi performans değerlendirmesi

T	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
1	94.89	58.83	61.38	15.33	99.14	27.61
0,8	95.37	62.99	66.95	28.22	98.71	38.87
0,6	93.61	66.71	71.27	39.02	97.71	46.56
0,4	91.12	71.88	76.45	53.14	95.57	55.21
0,3	90.12	75.45	79.59	61.85	94.14	60.52
0,2	86.90	78.70	81.40	69.69	91.00	63.07
0	**82.86**	**85.46**	**84.07**	**82.58**	**85.29**	**68.09**
-0,2	77.38	90.50	83.21	90.42	77.29	67.78
-0,3	73.73	92.47	81.48	93.38	71.71	65.63
-0,4	70.16	92.69	78.81	93.90	66.43	61.55
-0,6	62.59	94.66	72.22	96.69	52.14	52.84
-0,8	57.91	96.05	66.80	98.08	41.14	45.90
-1	53.70	97.45	60.99	98.95	29.86	38.25

Çizelge 4.2.4 DVM ile gerçekleştirilen amino asit +tripeptit bileşim yöntemi performans değerlendirmesi

T	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
1	96.47	57.82	59.81	11.31	99.57	23.51
0,8	96.25	61.47	65.22	23.67	99.29	35.89
0,6	95.42	66.23	71.03	37.43	98.57	46.64
0,4	93.18	70.92	75.74	50.49	96.43	54.56
0,3	92.98	73.01	77.63	55.55	95.71	57.92
0,2	90,90	75.66	79.75	62.00	94.29	61.10
0,1	90,33	79,21	82,65	69,67	93,29	66,13
0	88,45	81,41	83,60	74,21	91,29	67,62
-0,1	87,15	84,19	84,77	79,61	89,00	69,95
-0,2	**86,20**	**87,28**	**86,11**	**84,31**	**87,57**	**72,67**
-0,3	82,36	89,03	85,32	87,63	83,43	71,22
-0,4	80,37	90,32	84,85	89,72	80,86	70,63
-0,6	74.62	93.59	82.34	94.42	72.43	67.50
-0,8	68.46	95.01	77.79	96.51	62.43	61.10
-1	62.45	96.29	72.06	98.08	50.71	53.42

Çizelge 4.2.5 DVM ile gerçekleştirilen tüm dizilim verisi kullanılarak hesaplanan benzerlik skorları yöntemi performans değerlendirmesi

T	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
1	66,98	54,90	54,32	5,91	94,00	2,64
0,8	69,52	55,58	55,34	9,22	93,14	6,56
0,6	78,03	56,14	56,20	12,18	92,29	10,90
0,4	79,43	56,75	56,75	16,35	89,86	12,10
0,2	76,93	57,08	56,67	22,27	84,86	14,41
0	75,94	59,12	59,18	32,54	81,00	21,28
-0,2	**81,99**	**68,29**	**69,62**	**51,69**	**84,29**	**41,55**
-0,4	69,98	73,73	68,68	73,80	64,43	40,69
-0,6	66,66	72,74	67,98	83,06	55,57	38,97
-0,8	62,28	75,84	66,73	89,88	47,71	37,75
-1	50,80	60,82	55,03	88,16	27,86	15,86

Çizelge 4.2.6 DVM ile gerçekleştirilen ilk 20 amino asit kullanılarak hesaplanan benzerlik skorları yöntemi performans değerlendirmesi

T	Duyarlılık	Belirlilik	Doğruluk	PPV	NPV	MCC
1	89,94	57,83	59,58	12,88	97,86	22,07
0,8	88,59	59,58	61,74	17,80	96,86	24,28
0,6	85,02	61,44	64,52	27,18	95,14	32,03
0,4	85,29	63,66	67,11	34,50	93,86	36,98
0,2	85,10	67,01	70,33	44,08	91,86	42,96
0	84,08	70,56	73,16	53,31	89,43	48,13
-0,2	81,04	73,66	74,65	60,98	85,86	50,52
-0,4	**76,31**	**76,42**	**74,89**	**67,43**	**81,00**	**50,49**
-0,6	68,56	74,82	70,57	66,92	73,57	41,89
-0,8	61,21	77,34	67,19	74,58	61,14	37,02
-1	57,15	80,23	64,29	82,76	49,14	34,39

5. TARTIŞMA

Bu çalışmanın amacı, alerjen proteinlerin, özellikle gıda alerjenlerinin tahmini için değişik yöntemler denemektir. Dünya Sağlık Örgütü ve Gıda ve Tarım Örgütü kurumlarının bu amaçla hazırladıkları rehberlerde önerilen yöntemlerin çoğunlukla yarı-otomatik gerçekleştirilen ve tahmin yeterliliği düşük olan yöntemler olmasından dolayı bu çalışmada otomatik yöntemler denenerek kuvvetli ve zayıf yanlarının anlaşılması hedeflenmiştir.

