İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

MAKİNA ÖĞRENME YÖNTEMLERİYLE GENOM DİZİLİM VERİLERİNİN ANALİZİ

Bilg. Yük. Müh. Ergün GÜMÜŞ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman
Prof. Dr. Ahmet SERTBAŞ

Mayıs, 2013

İSTANBUL

İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

MAKİNA ÖĞRENME YÖNTEMLERİYLE GENOM DİZİLİM VERİLERİNİN ANALİZİ

Bilg. Yük. Müh. Ergün GÜMÜŞ
Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Programı

Danışman
Prof. Dr. Ahmet SERTBAŞ

Mayıs, 2013

İSTANBUL

2602080017 Öğrenci numaralı Ergün GÜMÜŞ tarafından hazırlanan bu çalışma 14/06/2013 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği programında Doktora Tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. Ahmet SERTBAŞ (Danışman)
İstanbul Üniversitesi
Mühendislik Fakültesi

Prof. Dr. A. Halim ZAİM
İstanbul Ticaret Üniversitesi
Mühendislik ve Tasarım Fakültesi

Prof. Dr. Nizamettin AYDIN
Yıldız Teknik Üniversitesi
Elektrik-Elektronik Fakültesi

Prof. Dr. Ahmet GÜL
İstanbul Üniversitesi
İstanbul Tıp Fakültesi

Doç. Dr. Olcay KURŞUN
İstanbul Üniversitesi
Mühendislik Fakültesi

ÖNSÖZ

Tez çalışmamda yardım ve tavsiyelerini benden esirgemeyen değerli hocalarım Prof. Dr. Ahmet SERTBAŞ ve Doç. Dr. Olcay KURŞUN'a, disiplinler arası bu çalışmada genetik alanındaki birikimini büyük bir sabırla benimle paylaşan değerli hocam Doç. Dr. Duran ÜSTEK'e ve çalışma arkadaşlarım Zeliha GÖRMEZ, Muhammed Erdem İSENKUL ve Ahmet DEVELİ'ye içtenlikle teşekkür ederim.

Doktora öğrenimim süresince sağladığı destekten ötürü Türkiye Bilimsel ve Teknolojik Araştırma Kurumu'na (TÜBİTAK) teşekkürü borç bilirim.

Mayıs, 2013 **Ergün GÜMÜŞ**

İÇİNDEKİLER

ŞEKİL LİSTESİ

TABLO LİSTESİ

SEMBOL LİSTESİ

λ_X, λ_Y	: KBA için Lagrange çarpanları
$sign(.)$	: İşaret fonksiyonu
p_{AB}	: AB alel çiftinin beraber görülme olasılığı
D' , r^2	: Bağlantı eşitsizliği ölçütleri
$\overline{X}$	: X veri kümesinin ortalaması
C	: Kovaryans matris
N	: Örnek sayısı
D	: Verideki boyut sayısı
$\|\|w\|\|$	: w vektörünün normu
λ_1, λ_2	: Birincil ve ikincil özdeğerler
$L(.)$	: Lagrange fonksiyonu
c^u	: DVM için u. örneğin sınıf bilgisi
a^u	: DVM için u. örneğe karşılık gelen Lagrange çarpanı
ξ^u	: DVM için u. örneğe karşılık gelen artık değişken
$\phi(.)$	: Daha yüksek boyutlu uzaya taşıma operatörü
$K(,)$	: Çekirdek fonksiyonu
γ	: DTF türü çekirdek için Gauss yüzeyinin genişlik katsayısı
μ	: Bir dağılımın ortalaması
σ	: Bir dağılımın standart sapması
r_{XY}	: X ve Y dağılımları arasındaki bağıntı
w_X , w_Y	: KBA için izdüşüm vektörleri
$H(Y)$	: Y dağılımındaki düzensizlik miktarı
$D(p \| q)$	: p ve q dağılımları arasındaki bağıl düzensizlik
α, β	: Alfabe
d_i^k	: k. dizilimdeki i. tür bazların/dimerlerin frekansı
L_k	: k. dizilimin uzunluğu
m_i^k	: k. dizilimdeki i. tür bazların dizilimdeki ortalama pozisyonu
v_i^k	: k. dizilimdeki i. tür bazların dizilimdeki pozisyonlarının varyansı
w	: Pencere genişliği
g	: Pencereler arası boşluk miktarı
p	: İstatistiksel anlamlılık

KISALTMA LİSTESİ

DVM : Destek Vektör Makineleri
KBA : Kanonik Bağıntı Analizi
TBA : Temel Bileşen Analizi
DTF : Dairesel Tabanlı Fonksiyon

ÖZET

MAKİNA ÖĞRENME YÖNTEMLERİYLE GENOM DİZİLİM VERİLERİNİN ANALİZİ

Geçtiğimiz yüzyılda biyoloji ve genetik alanında yaşanan ilerlemeler “Biyoinformatik” isimli yeni bir disiplinin oluşumuna ve insanoğlunun dünyadaki canlı çeşitliliğini, hastalıklara neden olan etmenleri ve çözümlerini daha iyi anlamasına yol açmıştır. Canlının yapı taşı olan genomu her yönüyle anlamayı hedefleyen genom araştırmalarının şüphesizki bu ilerlemede payı büyüktür. Bununla beraber her yeni dizileme sisteminde, genom dizileme maliyetlerinin düşmesiyle beraber genom araştırmalarının önemli bir çalışma alanı olan “kişiye özgü ilaç tasarımı” daha fazla uygulanabilir hale gelmiştir. Bu bağlamda makine öğrenme ve istatistiki analiz yöntemlerine bağlı genom analizi de önemli bir rol kazanmıştır.

Bu çalışma kapsamında, gen tedavisi araştırmalarında gen transferi amacıyla kullanılan ve HIV (İnsan Bağışıklık Yetmezliği Virüsü)’den türetilmiş lentivirüs vektörlerinin insan genomunda yerleşmeyi tercih ettiği bölgelerde gözlemlenen simetrik/palindromik davranışı yakalayan bir örüntü tarama aracı geliştirilmiştir. Örüntü tarama aracının, oluşturulan farklı test kümeleri üzerinde, çeşitli dizilim özellikleri ve değişken parametrelerle (pencere genişliği ve pencereler arası boşluk gibi) kullanılması sonucu probleme en uygun parametreler belirlenmiştir. Sonuçların anlamlılığı z-test ve Man-Whitney-Wilcoxon sıralama toplamı testi gibi istatistiki testlerle sınanmıştır.

Çalışmanın ikinci kısmında, söz konusu örüntü tarama aracında kullanılan Kanonik Bağıntı Analizi yöntemi, vaka ve kontrol gruplarındaki farklı “Bağlantı Eşitsizliği”ne sahip bölgelerin tespiti için kullanılmış ve böylece Behçet hastalığına neden olan aday mutasyonların genomdaki dağılımı incelenmiştir. Sonuçlar, söz konusu yöntemin hastalığa neden olan ve birbiriyle ilişkili mutasyonların tespiti için de kullanılabileceğini ortaya koymuştur.

Çalışmanın son aşamasında, milletlerin genetik çeşitlilikleri ve coğrafi dağılımları arasındaki ilişki incelenmiştir. Buna göre “İnsan Genomu Çeşitliliği Projesi” (Human Genome Diversity Project) kapsamında oluşturulmuş bir veri kümesinden faydalanılmış ve Temel Bileşen Analizi yöntemi yardımıyla insanların genetik çeşitliliğinin coğrafi çeşitlilikleri ile bağıntılı olduğu gösterilmiştir. Bunun yanı sıra, bu bağıntıyı ifade etme konusunda daha az sayıda tekli nükleotit çeşitliliğinin de yeterli olabildiği görülmüştür.

SUMMARY

ANALYSIS OF GENOME SEQUENCE DATA USING MACHINE LEARNING METHODS

Over the past century, the progress in biology and genetics fields has helped the birth of a new discipline called “Bioinformatics” and a better understanding of species variety, causes of diseases and along with their cures. Without a doubt, genome-wide studies which aim to understand genome with all of its aspects, have a major role in this progress. Nevermore, due to reduced sequencing costs by each sequencing system, “personalized medicine”, which is a core study field of genomic research, has become much more applicable. In this context, machine learning and genome analysis based on statistical methods have gained an important role.

Lentivectors derived from various types of viruses are used for gene transfer in gene therapy studies. In this study, a pattern search tool of which aim is to find symmetric/palindromic behavior observed in the integration regions of HIV (Human Immunodeficiency Virus) derived lentivirus vectors, has been developed. By using the pattern search tool on different test sets with different sequence features and parameters (like window width and gap between windows), optimal parameters specific to the problem have been determined. Significance of the results have been tested using statistical tests like z-test and Mann-Whitney-Wilcoxon ranksum test.

In second part of the study, Canonical Correlation Analysis method on which the developed pattern search tool depended, has been used to detect genomic regions with different “Linkage Equilibrium” values in case/control groups. By this way, distribution of candidate mutations causing to Behcet’s disease has been analyzed. Results proved that this methodology can be used to detect disease related and cross-correlated mutations.

In last part of the study, the relation between the genetic diversities and geographical locations of races has been studied. For this reason, the dataset which had been composed in context of Human Genome Diversity Project has been utilized and with the help of Principal Component Analysis method, a correlation (called as geo-genomic correlation) between the pairwise genetic distance and geographical distance of races has been found. Nevertheless, it is shown that much less number of Single Nucleotide Polymorphisms (SNP) are required to establish such correlation than using all SNPs.