Uygulanan yöntemler K-En Yakın Komşu, Bulanık K-En Yakın Komşu ve Destek Vektör Makineleridir. Ayrıca benzerlik skorlarından faydalanılırken bu yöntemlere ek olarak sınıflandırma, maksimum benzerlik skoru gözetilerek yapılmıştır.

Protein dizilimleri amino asit bileşim yöntemi, dipeptit bileşim yöntemi, iki yöntemin birlikte kullanılması ve tripeptit bileşim yöntemi ile amino asit bileşim yönteminin birlikte kullanılması ile vektörel şekilde ifade edilmişlerdir. Ayrıca dizilim benzerlik skorları ile de nitelik gösterimleri gerçekleştirilmiştir.

K-En Yakın Komşu yöntemi için amino asit ve dipeptit bileşim yöntemleri kullanılmıştır. Dipeptit bileşim yöntemi, amino asit bileşim yöntemine göre daha çok bilgi içermektedir. Sonuçlarda da görüldüğü gibi K-En Yakın Komşu yöntemi, dipeptit bileşim yöntemi ile uygulandığında, amino asit bileşim yöntemine göre daha iyi sonuçlar üretmiştir. Amino asit ve dipeptit bileşim yöntemleri birlikte kullanılarak daha çok bilgi içeren nitelik vektörlerinden daha iyi sonuçlar alınabileceği düşünülmüştür. Uygulama sonucunda, bu yöntemin doğruluk değerlerinin amino asit ve dipeptit bileşim yöntemleri arasında sonuçlar verdiği gözlemlenmiştir.

K-En Yakın Komşu yöntemi, benzerlik skorları kullanılarak uygulama gerçekleştirildiğinde amino asit ve dipeptit bileşim yöntemlerinden yaklaşık olarak %10 daha doğru sonuçlar verdiği performans sonuçlarında gözlemlenmiştir. Protein diziliminin tamamı, ilk 20 amino asit ve ilk 10 amino asit üzerinden hesaplanan benzerlik skorları için en iyi sonuç ilk 10 değer kullanılarak bulunmuştur. K-En Yakın Komşu için en iyi sonuç K=20 seçildiğinde elde

edilmiştir. Dizilim verisinin ilk 20 amino asiti tüm dizilim bilgisi kullanılarak elde edilen sonuçlardan daha performanslıdır. Benzerlik skorları kullanıldığında K değerinin artımına bağlı olarak sonuçlarda iyileşme saptanmıştır. Bileşim yöntemleri kullanıldığında K'nın azalan değerlerinde artan doğruluk oranları, benzerlik skorları kullanıldığında K'nın artmasına bağlı olarak artmıştır. K-En Yakın Komşu yöntemi benzerlik skorları kullanılarak uygulandığında, doğruluk değerinin amino asit ve dipeptit bileşiminden daha yüksek olduğu görülmüştür. Alerjen protein tahmin oranının, dizilim bilgisinin ilk 10 değeri kullanılarak gerçekleştirilen uygulama için diğer sonuçlara göre maksimum değerde çıktığı gözlemlenmiştir.

K-En Yakın Komşu yöntemi yerine benzerlik skorları kullanılarak maksimum benzerlik skoruna bağlı sınıflandırma işlemi yapıldığında, yine en iyi sonucun ilk 10 amino asit için alındığı gözlemlenmiştir. Ancak tüm sonuçlara bakıldığında K-En Yakın Komşu yönteminin performansının daha iyi olduğu sonucuna varılmıştır.

Bulanık K-En Yakın Komşu yöntemi için amino asit bileşim ve dipeptit bileşim uygulanarak alınan sonuçlara bakıldığında dipeptit bileşim, amino asit bileşime oranla daha fazla bilgi barındırmasına rağmen, amino asit bileşim kullanılarak gözlemlenen sonuçlarda %4 oranında daha doğru sınıflandırma yapılmıştır.