1. GİRİŞ

Bilgisayarların hayatımıza girdiği günden bu yana hesaplama güçlerinin yanı sıra kullanım alanlarındaki çeşitlilik de artmıştır. Temeli ikinci dünya savaşı yıllarında şifrelenmiş mesajları çözmek için inşa edilen büyük boyutlu mekanik şifre çözücülere (Davies, 1999) dayanan bu cihazlar, zaman içerisinde gelişerek kendilerine yeni görev alanları bulmuşlardır. Endüstriyel ürün tasarımı, finans sektörü, iletişim, havacılık, uzay araştırmaları, tıbbi görüntüleme teknolojileri bu alanlardan sadece bir kaçıdır. Ancak bu alanların çoğunda bilgisayarlar, sadece ağır hesaplama yükünü üstlenmekte ve son karar yine insan operatöre kalmaktadır. Bu durum, danışman olarak bir insan bulundurmayan ve doğru kararların hızlıca alınması gereken sistemler için insanın görevini üstlenecek yapılara ihtiyacı doğurmuştur. Bu gereksinim, günümüzdeki "Yapay Zeka" ve buna bağlı olarak "Yapay Öğrenme" ya da diğer bir deyişle "Makine Öğrenmesi" isimli çalışma alanlarının doğmasına neden olmuştur.

İngiliz matematikçi Alan Turing, "Turing Testi" olarak isimlendirilen teorisini (Turing, 1950) öne sürerek bilgisayarlar için yüksek maliyetli hesaplama işlemlerini gerçekleştirmenin yanı sıra insan davranışlarını ve düşüncelerini taklit etmek gibi yeni bir görev daha tayin etmiştir. Bu görevi yerine getirebilen, çevresindeki değişen parametrelere göre kendi kararlarını üretebilen hesaplama sistemleri yapay bir zekaya sahip olarak addedilir. Günümüzde kullandığımız pek çok sistem, biz "bozulduklarını" farkedene kadar insan operatörün vermesi gereken kararları bizim yerimize vermeye devam etmektedir. Ortamın sıcaklığını ve nemini ölçerek iklimlendirme yapan havalandırma cihazı, herhangi bir istasyonu bozulan ya da tıkanan iletişim ağındaki sorunu algılayabilen ve trafiği mümkün olan en düşük maliyetle yeniden düzenleyebilen yönlendirme algoritması, viraja hızlı giren bir aracı yolda tutmak için tekerleklere değişen zaman dilimlerinde farklı frenleme etkisi dağıtan çekiş kontrol sistemi, bir nükleer santraldeki reaktörün kararlı çalışması için gerekli ölçümleri yapan ve buna göre önlemler alan karar sistemi, yapay zekanın en basitinden en kritiğine kadar her türlü alanda kullanılan örnekleridir.

Yapay zeka sistemlerinin davranışlarındaki çeşitliliği sağlayan en önemli etmen “eğitim” ya da diğer bir deyişle “öğrenme” adımıdır. Öğrenme adımı için farklı amaçlara yönelik pek çok makine öğrenme yöntemi bulunur.

Öğrenme çeşitlerine göre makine öğrenme yöntemleri, danışmalı ve danışmasız olarak ikiye ayrılır. Danışmalı öğrenmede, sistemin aldığı girdilere (örneklere) karşı üretmesi gereken çıktıları (sınıfları) bir uzman önceden kesin bir şekilde belirler. Bu işleme “sınıflandırma” adı verilir. Buna karşın danışmasız öğrenmede bu uzman etkisi bulunmadığından veri, doğasında bulunan özellikler kullanılarak farklı alt kümelere bölünmeye çalışılır. Bu işleme de “kümeleme” ya da “öbekleme” adı verilir.

Öğrenme amaçlarına göre ise; makine öğrenme yöntemleri, “Sınıflandırma Yöntemleri” ve “Bağlanım Yöntemleri” olarak ikiye ayrılır. Önceden de belirtildiği gibi sınıflandırma yöntemlerinde, sistemin eğitiminde kullanılan örneklerin ait oldukları sınıflar (kategoriler) önceden bellidir ve öğrenilen model kullanılarak yeni bir örneğin hangi sınıfa ait olduğu tespit edilir. Bağlanım yöntemlerinde ise birbirinden kesin sınırlarla ayrılmış sınıflar/kategoriler bulunmamaktadır. Bunun yerine sisteme verilen bir girdiye karşı üretilecek sayısal çıktının ne olacağı, önceki girdilere bağlı olarak tahmin edilmeye çalışılmaktadır. Bir metal çubuğun değişen sıcaklıklarda ne kadar esneyeceği, bir aracın katettiği kilometre miktarına göre ikinci el pazarındaki satış fiyatı veya bir firmanın yıllık kâr tahmini bağlanım modelleri ile açıklanabilmektedir.

Makine öğrenmesi alanında günümüze kadar, Temel Bileşen Analizi, Doğrusal Ayırtaç Analizi, Kanonik Bağıntı Analizi, K-En Yakın Komşuluk Sınıflandırma, K-Merkez Öbekleme, Öz-örgütlemeli Haritalar, Yapay Sinir Ağları, Karar Ağaçları, Destek Vektör Makineleri vb. gibi pek çok teknik ve bu tekniklerin türevleri öne sürülmüştür. Bu denli geniş bir yelpazeye sahip olması nedeniyle makine öğrenmesi, bir çok disiplinden çok çeşitli veri kümeleri üzerinde başarıyla kullanılabilir hale gelmiştir. Bu disiplinlerden taşıdığı hayati önem nedeniyle günümüzde en çok ilgi çekeni ve hâlâ en çok bilinmeyene sahip olanı şüphesiz Biyoinformatik’tir.

Biyoinformatik, biyoloji çalışmalarından elde edilen ham verinin bilgisayarlarla işlenebilir formata getirilmesi, analizi ve uygun formatta saklanması için gerekli tüm

yöntemleri bünyesinde toplayan disiplinler arası bir çalışma alanıdır. Bu alanın günümüzdeki popüler çalışma konuları: genom dizileme, genom kapsamında ilişki çıkarımı, dizilim verilerinin analizi, protein sınıflandırma, tür içi ve türler arası çeşitliliklerin tespiti vb. şeklinde özetlenebilir. Bu çalışma konularına bağlı olarak hesaplamalı bilim disiplinlerinden makine öğrenmesi başta olmak üzere sinyal ve görüntü işleme, istatistik ve algoritma tasarımı alanlarındaki yöntemlere sıklıkla başvurulmaktadır.

1970'li yıllara kadar, dünyadaki tüm canlıların yaşam şifresi olan DNA'nın (Deoksirübo Nükleik Asit) üç boyutlu yapısı keşfedilmiş, tür içi çeşitliliğin proteinler ve bu proteinleri kodlayan genler aracılığıyla sağlandığı öğrenilmiştir. Bu döneme kadar yapılan genomik çalışmalar, çoğunlukla bakteri gibi ilkel organizmalar üzerinde kısıtlı tekniklerle sürdürülmüştür. Ancak bu tarihten itibaren, bütün canlı türlerinin genomik atlasını çıkarabilecek tüm-genom dizileme yöntemlerinin/cihazlarının öne sürülmesi ve paralelinde gelişen hesaplamalı bilim yöntemleri Biyoinformatik çalışmalarına ivme kazandırmıştır.

İlk nesil genom dizileme tekniklerinin yüksek maliyeti, moleküler biyoloji konusunda yapılan çalışmalar sonucunda bulunan yeni tekniklerle düşürülmüştür. Böylece bir canlı türünden ziyade türün her bireyinin tüm-genom atlasını çok daha ucuza çıkartabilmek mümkün hale gelmiştir. Ancak ne yazık ki, kullanılan yeni teknikler beraberinde hatalı ve daha kısa uzunluklu okumalar gibi handikaplar getirmiş ve bu durum, okumaların derlenerek (genome assembly) sıralı bir genom oluşturulmasını zorlaştırmıştır. Bu sorun, geliştirilen parçalı eşleşme algoritmaları ve okuma hatası düzeltmek için kullanılan istatistik testleri vasıtasıyla giderilmeye çalışılmaktadır. Geliştirilen tüm bu algoritmalar ve testler "Genom Derleme" adı altında ham verinin işlenebilir formata dönüştürülmesini hedefleyen yeni bir Biyoinformatik çalışma konusunu doğurmuştur.

Tür içi çeşitliliğin tespiti amacıyla birey bazında yapılmaya başlanan genom dizileme işlemleri, kişiye özgü ilaç tedavisini hedefleyen ve Biyoinformatikin yeni bir çalışma alanı olan "Gen Terapisi" araştırmalarına da olanak sağlamıştır. Gen terapisi/tedavisi alanında, canlının fenotipinde (dış görünüm) kendisini gösteren genotip (genetik yapı) kaynaklı hastalık veya özelliklerin genomdaki hangi bozulmalardan/değişikliklerden

kaynaklandığını belirlemeye yönelik çalışmalar yapılmaktadır. Bu maksatla, hasta bireylerden alınan genetik örneklerle oluşturulan bir dizi kümesinin, sağlıklı bireylere ait örneklerle oluşturulan dizi kümesiyle kıyaslanarak hastalığa neden olabilecek nükleotit değişikliklerinin tespiti sağlanabilmektedir. Bu tür çalışmalar "Genom Kapsamında İlişki Çıkarımı Çalışmaları" olarak isimlendirilir ve biyoinformatik'in t-testi, Fisher testi, Ki-kare testi, Hardy-Weinberg kuralı gibi istatistiksel anlamlılık testlerine sıklıkla başvurduğu bir alanıdır.

Biyoinformatik'in ilgilendiği bir diğer konu da, vücudumuzun yapı taşı olan proteinlerin sınıflandırılmasıdır. Örneğin, doğada mevcut bulunan veya laboratuar ortamında üretilen proteinler kullanılarak bir ilaç tasarlanmak istenildiğinde hangi proteinin bu iş için uygun erime sıcaklığına sahip olduğunu tespit etmek için aday proteinlerin tek tek laboratuar ortamında teste sokulması gerekmektedir. Ancak bu işlem, maliyetli olmasının yanı sıra zaman kaybına da yol açmaktadır. Bu noktada makine öğrenme yöntemlerinden faydalanarak uygun adayın tahmini de mümkün olabilmektedir (Gorania, 2010).