K-En Yakın Komşu yöntemi ve Bulanık K-En Yakın Komşu yöntemini karşılaştırmalı olarak değerlendirmek istersek; Bulanık K-En Yakın Komşu yöntemi alerjen proteinleri K-En Yakın Komşu yöntemine oranla daha doğru tahmin etmiştir. Amino asit bileşimi her iki yöntemle değerlendirildiği zaman Bulanık K-En Yakın Komşu yönteminin, alerjen proteinlerin sınıflandırılmasında, yaklaşık olarak %10 oranla daha doğru sınıflandırma yaptığı sonuçlardan görülmektedir. Amino asit bileşim yöntemi, Bulanık K-En Yakın Komşu için en iyi doğruluk değerini verirken, K-En Yakın Komşu yöntemi ile kullanıldığında en kötü doğruluk değerini, her iki yöntemde de K=5 iken vermiştir. K=5 için her iki yöntemde dipeptit bileşim kullanılarak alınan sonuçlarda, Bulanık K-En Yakın Komşu yönteminin alerjen olan proteinleri tahmin etme değeri daha yüksektir. Bu sonuçlar proteinlerin sınıflandırılmasında yalnızca sınıfa aittir ya da ait değildir verisi ile birlikte kullanılan üyelik değerinin sonuçları arttırdığını ifade etmektedir.

Bir makine öğrenme tekniği olan DVM, bu çalışmada alerjen proteinlerin tespiti için kullanılmıştır. Amino asit ve dipeptit bileşimleri kullanılarak DVM tabanlı yöntem geliştirilmiştir. Bu yaklaşımın en büyük avantajı kullanıcının gereksinimi olan eşik değerini seçerek yüksek belirlilik ve duyarlılık ile alerjen proteinlerin tahminini sağlamasıdır. DVM ile amino asit bileşim, dipeptit bileşim, ikisinin birlikte kullanılması, tripeptit bileşim ve amino asit bileşimin birlikte kullanılaması ve son olarak benzerlik skorları uygulamaları gerçekleştirilmiştir. Benzerlik skorları uygulanırken dizilim verisinin bütünü ve ilk 20 amino asit için hesaplanan benzerlik değerleri kullanılmıştır.

Amino asit bileşim yöntemi kullandığında en iyi performans için seçilen eşik değerinde, alerjen ve alerjen olmayan proteinlerin tahmin edilme oranları hemen hemen aynı çıkmıştır. Aynı eşik değeri için dipeptit bileşim yöntemi uygulandığında alerjen olmayan proteinlerin tahmin oranlarının yaklaşık olarak değişmediği gözlemlenirken, alerjen protein tahmin oranının %10 azaldığı görülmüştür. Yine aynı eşik değeri için tripeptit ve amino asit birlikte kullanıldığında, alerjen olmayan proteinlerin tahmin oranının arttığı gözlemlenmiştir.

Dipeptitler amino asit bileşimlerinden daha fazla bilgi vermelerine rağmen dipeptit bileşim yönteminin performansının amino asit bileşim yönteminden daha düşük olduğu gözlemlenmiştir. Bu çalışmada kullanılan protein dizilimlerinin kısa olmasından dolayı dipeptitlerin sıklıklarının tam olarak belirlenememesinden kaynaklı olarak sonuçların bu şekilde çıktığı varsayılmıştır. Amino asit bileşimi ve dipeptit bileşimi kullanılarak yapılan çalışmalara [59] ek olarak daha fazla özellik barındırması açısından dipeptit bileşim yöntemi amino asit bileşim yöntemi ile birarada kullanılarak alerjen tahmin oranının yükseldiği gözlemlenmiştir. Ayrıca aynı şekilde, daha fazla özellik barındırması açısından tirpeptit bileşim yöntemi amino asit bileşim yöntemi ile birarada kullanılarak, doğruluk değerinde artış gözlemlenmiştir.

DVM, K-En Yakın Komşu ve Bulanık K-En Yakın Komşu yöntemleri ile karşılaştırıldığında daha etkili sınıflandırma yapmaktadır. DVM'nin çok boyutlu uzayda hiper düzlemler oluşturarak farklı sınıf etiketlerini birbirinden ayırarak

sınıflandırma işlemini gerçekleştiren bir sınıflandırma metodudu olmasının daha doğru tahminler yapmasında etkili olduğu sonucuna varılmıştır.