Bu çalışma kapsamında, bir virüsten türetilerek gen terapisinde kullanılacak olan bir vektörün insan genomunda yerleşmeyi tercih ettiği bölgelerde görülen bir karakteristiği tespit eden bir örüntü tarama aracı tasarlanmıştır. Bu araç, iki farklı veri kümesinde değişik dizilim özellikleriyle (baz-dimer frekansları ve moment değişkenleri) beraber kullanılmış sonuçların anlamlılığı istatistik testleri ile sınanmıştır. Bunun yanı sıra bu aracın temelini oluşturan kanonik bağıntı analizi yöntemi, İstanbul Üniversitesi Deneysel Tıp Araştırma Enstitüsü'nce (DETAE) hazırlanan ve Türkiye'deki Behçet hastalarının ve bir grup sağlıklı bireyin genomik verisini içeren bir veri kümesi üzerinde kullanılmıştır. Bu sayede, Behçet hastalığına neden olan tekil nükleotit değişimlerinin yalnız başlarına ve bir arada bulunmalarının hastalığa etkisi incelenmiştir. Bir başka çalışma konusu olarak "İnsan Genomu Çeşitliliği Projesi"nden elde edilen veri kümesi kullanılarak toplumların birbirlerine olan genomik ve coğrafi mesafeleri arasındaki ilişkinin Temel Bileşen Analizi yöntemiyle nasıl daha iyi bir şekilde ifade edilebileceği araştırılmıştır.

Bu çalışmanın kalan kısmı şu şekilde düzenlenmiştir: İkinci bölümde; genom araştırmalarının tarihçesi ve genom dizileme yöntemlerinin gelişimi gibi konulara yer verilmiş, bunun yanı sıra gen terapisi, genom kapsamında ilişki çıkarımı ve jeo-genomik ilişki çıkarımı gibi Biyoinformatik konuları işlenmiştir. Üçüncü bölümde; çalışmada kullanılan makine öğrenme yöntemleri, dizilim eşleştirme algoritmaları, istatistik testleri ve veri kümelerine değinilmiştir. Tasarlanan örüntü tarama aracı, kanonik bağıntı analizi ile genom kapsamında ilişki çıkarımı ve jeo-genomik ilişkinin temel bileşen analizi ile incelenmesi çalışmalarının sonuçları dördüncü bölümde verilmiştir. Son olarak; tartışma ve sonuç bölümünde, elde edilen deney çıktıları yorumlanarak çalışma sonlandırılmıştır.

2. GENEL KISIMLAR

2.1 GENOM ARAŞTIRMALARININ TARİHÇESİ

Yirminci yüzyılın başlarına kadar, genler ve önemleri hakkında pek az şey bilinmekteydi. Avery ve çalışma arkadaşları (Avery, 1944), genetik bilginin yeni hücrelere "Deoksirübo Nükleik Asit" (DNA) aracılığıyla aktarıldığını kanıtlayana kadar genetik çeşitliliğin proteinler aracılığıyla sürdürüldüğü düşünülüyordu. Bundan yaklaşık on yıl sonra 1953'te Watson ve Crick'in DNA'nın üç boyutlu yapısını modellemesiyle beraber bilim insanları DNA iplikçiğinin eşlenmesi ve genetik materyal transferi gibi konularda yeni keşifler yapmaya, takip eden yıllarda da DNA sarmalını istedikleri bölgeden kırabilen enzimler geliştirmeye başladılar. Böylece gelecekteki "Hedefe Yönelik Gen Terapisi" çalışmaları için uygun zemin hazırlanmış oldu.

1970'li yıllara doğru, tür içi genetik çeşitliliğin ve türler arası benzerliklerin ortaya çıkartılması için canlılara ait tüm genom atlasının belirlenmesi ihtiyacı kendini hissettirmeye başlamıştır. Bu amaçla 1972 yılında Tüm Genom Dizilimini (Whole Genome Sequencing) belirlemeye yönelik ilk otomatize yöntem Frederick Sanger tarafından öne sürülmüştür. "Sanger Sequencing" (Sanger, 1977) olarak da bilinen bu yöntemde birbirine yapışık DNA sarmal çiftleri ısı yardımıyla birbirinden ayrılmakta ve "DNA polimeraz" adı verilen özel bir enzim yardımıyla seçilen bir iplikçiğin bir bölgesini tamamlayan yeni bir dizilim (sekans) elde edilmektedir. Burada dikkat edilmesi gereken nokta, üretilen dizilimlerin genomun aynı bölgesinden gelebileceği gibi farklı bölgelerinden de gelebileceğidir. Bu nedenle "okunan" (içerdiği nükleotit sırası tespit edilen) dizilimlerin, daha önceden yapısı bilinen bir referans genom dizilimi ile karşılaştırılarak "konsensus" dizisinin oluşturulması gerekmektedir. Bu noktada karşılaşılan üç temel sorun şunlardır:

- Okunan dizilimin konsensusta kullanılamayacak kadar kısa olması,
- Genomun her bölgesinde okuma yapmanın garanti edilememesi,
- Evrim süresince genomun farklı bölgelerinde kendini tekrarlayan dizilimlerin konsensus dizisini oluşturmada fayda sağlamaması.

1988 yılında Amerikan Ulusal Araştırma Konseyi tarafından hazırlanan “Mapping and Sequencing the Human Genome” adlı bir çalışma ile DNA’nın keşfinin 50. yıl dönümü olan 2003 yılına kadar tamamlanması ön görülen bir plan açıklandı (Baldi, 2002). Bu plana göre; 2001 yılı sonuna dek insan genomunun üçte birlik bir kısmının, 2003 yılının sonuna kadar da tüm insan genomunun çıkartılması hedef koşuldu. Ayrıca çıkartılan genom dizilerinin herkesin erişimine açık olması gerekmekteydi. Plan öngörüldüğü gibi 2003 yılında tamamlanmış ve tüm insan genomunun %90’lık bir bölümü (özellikle dizilemesi oldukça güç olan tekrar bölgeleri genoma dahil edilmemiştir) NCBI (National Center for Biotechnology Information) adlı kuruluşun web sitesinden tüm dünyadaki araştırmacıların erişimine açılmıştır.

Proje tamamlandığında, insana ait referans genomu çıkarmak için gereken toplam maliyetin iki milyar doları geçtiği görülmüştür. Gelecekteki kişiye özel ilaç tasarımı çalışmaları için bu maliyetin düşürülmesi gerektiği anlaşılmıştır. Bu gereksinimden ötürü son on yıl içerisinde dizileme teknolojileri bir dizi gelişim göstermiştir. Böylece Sanger dizileme yönteminden günümüze kadar kullanılan dizileme yöntemindeki farklılıklardan kaynaklanan üç adet genom dizileme nesli doğmuştur.

2.1.1 İlk Nesil Dizileme

Bu nesil bilinen ilk dizileme yöntemi olan Sanger yönteminden (Shotgun Sequencing) oluşmaktadır. Bu yöntemi kullanan ilk dizileme makinesi 1987 yılında ABi (Applied Biosystems) firması tarafından piyasaya sürülmüştür. Bu yöntem ile ortalama 500-1000bp (base-pair) dizilimler elde edilmekteydi. Elde edilen dizilimlerdeki hatalı okuma olasılığı yüzde birden azdı. Bunlar dizilemeyi takip eden derleme (assembly) aşamasını kolaylaştıran özelliklerdi. Ancak işlem sırasında kullanılan materyalin okunan her bin baz çifti için yaklaşık bir dolar gibi bir maliyeti vardı. İnsan genomu projesi 2003 yılında tamamlandığında 20 bireyden alınan kan örnekleri birden çok dizileme işlemine tabi tutulmuş ve toplam maliyet önceden de belirtildiği gibi iki milyar doları bulmuştu.

Sanger yöntemi, ırklar arası farklılıkların (özellikle SNP-Single Nucleotide Polymorphism) tespiti ve buna bağlı olarak kişiye özel ilaç tasarımı gibi konular için kullanılamayacak kadar pahalı olduğundan yeni dizileme yöntemlerine ihtiyaç duyuldu.

Böylece ikinci nesil teknikleri barındıran “Gelecek Nesil Dizileme” (Next Generation Sequencing) dönemine geçildi.

2.1.2 İkinci Nesil Dizileme

Sanger dizileme yönteminde maliyeti artıran en önemli unsur, kullanılacak olan DNA örneğinin bakteriyel klonlama tekniğiyle güçlendirilmesiydi (Shotgun sequencing). İkinci nesil dizileme yöntemlerinde bunun yerine PCR (Polymerase Chain Reaction) tekniği kullanılmış ve daha düşük maliyetle daha fazla dizilimin elde edilmesi hedeflenmiştir. Bu nesilde temel olarak üç ürün ön plana çıkmıştır:

- *Roche 454 (GS FLX Titanium Series)*
  Dünyanın en büyük ilaç üreticilerinden Roche firması 2005 yılında 454-GS FLX isimli ilk ikinci nesil dizileme cihazını üretti. Cihazın ilk sürümü, uzunlukları 200-300bp arasında değişen ortalama bir milyon adet dizilim üretmekteydi. Her bir çalıştırma turu yaklaşık on saat sürmekteydi.

- *Illumina (HiSeq 2000)*
  Illumina firmasının ürettiği HiSeq2000 cihazı, 100bp uzunluklu ortalama altı milyar adet dizilim üretmekteydi. Cihazın her bir çalışma turu on bir gün sürmekte ve insan genomu için ortalama çalışma maliyeti 8.000 doları bulmaktaydı.

- *Applied Biosystems (SOLiD)*
  Sanger dizileme yöntemiyle çalışan ilk makineyi üreten ABi firması ikinci nesil dizileme için SOLiD isimli bir ürünü piyasaya sürdü. Cihaz bir çalıştırmada, 50bp uzunluklu ortalama iki milyar adet dizilim üretebilmekteydi. Her çalıştırmanın maliyeti ortalama 8000 doları bulmaktaydı.

İkinci nesil dizileme ürünlerinin ortak özelliği doğruluktan ve uzun dizilim okumalarından taviz vermiş olmalarıdır. Ancak Sanger yönteminin yüksek maliyetiyle kıyaslandığında bu kabul edilebilir bir kusurdur.

2.1.3 Üçüncü Nesil Dizileme

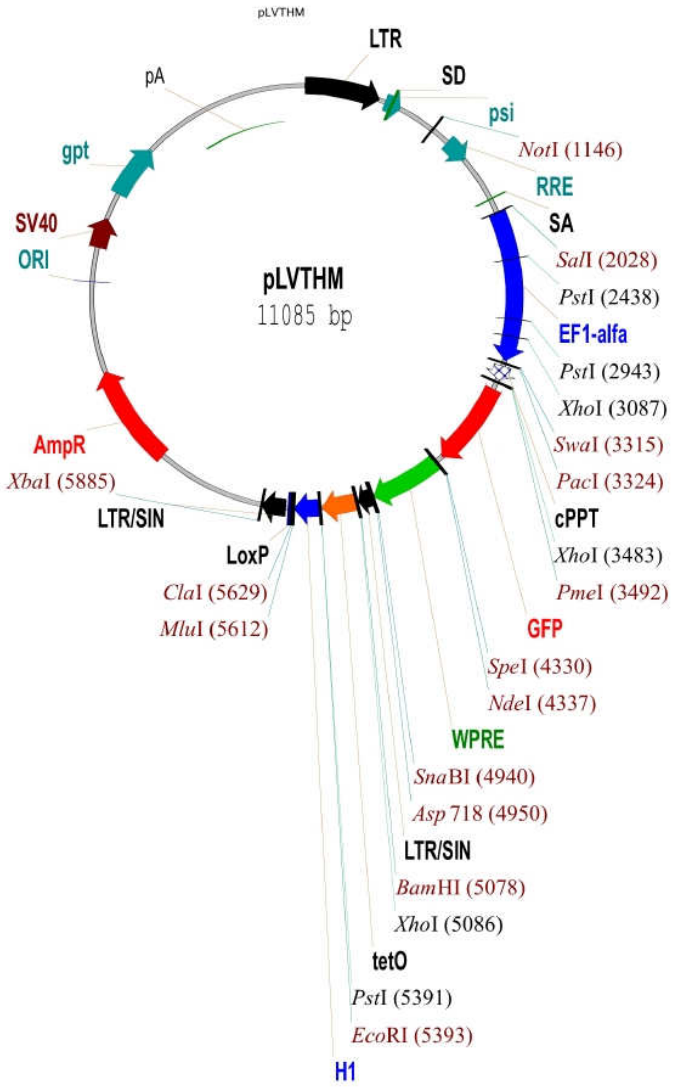
Bu nesilde öne sürülen ürünlerin temel özelliği, dizileme maliyetini düşürmek için "Tekil Molekül Dizileme" (Single molecule sequencing) tekniğinin kullanılmasıdır. Ticari olarak Helicos firmasının "Heliscope Sequencer" isimli ürünü ön plana çıkmıştır. Bu cihazın bir çalışma turu altı gün sürmekte ve ortalama 700 milyon adet 35-50bp uzunluklu dizilimler üretmektedir. Bu teknoloji ile genom dizileme maliyeti 2000 dolara kadar düşmüş ancak okunan dizilimlerdeki hata oranı da ikinci nesil ürünlere göre artmıştır. Ayrıca elde edilen dizilimlerin çok kısa olması bir sonraki aşama olan derleme aşamasını güçleştirmiştir.

2.2 GEN TEDAVİSİ ÇALIŞMALARI

Gen tedavisi, geçmişi 1990'lı yıllara dayanan, kalıtsal ya da sonradan ortaya çıkan hastalıkların tedavisi maksadıyla gen, DNA ve RNA moleküllerinin insan hücrelerine, organ ve dokularına transfer işlemini içeren bir tedavi şeklidir (Şanlıoğlu, 2010). Tedavide kullanılacak hedef genin ürününe sürekli ihtiyaç duyuluyorsa, genin kromozomal yerleşimi için virüs vektörlerinin kullanılması avantajlı gibi gözükmektedir. HIV'den türetilmiş lentivirüs vektörleri genoma eklenmeleri ve bağışıklık sisteminden kaçmaları nedeniyle tercih edilen sistemlerdir. Yöntem temel olarak, yoksunluğu hastalığa neden olan proteinleri kodlayacak RNA (Ribo Nükleik Asit) dizilimlerini özel olarak tasarlanmış retrovirüslerin içerisine yerleştirmek ve bu retrovirüsleri hastaya transfer etmeye dayalıdır. Transfer işlemi sonucunda retrovirüs, özel bir enzim grubu kullanarak önce içeriğindeki RNA'yı DNA'ya çevirir sonra da üretilen DNA'yı konak hücrenin genomuna ekler. Konak genoma bağlanma işlemi bu virüslerin baş ve son ucunda bulunan "Long Terminal Repeat" (LTR) isimli bölümler aracılığıyla sağlanmaktadır. Şekil 2.1'de (Addgene, 2012) gen tedavisi yönteminde kullanılan örnek bir retrovirüs (vektör) görülmektedir.

Bu vektör aynı zamanda İstanbul Üniversitesi Deneysel Tıp Araştırmaları Enstitüsü (DETAE) Genetik Anabilim dalındaki bir çalışmada kullanılmıştır. Bu çalışmada, insan konağına ait 293T hücrelerinden alınan örneklere, tasarlanmış retrovirüs yerleştirilmiş (transfection) ve çoğaltılmıştır (transduction) (Ustek, 2012). Böylelikle dizileme işlemi sonucunda, hücrelere yerleştirilmiş bu retrovirüsün izlerini (yerleşim için tercih ettiği

noktaları) bulmak kolaylaşmıştır. Dizileme işlemi, Bölüm 2.1.2'de bahsedilen GS FLX cihazıyla gerçekleştirilmiş, yine Bölüm 3.6.2'de bahsedilen Smith-Waterman hizalama tekniği ile içinde bu retrovirüse ait parçaların bulunduğu okumalar tespit edilmiştir. Sonrasında, arda kalan okumalar Bölüm 3.6.3'de bahsedilen BLAST tekniği ile insan referans genomu ile karşılaştırılmış ve retrovirüsün tercih ettiği toplam 76 farklı bölge bulunmuştur. Bulunan bölgeler, retrovirüsün yerleştiği başlangıç noktasına göre hizalanmış ve bu çalışma kapsamında geliştirilmiş bir örüntü tarama aracı vasıtasıyla bu dizilimde okumaların belirli bir bölgesinde nükleotit-dimer frekansları gibi dizilim özellikleri açısından benzer bir bölge olup olmadığı incelenmiştir. (Bkz: Bölüm 4.2).



Şekil 2.1: Bir retrovirüs örneği

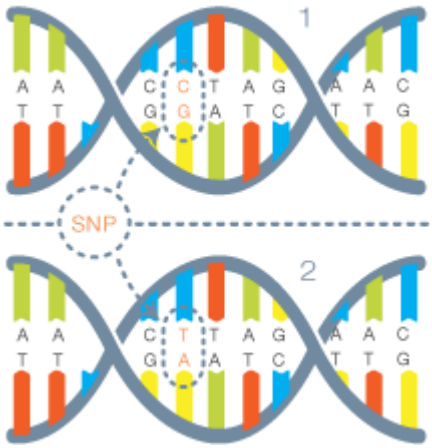
Benzer bir çalışmada (Schroder, 2002), HIV'in insan genomunda yerleşmeyi tercih ettiği bölgeleri tespit etmek amacıyla insan konağına ait bir T hücresi, HIV'den türetilmiş bir vektörle enfekte edilmiş ve çoğunluğu gen bölgesi olan toplam 524 alanda söz konusu vektörün yerleşimi tespit edilmiştir. Bu çalışmada elde edilen kimerik okumalar incelenip insan genomundaki yerleşim noktaları alt alta getirilip hizalandığında, yerleşim noktasını çevreleyen küçük bir alanda nükleotitlerin görülme olasılığına göre simetrik bir değişim olduğu gözlemlenmiştir (Holman, 2005 – Wu, 2005). Bu simetrik/palindromik özelliğin tanımı Bölüm 3.7'de verilmiş ve söz konusu özellik yine bu tez çalışması kapsamında hazırlanan örüntü tarama aracıyla incelenmiş, sonuçların istatistiksel anlamlılığı z testi ile sınanmıştır (Bkz: Bölüm 4.1).

2.3 GENOM KAPSAMINDA İLİŞKİ ÇIKARIMI

Doğuştan gelen fizyolojik özelliklerimizin (Fenotip) tamamı genetik yapımız (Genotip) sayesinde şekillenmektedir. Bu şekillenme kendisini, boy ya da göz rengi olarak gösterebileceği gibi bazı kanser türleri, diyabet, yüksek tansiyon vb. hastalıklar şeklinde de gösterebilir. Bu sonuçlara bağlı parametrelerin (örneğin diyabet hastalarındaki insülin miktarı), genomdaki SNP (Single Nucleotide Polymorphism - Tekil Nükleotit Çeşitliliği) adı verilen nükleotit değişiklikleriyle (varyasyon) ilişkisini çözmek için yapılan çalışmaların tamamına "Genom Kapsamında İlişki Çıkarımı Çalışmaları" (Genome Wide Association Studies- GWAS) adı verilir.