Benzerlik skorları ile uygulanan DVM yönteminde, dizilim verisinin tümü ve ilk 20 dizilim verisi kullanılarak benzerlik skorları hesaplanmıştır. İlk 20 dizilim verisi kullanılarak gerçekleştirilen uygulama sonucunda doğruluk değerinin daha iyi olduğu gözlemlenmiştir. İlk 20 dizilim verisi kullanıldığında alerjen tahmin oranı, tüm dizilim verisi kullanıldığında alınan orandan daha yüksektir. Bu sonuçlara göre alerjenliği belirleyen özelliklerin dizilim verisinin başlarında yeraldığı varsayımı yapılabilir.

Genetiği değiştirilmiş organizmaların ve biyo-ilaçların artmasından dolayı alerjen proteinlerin tahmin edilmesi önem kazanması ve son yıllarda bu yönde çalışmaların sıkça uygulanmasından dolayı yapılan bu çalışmada, farklı sınıflandırma yöntemleri, çeşitli dizilim verileri için uygulanmıştır. Yapılan çalışma ile alerjen proteinlerin otomatik sınıflandırılması problemine farklı çözüm yolları sunulmuş ve sonuçlar karşılaştırmalı olarak değerlendirilmiştir. Bu çalışmada uygulanan yöntemler farklı sınıflandırma problemleri için de gerçekleştirilebilir.

KAYNAKLAR

[1] Aalberse, R.C., Structural biology of allergens, J. Allergy Clin. Immunol., vol.106, s.228–238, 2000.

[2] Aalberse, R.C., Akkerdaas, J.H., van Ree, R., Crossreactivity of IgE antibodies to allergens, Allergy, vol.56, no.6, s.478–490, 2001.

[3] Alinorm 03/34, Joint FAO/WHO Food Standard Programme, Codex Alimentarius Commission, Twenty-Fifth Session, Appendix III, Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants and Appendix IV, Annex on the assessment of possible allergenicity, Rome, Italy, s.47–60, 30 June-5 July 2003.

[4] Altschul, S., Gish, W., Miller, W., Myers, E.W., Lipman, D., A basic local alignment search tool, Journal of Molecular Biology, vol.251, s.403-410, 1990.

[5] Ayat, N. E., Cheriet, M., Remaki, L., Suen, C. Y., KMOD- A new support vector machine kernel with moderate decreasing for pattern recognition, Proceedings of the 6th Int. Conference on Document Analysis and Recognition, s.434-438, 2001.

[6] Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches, Bioinformatics, vol.14, s.48-54.

[7] Banerjee, B., Greenberger, P.A., Fink, J.N., Kurup, V.P., Conformational and linear B-cell epitopes of Asp f 2, a major allergen of Aspergillus fumigatus, bind differently to immunoglobulin E antibody in the sera of allergic bronchopulmonary aspergillosis patients, Infect. Immun., vol.67, s.2284–2291, 1999.

[8] Beezhold, D.H., Hickey, V.L., Slater, J.E., Sussman, G.L., Human IgE-binding epitopes of the latex allergen Hev b 5, J. Allergy Clin. Immunol., vol.103, s.1166–1172, 1999.

[9] Bendtsen, J.D., Jensen, L.J., Blom, N., Von, H.G. and Brunak, S., Feature-based prediction of non-classical and leaderless proteinsecretion, Protein Eng. Des. Sel., vol.17, s.349–356, 2004.

[10] Bjorklund, A.K., Soeria-Atmadja, D., Zorzet, A., Hammerling, U., and Gustafsson, M.G., Supervised identification of allergenrepresentative peptides for in silico detection of potentially allergenic proteins, Bioinformatics, vol.21, s.39–50, 2005.

[11] Boser, B.E., Guyon, I.M., and Vapnik, V., A Training Algorithm for Optimal Margin Classifiers, Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh, PA, s.144-152, 1992.

[12] Bottou, L., Cortes, C., Denker, J., Drucker, H., Guyon, I., Jackel, L., LeCun, Y., Muller, U., Sackinger, E., Simard, P., and Vapnik, V., Comparison of classifier methods: A case study in handwriting digit recognition, Proc. Int. Conf. Pattern Recognition, vol.2, s.77–87, 1994.

[13] Brusic,V., Millot, M., Petrovsky, N., Gendel, S.M., Gigonzac, O. and Stelman, S.J., Allergen databases, Allergy, vol.58, s.1093–1100, 2003.

[14] Burges, C., A Tutorial on Support Vector Machines for Pattern Recognition, Proc.Data Mining and Knowledge Discovery, U. Fayyad, Ed., Kluwer Academic, s.1–43, 1998.

[15] Cao, L.J. and Tay, F., Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting. IEEE Transactions On Neural Networks, vol. 14, no. 6, s.1506–1518, 2003.

[16] Christianini, N. and Taylor, J., An Introduction to Support Vector Machines and Other Kernel Methods. Cambridge University Press, Cambridge, 2000.

[17] Cortes, C. and Vapnik, V.N., Support vector networks, Machine Learning, vol.20, no.3, s.273-297, 1995.

[18] Cui, J., Han, L.Y., Li, H., Ung, C.Y., Tang, Z.Q., Zheng, C.J., Cao, Z.W., Chen, Y.Z., Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. Mol. Immunol., vol.44, no.4, s.514-520, 2007.

[19] Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., A model of evolutionary change in proteins, Atlas of Protein Sequence and Structure, vol.5, s.345-352, 1978.

[20] Doğan, H., Gradient networks design for clustering in novel optimization frameworks, PhD. Thesis, Dokuz Eylül University, Izmir, Turkey, 2004.

[21] FAO/WHO, Evaluation of allergenicity of genetically modified foods. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, 2001.

[22] FAO/WHO, Report of the fourth session of the codex ad hoc intergovernmental task force on foods derived from biotechnology, 2003. (http://www.codexalimentarius.net/download/report/46/Al0334ae.pdf).

[23] Gendel, S.M., The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods, Adv. Food Nutr. Res., vol.42, s.45–62, 1998.

[24] Gendel, S.M., Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods, Adv. Food Nutr. Res., vol.42, s.63–92, 1998.

[25] Gendel, S.M., Sequence analysis for assessing potential allergenicity. Ann. NY Acad. Sci., 964, s.87–98, 2002.

[26] Georgoulas, G., Georgopoulos, V.C. and Stylios, C.D., Speech Sound Classification and Detection of Articulation Disorders with Support Vector Machines and Wavelets, Engineering in Medicine and Biology Society, 28th Annual International Conference of the IEEE, s.2199-2202, 2006.

[27] Goodman, R.E., Silvanovich, A., Hileman, R.E., Bannon, G.A., Rice, E.A., Astwood, J.D., Bioinformatic methods for identifying known or potential allergens in the safety assessment of genetically modified crops, Comments Toxicol., vol.8, s.251–269, 2002.

[28] Goodman, R.E., Hefle, S.L., Taylor, S.L. and Ree, R.V., Assessing genetically modified crops to minimize the risk of increased food allergy: a review, Int. Arch. Allergy Immunol., vol.137, s.153–166, 2005.

[29] Guidance document of the Scientific Panel on Genetically Modified Organisms for the risk assessment of genetically modified plants and derived food and feed, EFSA J., vol.99, s.1– 94, 2004.

[30] Hileman, R.E., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D. and Hefle, S.L., Bioinformatic methods for allergenicity assessment using a comprehensive allergen database, Int. Arch. Allergy Immunol., vol.128, s.280–291, 2002.

[31] Hösel, V., Walcher, S., Clustering Techniques: A Brief Survey, AMS Subject Classification, 62H30, 68T10, s.62-07, Germany, 2000.

[32] Hsu, C.W., Lin, C.J., A comparison of methods for multiclass support vector machines, IEEE Trans. on Neural Networks, vol.13, s.1026 – 1027, 2002.

[33] Huang, H. and Liu, Y.H., Fuzzy support vector machines for pattern recognition and data mining, International Journal of Fuzzy Systems, vol.4, no.3, s.826-835, 2002.

[34] Hyungkeun, J., Kyunghee, L. and Sungbum, P., Eye and face detection using SVM, Intelligent Sensors, Sensor Networks and Information Processing Conference, ISSNIP '04., s.577-580, 2004.