Tekil Nükleotit Çeşitliliği:

Aynı tür içindeki bireylerin DNA dizilimleri kıyaslandığında aynı nükleotit pozisyonunda (locus) farklı bazların bulunması durumuna denir. Söz konusu nükleotit pozisyonunda bulunan bu farklı bazların her biri "Alel" olarak isimlendirilir. Şekil 2.2'de (Sirius Genomics, 2011), görülen SNP konumunda C ve T alelleri (tamamlayıcı iplikçik için G ve A alelleri) görülmektedir. Bir gen bölgesinde (kodlayan bölge) bulunan SNP, o genin davranışını bozup bir düzensizliğe neden olabilir.



Şekil 2.2: Aynı türden iki birey arasında tespit edilen bir SNP

Genom kapsamında ilişki çıkarımı çalışmaları, genellikle aynı tür içinde ortaya çıkan hastalığın kaynağı olan SNP'leri belirleyebilmek için yapılmaktadır. Bu kapsamda söz konusu hastalıktan etkilenen hasta ve etkilenmeyen sağlıklı grupların genom dizilimleri karşılaştırılır. Bu karşılaştırma sırasında SNP bölgesindeki alel frekanslarının hasta ve sağlıklılara göre dağılımı Ki-kare testi veya Hardy-Weinberg kuralı (Weinberg, 1908) gibi temel istatistik testleriyle incelenir. Eğer incelenen SNP'lerden hasta grubunda daha sık görülen bir tanesi belirlenirse bu SNP'nin hastalıkla ilişkili olduğu kabul edilir. Ancak bir hastalığı her zaman bir SNP ile açıklamak mümkün olmayabilir. Bu gibi durumlarda genomun farklı bölgelerindeki SNP gruplarının hastalıkla ilişkisi araştırılır. Söz konusu bu ilişkilerin belirlenmesi için Bölüm 3.3'de değinilen "Kanonik Bağıntı Analizi" (KBA) yönteminden faydalanılabilmektedir.

2.3.1 KANONİK BAĞINTI ANALİZİ İLE GENOM KAPSAMINDA İLİŞKİ ÇIKARIMI

KBA yöntemi, incelenen görüler için ayrı ayrı oluşturulan kovaryans matrislerin tekil olması (tersinin alınamaması) durumunda özproblem çözülemediğinden çalışmamaktadır. Bu durumda kanonik ağırlık katsayıları (v ve u ağırlık vektörleri) "Çoklu Azaltım" (Multiple Regression) yöntemiyle hesaplanabilir. Bu amaçla Waaijenborg ve çalışma arkadaşları (Waaijenborg, 2008), "Cezalı KBA" (Penalized CCA) yöntemini öne sürmüşlerdir. Algoritma elastik ağ yöntemindeki cezalandırma mekanizmasını kullanarak gen ifadesi-kopya sayısı gruplarını modelin içinde beraber kalmaya ya da modelden beraber çıkmaya (sıfır katsayısı ile ağırlıklandırılma)

zorlamaktadır. v ve u ağırlık vektörleri k indisiyle kontrol edilen ve aşağıda görülen bir en iyileme döngüsüyle elde edilmektedir. n adet gözlem için X, $n \times q$; Y de $n \times p$ boyutlu veri kümeleridir. λ_X , X kümesi için ; λ_Y ise Y kümesi için azaltımda kullanılan ceza katsayılarıdır.

1. $k \leftarrow 0$
2. İlk ağırlık vektörleri olan $\hat{v}^{(0)}$ ve $\hat{u}^{(0)}$ için başlangıç değerleri ata. Örneğin $\hat{v}^{(0)} \leftarrow \frac{1}{q}$, $\hat{u}^{(0)} \leftarrow \frac{1}{p}$ gibi. $\hat{v}^{(0)}$ ve $\hat{u}^{(0)}$'yi normalize et.
3. ξ ve ω hesaplanan çıktılar olmak üzere, $\hat{\xi}^k - \hat{\xi}^{k-1} > 10^{-3}$ veya $\hat{\omega}^k - \hat{\omega}^{k-1} > 10^{-3}$ olduğu sürece (a-d) tekrarla,

   (a) $k \leftarrow k + 1$

   (b) $\hat{\xi}^k \leftarrow X\hat{v}^{(k-1)}$, $\hat{\omega}^k \leftarrow Y\hat{u}^{(k-1)}$

   (c) $\hat{u}_i^{(k)} = \left(\left| \hat{\xi}^{k\,T} Y_i \right| - \frac{\lambda_Y}{2} \right)_+ sign\left(\hat{\xi}^{k\,T} Y_i \right)$, $i = 1, 2, \ldots, p$

   $\hat{v}_i^{(k)} = \left(\left| \hat{\omega}^{k\,T} X_i \right| - \frac{\lambda_X}{2} \right)_+ sign\left(\hat{\omega}^{k\,T} X_i \right)$, $i = 1, 2, \ldots, q$

   Eğer $f > 0$, $f_+ = f$ aksi takdirde $f_+ = 0$

   (d) $\hat{v}^{(k)}$ ve $\hat{u}^{(k)}$'yı normalize et.

Ağırlık katsayılarının iteratif şekilde değiştirilmesiyle hem kovaryans matrislerin tersinin alınamaması sonucu doğan çözümsüzlük engellenmekte hem de katsayısı sıfır olarak atanan değişken grupları ilişki kümesinden çıkartılabilmektedir (Değişken azaltımı). Waaijenborg ve çalışma arkadaşları takip eden çalışmalarında da (Waaijenborg, 2009), (Waaijenborg, 2010) yöntemin değişik veri kümeleri üzerine uygulanabilirliğini göstermişlerdir.

Parkhomenko ve çalışma arkadaşları (Parkhomenko, 2009) da benzer bir şekilde, çok değişkenli iki veri kümesi arasındaki grup bağıntısını en yüksek kılmak için değişken azaltımı yöntemine başvurmuşlardır. Bu şekilde değişken gruplarından grup bağıntıları

en yüksek olacak alt kümelerin seçimi işlemine “Ayrık KBA” (Sparse CCA) adı verilmektedir. Parkhomenko ve arkadaşları çalışmalarında, *X* ve *Y* görüleri için uygun ağırlık vektörlerini hesaplarken $K = C_{XY}$ (Bkz: Denklem 3.40) matrisinin “Tekil Değer Ayrışımı” (Singular Value Decomposition) yöntemiyle tekil vektörlerini bulmuş ve bunları ağırlık vektörleri olarak kullanmışlardır. Önerdikleri algoritma şu şekildedir:

1. λ_u ve λ_v ayrıklık parametreleri için 0 ile 0.2 arasında başlangıç değerleri seç.
2. $i \leftarrow 0$. *K* matrisinin sol ve sağ tekil vektörleri için başlangıç değerleri (u^0, v^0) ata. Örneğin $v^0 \leftarrow \frac{1}{q}$, $u^0 \leftarrow \frac{1}{p}$ gibi.
3. *u* vektörünü güncelle.
   (a) $u^{i+1} \leftarrow Kv^i$
   (b) $u^{i+1} \leftarrow \frac{u^{i+1}}{\left\| u^{i+1} \right\|}$ (Normalizasyon)
   (c) $u_j^{i+1} \leftarrow \left(\left| u_j^{i+1} \right| - \frac{\lambda_u}{2} \right)_+ sign\left(u_j^{i+1} \right),\ j = 1, 2, \ldots, p$
   Eğer $f > 0$, $f_+ = f$ aksi takdirde $f_+ = 0$
   (d) $u^{i+1} \leftarrow \frac{u^{i+1}}{\left\| u^{i+1} \right\|}$ (Tekrar normalizasyon)
4. *v* vektörünü güncelle.
   (a) $v^{i+1} \leftarrow K^T u^{i+1}$
   (b) $v^{i+1} \leftarrow \frac{v^{i+1}}{\left\| v^{i+1} \right\|}$ (Normalizasyon)
   (c) $v_j^{i+1} \leftarrow \left(\left| v_j^{i+1} \right| - \frac{\lambda_v}{2} \right)_+ sign\left(v_j^{i+1} \right),\ j = 1, 2, \ldots, q$
   Eğer $f > 0$, $f_+ = f$ aksi takdirde $f_+ = 0$
   (d) $v^{i+1} \leftarrow \frac{v^{i+1}}{\left\| v^{i+1} \right\|}$ (Tekrar normalizasyon)
5. $i \leftarrow i + 1$

6. 3 ve 4 nolu adımları yakınsama sağlanana dek ($\left\|v^{i+1}-v^{i}\right\|$ ve $\left\|u^{i+1}-u^{i}\right\|$ seçilen eşik değerinden küçük olana dek) gerçekleştir.

Söz konusu algoritma aynı eğitim-test kümesi grubu için farklı λ_u ve λ_v ayrıklık parametreleri ile denenmekte ve görüler arasındaki bağıntıyı en yüksek kılacak parametreler tespit edilmektedir.

Witten ve çalışma arkadaşları (Witten, 2009), Ayrık KBA yöntemine danışmalı (supervised) öğrenme tekniğini uygulamış ve aynı gözlemin ikiden fazla görüsü için çalışan bir mekanizma geliştirmişlerdir. $\|w_1\|^2 \le 1$, $\|w_2\|^2 \le 1$, $w_{1j}=0 \ \forall j \notin Q_1$ ve $w_{2j}=0 \ \forall j \notin Q_2$ kriterlerine göre $w_1^T X_1^T X_2 w_2$ ifadesini en büyük kılacak w_1 ve w_2 ağırlık vektörlerini elde etmeye çalışmışlardır. *y* her örneğe karşılık gelen sınıf bilgisi olmak üzere; Q_1, X_1 görüsündeki *y* ile bağımlı değişkenler kümesini ve Q_2 de X_2 görüsündeki *y* ile bağımlı değişkenler kümesini göstermektedir. Böylece ayrık ağırlıklandırma sırasında her değişkenin sınıf etiketiyle olan ikili bağıntısı da incelenmektedir.

KBA yönteminin diğer genom kapsamında ilişki çıkarımı çalışmalarındaki kullanımı kısaca şu şekilde özetlenebilir:

- Boutte ve çalışma arkadaşları (Boutte, 2010), SNP gibi genetik işaretçilerin beyin fonksiyonları ile ilişkisini incelemek için fonksiyonel manyetik rezonans görüntüleme (fMRI) sistemini kullanmışlardır. Yüksek aktivite gösteren bölge çıktıları ile simüle edilmiş genetik işaretçiler arasındaki ilişki KBA yöntemi ile incelenmiştir.
- Naylor ve çalışma arkadaşları (Naylor, 2010), kromozom 2'deki üç genin gen ifade seviyeleri ile seçtikleri yirmi SNP arasındaki ilişkiyi KBA yöntemi ile gözlemlemiş, yöntemin hesaplama maliyeti açısından çiftler arası regresyon modelinden daha iyi olduğunu göstermişlerdir.
- Peng ve çalışma arkadaşları (Peng, 2010), hasta ve kontrol grupları arasında gen-gen ikili ilişkilerini değerlendirmek için KBA tabanlı bir istatistik ölçütü geliştirmişlerdir. Ölçütün bazı veri kümeleri için iyi çalıştığı gözlemlenirken

değişkenler arasındaki doğrusal olmayan ilişkilerin tespiti için yetersiz kaldığı belirtilmiştir.

2.3.1.1 ÖRNEK ÇALIŞMA: BEHÇET ARAŞTIRMASI

Behçet hastalığı, 1939 yılında Prof. Dr. Hulusi Behçet tarafından tanımlanan bir damar iltihabı hastalığıdır. Hastalık kendisini, tekrarlayan ağız yaraları (aft), genital bölge yaraları, deri, göz, eklem, damar tutulumu gibi belirtilerle (Pediatric Rheumatology, 2003) göstermektedir. Hastalık süresince her hastada aynı semptomlar gözlenmediğinden ortak bir tedavi yöntemi yoktur. Hastalığın nedeni kesin olarak bilinmemekle beraber yapılan genom kapsamında ilişki çıkarımı çalışmalarında (Mizuki, 2001), (Remmers, 2010) aileden gelen genetik yatkınlığın rolü ortaya konmuştur.

İstanbul Üniversitesi Deneysel Tıp Araştırmaları Enstitüsü'nce yapılan bir çalışmada (Remmers, 2010), Türkiye'deki 1215 Behçet hastası ve 1278 sağlıklı bireyden alınan örneklerle genom haritaları çıkartılmış ve 311459 adet SNP (Single Nucleotide Polymorphism) üzerinden genom kapsamında ilişki çıkarımı çalışması (GWAS) gerçekleştirilmiştir. SNP bazında yapılan ki-kare anlamlılık testleri ile 6 numaralı kromozomda bulunan MHC (Major Histocompatibility Complex) bölgesinin hastalıkla ilişkili SNP'ler bulundurduğu gözlemlenmiştir. Bu bölgede tespit edilen SNP'ler arasındaki "Bağlantı Eşitsizliği"nin (Linkage Disequilibrium) hasta ve sağlıklı gruplar arasında farklı değerlere sahip olduğu gözlemlenmiştir. Bu noktadan yola çıkarak hastalığa neden olan bölgelerin tespiti Bölüm 4.3'de değinildiği gibi sağlanabilir.

Bağlantı Eşitsizliği (Linkage Disequilibrium):

Bağlantı eşitsizliği, genom üzerinde birbirleriyle ilişkili iki SNP bölgesindeki (locus) alel frekansının birbirlerinden bağımsız olmaması durumudur. Ortak mutasyon, küçük gen havuzu, doğal seçilimle oluşan "süper genler" vb. söz konusu bu iki bölgedeki ilişkinin nedenleri arasında sayılabilir. Bağlantı eşitsizliğini ve ölçümünde kullanılan metrikleri daha iyi anlamak için Tablo 2.1 incelenmelidir.

Tablo 2.1: X ve Y bölgelerindeki alellerin görülme olasılıkları

		Y Bölgesi		
		B	**b**	
X Bölgesi	**A**	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
		p_B	p_b	

X ve Y isimli bölgelerde görülebilen alel türleri sırasıyla A, a, B, b şeklindedir. A ve B'nin baskın, a ve b'nin de çekinik aleller olduğunu kabul edelim. p_A ve p_a ilgili alellerin X bölgesinde, p_B ve p_b de ilgili alellerin Y bölgesindeki görülme olasılıkları olsun. İki bölgedeki alellerin birleşik olasılıkları p_{AB}, p_{Ab}, p_{aB} ve p_{ab} şeklindedir (Models in Human Genetics, 2006).

Buna göre X ve Y bölgelerinin birbirlerinden tamamen bağımsız olmaları durumunda (Bağlantı Eşitliği),

$$p_{AB} = p_A p_B \qquad (2.1)$$

$$p_{Ab} = p_A p_b = p_A(1 - p_B) \qquad (2.2)$$

$$p_{aB} = p_a p_B = (1 - p_A) p_B \qquad (2.3)$$

$$p_{ab} = p_a p_b = (1 - p_A)(1 - p_B) \qquad (2.4)$$

eşitliklerinin sağlanması beklenir. Ancak, eğer bu iki bölge birbiriyle ilişkili ise (Bağlantı Eşitsizliği) söz konusu eşitlikler sağlanmaz. Bunun yerine,

$$D_{AB} = p_{AB} - p_A p_B \qquad (2.5)$$

$$p_{AB} = p_A p_B + D_{AB} \qquad (2.6)$$

$$p_{Ab} = p_A p_b - D_{AB} \qquad (2.7)$$

$$p_{aB} = p_a p_B - D_{AB} \qquad (2.8)$$

$$p_{ab} = p_a p_b + D_{AB} \qquad (2.9)$$

durumu söz konusudur. Burada görülen D = D_{AB} değeri bağlantı eşitsizliğini ölçmekte kullanılan ilk metriktir. Ancak bu metriğin sabit bir aralığı olmadığından bunun yerine genellikle Denklem 2.10'da görülen D' metriği kullanılır.

$$D' = D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\min(p_A p_B, p_a p_b)} & eger \quad D_{AB} < 0 \quad ise \\ \dfrac{D_{AB}}{\min(p_A p_b, p_a p_B)} & eger \quad D_{AB} > 0 \quad ise \end{cases} \qquad (2.10)$$

D' metriği [-1 : +1] aralığından değerler alabilir. Değer sıfıra yaklaştıkça X ve Y bölgelerinin birbirinden bağımsız olduğu sonucuna varılır. Değerin pozitif olması bağımlılığın homozigot alellerce, negatif olması ise heterozigot alellerce sağlandığını göstermektedir.

Bir diğer metrik olan r^2 metriği ise,

$$r^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)} \qquad (2.11)$$

şeklindedir. Söz konusu bu metrik [0 : +1] aralığından değerler alabilir. Değer sıfıra yaklaştıkça X ve Y bölgelerinin birbirinden bağımsız olduğu sonucuna varılır. +1 değeri ise tam bağlantı eşitsizliğine işarettir.

2.3.2 TEMEL BİLEŞEN ANALİZİ İLE JEO-GENOMİK İLİŞKİ ÇIKARIMI

Çoğu toplum, gelişimi boyunca ilk coğrafi konumlarını terk ederek dünyanın farklı bölgelerine göç etmiştir. Göç yolları boyunca gerek maruz kaldıkları coğrafi etkenler (özellikle iklim koşulları) gerekse de karşılaştıkları toplumlarla kurdukları kan bağları gen havuzlarını şekillendirmiş ve genetik açıdan ilk durumlarına göre daha farklı bir hale gelmişlerdir (Kittles, 2003). Bu farklılıklar kendini "tekil nükleotit çeşitlilikleri" (SNP) ile göstermektedir. Bu çalışmanın bir ayağında farklı toplumlardaki bireylerin birbirlerine olan ortalama genomik mesafesi (genomlarındaki birbirinden farklı SNP'lerin toplam sayısı) ile coğrafi mesafeleri arasındaki (jeo-genomik) ilişki Temel Bileşen Analizi (TBA) yöntemiyle incelenmiştir (Bkz: Bölüm 4.4). TBA yöntemi,

yüksek boyutlu bir uzaydaki dağılımın daha düşük boyutlu bir uzaya taşınması ve buna bağlı olarak sınıfları ayırt edici özelliklerin çıkarımı için sıklıkla kullanılan bir makine öğrenme yöntemidir. Yöntem, veri uzayındaki örneklerin dağılımını mümkün olduğunca geniş bir alana yayan (örnek kümesindeki varyansı en büyük kılan) en uygun izdüşüm vektörlerini bulmayı amaçlamaktadır. Bu özelliği sayesinde sınıflandırmanın yanı sıra veri dağılımının çoğunlukla iki ya da üç boyutta görsel olarak incelenmesine de olanak sağlamaktadır (Bkz: Bölüm 3.1).