[35] İbrikçi, T., Çakmak, A., Ersöz, İ. and Açıkkar, M., Destek Vektörlerinin Proteinlerin İkincil Yapılarını Tahmin Etmek İçin Uygulanması. BİYOMUT, National Meeting on Biomedical Engineering, İstanbul, 2004.

[36] Ivanciuc, O., Schein, C.H. and Braun, W., SDAP: database and computational tools for allergenic proteins, Nucleic Acids Res., vol.31, s.359–362, 2003.

[37] Jain, A.K., Murty, M.N., Flynn, P.J., Data Clustering: A Review, ACM Computing Surveys, vol.31, no. 3, 1999.

[38] Jansen, J.J., Kardinaal, A.F., Huijbers, G., Vlieg-Boerstra, B.J., Martens, B.P. and Ockhuizen, T., Prevalence of food allergy and intolerance in the adult Dutch population, J. Allergy Clin. Immunol, vol.93, s.446–456, 1994.

[39] Kanny, G., Moneret-Vautrin, D.A., Flabbee, J., Beaudouin, E., Morisset, M. and Thevenin, F., Population study of food allergy in France, J. Allergy Clin. Immunol., vol.108, s.133–140, 2001.

[40] Kleter, G.A. and Peijnenburg, A.A., Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE-binding linear epitopes of allergens, BMC Struct. Biol., vol.2, no.8, 2002.

[41] Knerr, S., Personnaz, L. and Dreyfus, G., Single-layer learning revisited: A stepwise procedure for building and training a neural network, Neurocomputing: Algorithms, Architectures and Applications, J. Fogelman, Ed. New York: Springer-Verlag, 1990.

[42] Lee, Y.H. and Sinko, P.J., Oral delivery of salmon calcitonin, Adv. Drug Deliv. Rev., vol.42, s.225–238, 2000.

[43] Li, K.B., Issac, P. and Krishnan, A., Predicting allergenic proteins using wavelet transform, Bioinformatics, vol.20, s.2572–2578, 2004.

[44] Li, S., Kwok, J.T., Tsang, I.W., Wang, Y., Fusing Images With Different Focuses Using Support Vector Machines, IEEE Transactions On Neural Networks, vol.15, no.6, s.1555–1561, 2004.

[45] Li, Y., Liu Q., Ruan, X., Cancer molecular classification based on support vector machines, Fifth World Congress on Intelligent Control and Automation, (WCICA), vol.6, s.5521 – 5524, 2004.

[46] Lipman, D.J. and Pearson, W.R., Rapid and sensitive protein similarity search, Science, vol.227, s.1435-1441, 1985.

[47] Mao, K.Z. and Huang, G., Neuron selection for RBF neural network classifier based on data structure preserving criterion, IEEE Trans. on Neural Networks, vol.16, no.6, s.1531-1540, 2005.

[48] Markowska-Kaczmar, U., Kubacki, P., Support vector machines in handwritten digits classification, Proceedings. 5th International Conference on Intelligent Systems Design and Applications, vol.1, s.252 – 256, 2005.

[49] Mayoraz, E. and Alpaydin, E.,, Support vector machines for multi-class Classification, IWANN'99 (June), Alicante, Spain, 833-842, 1999.

[50] Mekori,Y.A., Introduction to allergic diseases, Crit. Rev. Food Sci. Nutr., vol.36, s.1–18, 1996.

[51] Metcalfe, D.D., Astwood, J.D., Townsend, R., Sampson, H.A., Taylor, S.L., Fuchs, R.L., Assessment of the allergenic potential of foods derived from genetically engineered crop plants, Crit. Rev. Food Sci. Nutr., vol.36, s.165–186, 1996.

[52] Nieuwenhuizen, N.E. and Lopata, A.L., Fighting food allergy, Curr. Approaches, Ann. N.Y. Acad. Sci., vol.1056, s.30–45, 2005.

[53] Oğul, H. and Mumcuoğlu, E., A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets, BioSystems, 87, 2007, s.75-81, 2007.

[54] Platt, J.C., Cristianini, N. and Shawe-Taylor, J., Large margin DAG's for multiclass classification, Advances in Neural Information Processing Systems, MA, MIT Press, Cambridge, vol.12, s.547-553, 2000.

[55] Qin, Jun., He, Zhong-Shi., A SVM face recognition method based on Gaborfeatured key points, International Conference on Machine Learning and Cybernetics, vol.8, s.5144 -5149, 2005.