Yapılan incelemelerde, “İnsan Genomu Çeşitliliği Projesi” (Human Genome Diversity Project - HGDP) kapsamında oluşturulan veri kümesi kullanılmıştır. Bu proje, Stanford üniversitesi genetik araştırmacılarından Luigi Luca Cavalli-Sforza tarafından yürütülmüş ve dünya çapında birçok genetik araştırmacı tarafından da desteklenmiştir. Projenin temel amacı, farklı kökenlere sahip insan grupları arasındaki evrim kaynaklı çeşitliliği belirlemek ve toplum genetiğindeki değişiklikleri takip etmektir. Projeye başlangıçta, “bilimsel ırkçılık” ve “bio-korsanlık” gibi çekinceler nedeniyle soğuk yaklaşılmıştır ancak zaman içerisinde yapılan konferanslarda projenin genetik kökenli hastalıklar için ilaç yapımına olumlu katkısı öne sürülmüş ve proje çalışanları gerekli etik izinleri almışlardır. Proje sonunda elde edilen genetik materyal, Paris’deki “Jean Dausset” vakfına bağlı CEPH (Center for Study of Human Polymorphism) laboratuvarlarında toplanmış ve tüm dünyadan kâr amacı gütmeyen araştırmacıların kullanımına açılmıştır (Sforza, 2005). HGDP veri kümesinin detayları ve ön işleme basamakları Bölüm 3.8’de detaylı olarak verilmiştir.

3. MALZEME VE YÖNTEM

Bu başlık altında veri kümelerinin çıkarımında, analizinde ve elde edilen sonuçların istatistiksel anlamlılığının ölçümünde kullanılan yöntemlerden bahsedilmiştir.

3.1 TEMEL BİLEŞEN ANALİZİ

Temel Bileşen Analizi (Principal Component Analysis) yöntemi, yüksek boyutlu veri kümelerinde boyut azaltımı, yeni ve sınıflandırmada daha etkili özelliklerin çıkartılması, çok boyutlu verinin iki ya da üç boyutla ifade edilerek görüntülenebilmesi (Novembre, 2008) gibi değişik amaçlarla kullanılmaktadır. Bu yöntem aynı zamanda "Karhunen-Loeve Dönüşümü" olarak da bilinir.

Yöntemin temel amacı *S* uzayında bulunan *D* boyutlu bir veri kümesindeki örnekleri yine *D* boyutlu yeni bir *S'* uzayına aktarmaktır. Bu aktarma işlemi sırasında *S'* uzayını geren ve birbirine dik olan her baz vektörü, verideki varyansı en büyük kılacak şekilde seçilir. *S* uzayındaki veri noktalarının yeni *S'* uzayındaki baz vektörlere izdüşümünün alınmasıyla aktarma işlemi tamamlanmış olur. Bu işlem temelde bir özdeğer-özvektör probleminin çözümüne dayanmaktadır.

Problemin Tanımı:

S uzayında, *N* adet örnek içeren *D* boyutlu bir veri kümemiz ($X = [x_1, x_2, ..., x_N]$) olduğunu ve bu örneklerin hepsinin yine *D* boyutlu bir *S'* uzayına izdüşümünün alınacağını kabul edelim. *S'* uzayını geren ilk baz vektörü v_1 olsun. Mevcut tüm örneklerin bu baz vektörü üzerine izdüşümü alındığında yeni örnek noktalarını içeren $v_1^T X$ elde edilir. *X* kümesindeki örneklerin ortalaması olan $\overline{X}$ değerinin de (Denklem 3.1) aynı v_1 vektörü üzerine izdüşümü alındığında $v_1^T \overline{X}$ elde edilir.

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (3.1)$$

Örneklerin S' uzayında alınan izdüşümleri ile örneklerin ortalamasının S' uzayında alınan izdüşümü arasındaki fark (Denklem 3.2) toplam varyans miktarını verir.

$$\frac{1}{N}\sum_{i=1}^{N}\left\{ v_1^T x_i - v_1^T \overline{X} \right\}^2 = v_1^T C v_1 \qquad (3.2)$$

Burada C, $D \times D$ boyutlu kovaryans matris olup

$$C = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{X} \right)\left(x_i - \overline{X} \right)^T = \frac{1}{N}\sum_{i=1}^{N} y_i y_i^T \qquad (3.3)$$

şeklinde tanımlıdır. İzdüşüm işlemi sonunda elde edilen toplam varyansı en büyük kılmak için en uygun v_1 baz vektörünü bulmak gerekir. Bu da bir optimizasyon problemi olup, toplam varyansı gösteren $v_1^T C v_1$ ifadesinin v_1'e göre türevinin alınmasıyla sağlanabilir (Bishop, 2006). Burada dikkat edilmesi gereken nokta v_1 baz vektörünün uzunluğunun (normunun) 1 olması gerektiğidir. v_1'in orijine göre uzaklığı (uzunluğu),

$$L = \sum_{i=1}^{D}\left(v_1^{\,i} - 0 \right)^2 = v_1^T v_1 = 1 \qquad (3.4)$$

şeklinde bulunur. En iyileme problemine bu kısıt λ_1 gibi bir Lagrange çarpanı kullanarak eklendiğinde,

$$v_1^T C v_1 + \lambda_1\left(1 - v_1^T v_1 \right) \qquad (3.5)$$

ifadesinin v_1'e göre türevinin sıfıra eşitlenmesi gerektiği görülür. Bu işlem sonunda,

$$C v_1 = \lambda_1 v_1 \qquad (3.6)$$

eşitliğini sağlayan v_1 vektörü, S' uzayına izdüşümü alınan verinin ilk boyuttaki varyansını en iyi şekilde açıklayan baz vektörüdür. Görüleceği üzere optimizasyon, v_1 için gerekli norm kısıtı kullanılarak temel bir özdeğer-özvektör problemine indirgenmiştir. Bu problemin çözümüyle elde edilen λ_1 özdeğerinin büyüklüğü, v_1 baz vektörünün verideki varyansı ne kadar iyi açıkladığını göstermektedir.

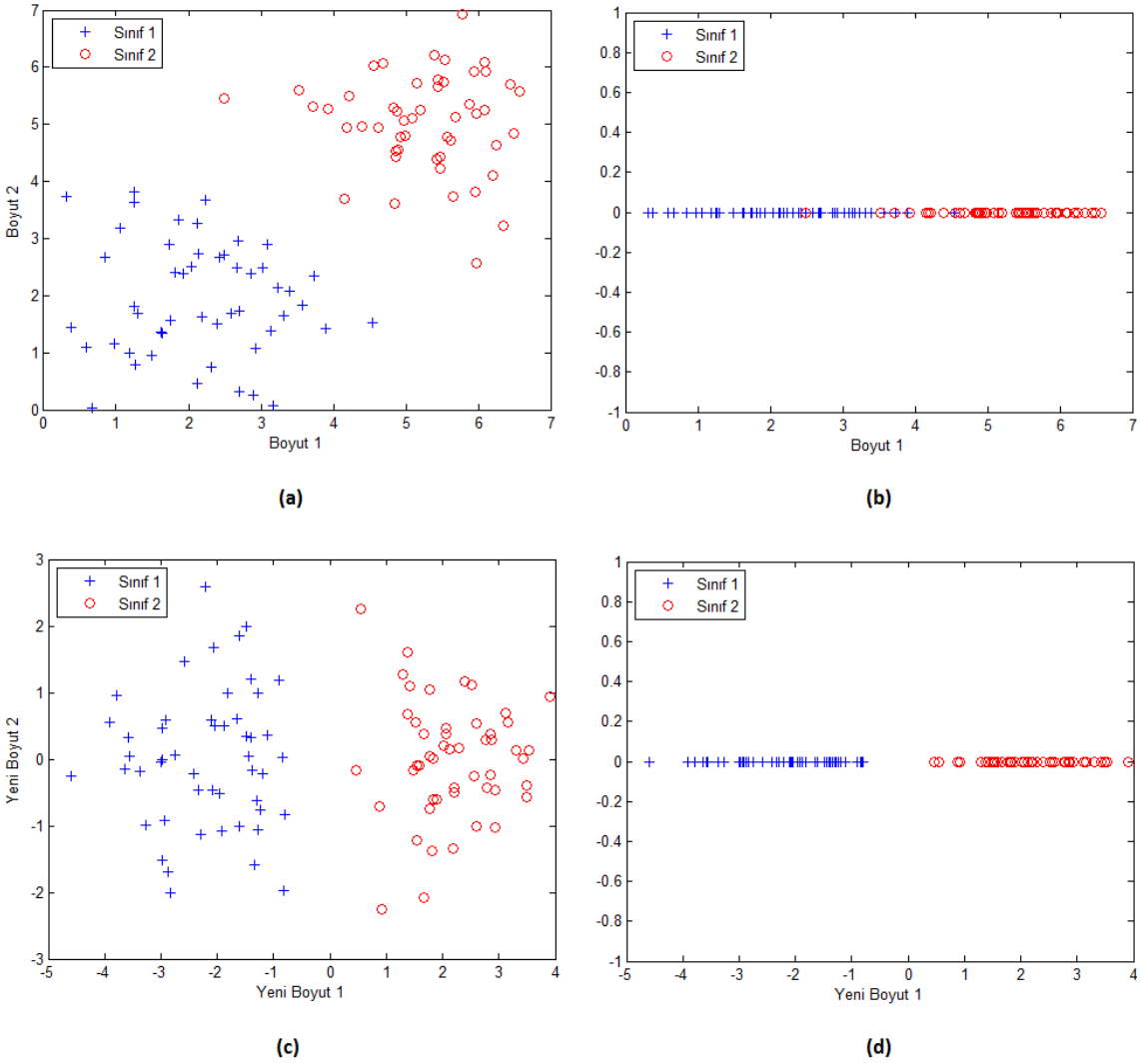
Kalan $D-1$ adet baz vektörü de aynı şekilde hesaplanarak $V=[v_1,v_2,...,v_D]$ baz vektör matrisi elde edilir. Bir sonraki aşama örnekleri merkeze çekmektir. Kümedeki her örneğin ortalamadan farkı olan y_i,

$$y_i = x_i - \overline{X} \quad , \quad i=1,...,N \qquad (3.7)$$

hesaplanır ve $Y=[y_1,y_2,...,y_N]^T$ olmak üzere örneklerin S' uzayına izdüşümü (X') alınır (Denklem 3.8).

$$X' = YV \qquad (3.8)$$

Şekil 3.1'de bu işlemin iki boyutlu bir uzayda 50'şer örnek içeren iki farklı sınıftan verinin tek boyutta (boyut azaltımı) ideal ayrımının sağlanması için kullanımı görülmektedir. S uzayındaki veri üzerinde temel bileşen analizi yapılmadan boyut azaltımı işlemi yapıldığında (Şekil 3.1b'de görüldüğü gibi *Boyut2* atıldığında) iki sınıfın örneklerinin *Boyut1* üzerinde çakıştığı görülmektedir. Ancak veri, TBA ile işlendikten sonra S' uzayına aktarılıp boyut azaltımı işlemi yapıldığında (Şekil 3.1d'de görüldüğü gibi *Yeni Boyut2* atıldığında) iki sınıfın örneklerinin birbirinden kolaylıkla ayırt edilebildiği görülmektedir. Bu işlem sonucunda bulunan *Yeni Boyut1* baz vektörü ($v_1=[0.7012 \quad 0.713]$) için $\lambda_1=5.5766$, *Yeni Boyut2* baz vektörü ($v_2=[-0.713 \quad 0.7012]$) için ise $\lambda_2=0.9054$ olarak bulunmuştur. $\lambda_1>\lambda_2$ olduğundan verinin gösterimi, varyansı en iyi açıklayan boyut olan *Yeni Boyut1* (v_1 vektörü) üzerinden gerçekleştirilmiştir.