[56] Rabjohn, P., Helm, E.M., Stanley, J.S., West, C.M., Sampson, H.A., Burks, A.W., Bannon, G.A., Molecular cloning and epitope analysis of the peanut allergen Ara h 3, J. Clin. Invest., vol.103, s.535–542, 1999.

[57] Saha, S. and Raghava, G.P.S., AlgPred: prediction of allergenic proteins and mapping of IgE epitopes, Nucleic Acids Research, vol.34, W202-W209, 2006.

[58] Salcedo, G., Sanchez-Monge, R., Diaz-Perales, A., Carcia-Casado, G., Barber, D., Plant non-specifi c lipid transfer proteins as food and pollen allergens, Clin. Exp. Allergy, vol.34, s.1336–1341, 2004.

[59] Scheurer, S., Son, D.Y., Boehm, M., Karamloo, F., Franke, S., Hoffman, A., Haustein, D., Vieths, S., Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen, Mol. Immunol., vol.36: s.155–167, 1999.

[60] Schölkopf, B. and Smola, A. J., Learning With Kernels: Support Vector Machines, Regularization and Beyond, The MIT Press, Cambridge, 2002.

[61] Silvanovich, A., Nemeth, M.A., Song, P., Herman, R., Tagliani, R., and Bannon, G.A., The value of short amino acid sequence matches for prediction of protein allergenicity, Toxicol. Sci., vol.90, s.252–258, 2006.

[62] Soeria-Atmadja, D., Zorzet, A., Gustafsson, M.G. and Hammerling, U., Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms, Int. Arch. Allergy Immunol., vol.133, s.101–112, 2004.

[63] Soltero, R. and Ekwuribe, N., The oral delivery of protein and peptide drugs, Innovat. Pharmaceut. Technol., vol.1, s.106–110, 2002.

[64] Stadler, M.B. and Stadler, B.M., Allergenicity prediction by protein sequence, FASEB J., vol.17, s.1141–1143, 2003.

[65] Taylor, S.L., Protein allergenicity assessment of foods produced through agricultural biotechnology., Annu. Rev. Pharmacol. Toxicol., vol.42, s.99-112, 2002.

[66] Van Regenmortel, M.H., Pellequer, J.L., Predicting antigenic determinants in proteins: Looking for unidimensional solutions to a threedimensional problem?, Pept. Res., vol.7, s.224–278, 1994.

[67] Vapnik, V.N., The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[68] Vapnik, V.N., Statistical Learning Theory, John Wiley & Sons, New York, 1998.

[69] Viswanathan, M., Kotagiri, R., Comparing the performance of support vector machines to regression with structural risk minimization, Proceedings of International Conference on Intelligent Sensing and Information Processing, s.445 -449, 2004.

[70] Wensing, M., Akkerdaas, J.H., van Leeuwen, W.A., Stapel, S.O., Bruijnzeel-Koomen, C.A., Aalberse, R.C., Bast, B.J., Knulst, A.C., van Ree, R., IgE to Bet v 1 and profiling: cross-reactivity patterns and clinical relevance, J. Allergy Clin. Immunol., vol.110, s.435–442, 2002.

[71] Weston, J. and Watkins, C., Support vector machines for multi-class pattern recognition, Proceedings of the Seventh European Symposium on Artificial Neural Networks, Bruges, April 21–23, 1999.

[72] Zhang, B., Is the maximal margin hyperplane special in a feature space?, Technical Report, HP Laboratories Palo Alto, 2001.

[73] Zorzet, A., Gustafsson, M. and Hammerling, U., Prediction of food protein allergenicity: a bioinformatic learning systems approach, In Silico Biol., vol.2, s.525–534, 2002.

ÖZGEÇMİŞ

Öykü Eren, 1981 yılında Ankara'da doğdu. İlköğretimini Yenimahalle Fatih İlkokulunda, ortaokul ve lise eğitimini Çankaya Milli Piyango Anadolu Lisesinde tamamladı. 2001 yılında kazandığı Başkent Üniversitesi Bilgisayar Mühendisliği bölümünden 2006 yılında mezun oldu. 2006 yılından bu yana, Başkent Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği bölümünde araştırma görevlisi olarak görev yapmaktadır. Araştırma konuları, bilgi geri getirim sistemleri, biyoenformatik ve veritabanı yönetim sistemleridir.