Şekil 3.1: 50'şer örnek içeren iki sınıflı verinin TBA yöntemi ile tek boyutta ayrımı

(a) Verinin 2 boyutlu S uzayındaki görünümü
(b) Verinin S uzayında sadece Boyut1 ile ifadesi
(c) Verinin TBA ile işlenerek oluşturulan 2 boyutlu S' uzayındaki görünümü
(d) Verinin S' uzayında sadece Yeni Boyut1 ile ifadesi

3.1.1 Yüksek Boyutlu Veri Kümeleri İçin Boyut Hilesi

Temel Bileşen Analizi yöntemi, $N \times D$ boyutlu bir veri kümesinden elde edilen $D \times D$ boyutlu kovaryans matris üzerinde tanımlı bir özdeğer-özvektör problemini çözmektedir. Bu işlem $O(D^3)$ gibi bir hesaplama maliyetine sahiptir. Bu maliyet, boyut sayısının (D) örnek sayısını (N) büyük farkla geçtiği $(N << D)$ uygulamalarda (örneğin 100 kişinin 500×500 piksellik resimlerini içeren bir görüntü kümesi için) artmakta ve doğal olarak bir değerden sonra problemin çözümü mümkün olmamaktadır.

TBA, bu sorunun üstesinden “Boyut hilesi” (dimension trick) adında bir yöntemle gelmektedir (Turk, 1991).

$N \times D$ boyutlu X kümesindeki örneklerden $D \times D$ boyutlu C kovaryans matrisi Denklem 3.3’de görüldüğü gibi elde edilir. C kovaryans matrisinin *i.* özvektörü olan k_i,

$$\frac{1}{N} Y^T Y k_i = \lambda_i k_i \qquad (3.9)$$

yoluyla bulunur. Sonrasında bu denklemin iki tarafı da soldan Y ile çarpılır.

$$\frac{1}{N} Y Y^T (Y k_i) = \lambda_i (Y k_i) \qquad (3.10)$$

$v_i = Y k_i$ şeklinde tanımlanırsa Denklem 3.11,

$$\frac{1}{N} Y Y^T v_i = \lambda_i v_i \qquad (3.11)$$

halini alır. v_i vektörü, $N \times N$ boyutlu $P = \frac{1}{N} Y Y^T$ kovaryans matrisi için *i.* özvektördür. Daha küçük boyutlu bu yeni kovaryans matrisin özvektörlerini hesaplamak mümkündür. Bu hesaplama işleminden sonra Denklem 3.11 soldan Y^T ile çarpılır. Böylece,

$$\left(\frac{1}{N} Y^T Y \right) (Y^T v_i) = \lambda_i (Y^T v_i) \qquad (3.12)$$

$C = \left(\frac{1}{N} Y^T Y \right)$ kovaryans matrisinin *i.* özvektörü $(Y^T v_i)$ şeklinde hesaplanabilir ve bu Denklem 3.9’da belirtilen k_i ile aynı yönde ancak farklı büyüklüktedir. Bu özvektörün Denklem 3.13’de görüldüğü gibi normalize edilmesi gerekir (Bishop, 2006).

$$k_i = \frac{1}{\sqrt{N \lambda_i}} Y^T v_i \qquad (3.13)$$

3.2 DESTEK VEKTÖR MAKİNELERİ

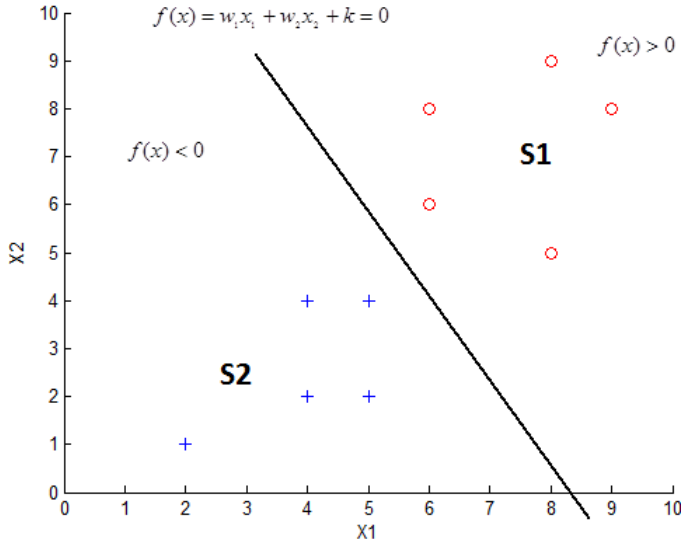
İki boyutlu sınıflandırma probleminde farklı sınıfların örneklerini birbirinden ayırt etmenin en kolay yolu, bu örnekleri birbirinden ayıran en uygun doğrunun denklemini bulmaktır. Bu şekilde yapılan sınıflandırmaya "doğrusal sınıflandırma" adı verilir. Destek Vektör Makineleri (Support Vector Machines) yöntemi, bu amaçla veriyi gerekirse yeni uzaylara taşıyabilen ve en uygun ayırıcı düzlemi (optimal separating hyperplane) arayan bir sınıflandırıcıdır.

3.2.1 Doğrusal Sınıflandırma

İki boyutlu bir uzayda mevcut olan iki sınıfın örneklerini birbirinden ayıran bir doğrunun bulunduğunu kabul edelim (Şekil 3.2). Bu doğrunun denklemi Denklem 3.14'deki gibi verilmiş olsun.

$$f(x) = w_1 x_1 + w_2 x_2 + k = w^T x + k = 0 \qquad (3.14)$$

Bu doğru, w ağırlık vektörü ve k öteleme miktarı gibi iki parametre ile tanımlanır. Herhangi bir x örneği için $f(x) > 0$ ise bu örnek S1 sınıfından aksi takdirde S2 sınıfındandır (Alpaydın, 2007).



Şekil 3.2: İki boyutlu uzayda iki sınıfın örneklerinin doğrusal olarak ayrımı

DVM yöntemi, bu şekilde tanımlanmış doğrusal sınıflandırıcıya "kenar payı" (margin) kavramını katar. Buna göre iki sınıflı ve N adet örnekli problemde u. örneğin x^u gibi

bir değeri ve $c^u \in \{-1\ +1\}$ gibi bir sınıf etiketi (S1 sınıfı için -1, S2 sınıfı için +1) olsun. Yeni sınıflandırıcının sağlaması gereken doğru denklemi

$$\begin{aligned} w^T x^u + k \geq +1 \quad & eger \quad c^u = +1 \\ w^T x^u + k \leq -1 \quad & eger \quad c^u = -1 \end{aligned} \qquad (3.15)$$

şeklindedir ve

$$c^u (w^T x^u + k) \geq +1 \qquad (3.16)$$

olarak genelleştirilebilir. Bu şekilde, *w* ağırlık vektörü ile yeni bir düzleme izdüşümü alınan örneklerin (*w* bu yeni düzlem için bir baz vektörüdür) bu düzlemi dik olarak kesen $f(x) = 0$ doğrusuna en az bir birim uzaklıkta olması hedeflenmektedir. Buradaki temel problem, *w* baz vektörünün ve *k* öteleme miktarının en uygun şekilde seçilerek örneklerin doğruya olan mesafesini yani kenar payını en büyük kılmaktır.

u. örneğin $f(x) = 0$ doğrusuna olan mesafesi,

$$\frac{c^u (w^T x^u + k)}{\|w\|} \qquad (3.17)$$

şeklinde tanımlıdır. Görüleceği üzere bu mesafeyi en büyük kılmak için $\|w\|$ değerinin mümkün olduğunca küçük seçilmesi gerekmektedir. Bu da,

$$\arg\min_{w,k} \frac{1}{2}\|w\|^2 \qquad (3.18)$$

şeklinde tanımlı bir kuadratik problemin $c^u (w^T x^u + k) \geq +1$ kısıtına bağlı olarak çözümü ile mümkündür. Bu, aynı zamanda bir en iyileme problemi olup $a^u \geq 0,\ u = 1...N$ şeklinde pozitif tanımlı Lagrange çarpanlarının kullanımıyla çözülür. Özel kısıtın ve Lagrange çarpanlarının eklenmesiyle Denklem 3.19'da görülen Lagrange fonksiyonu elde edilir.

$$L(w,k,a) = \frac{1}{2}\|w\|^2 - \sum_{u=1}^{N} a^u \{c^u (w^T x^u + k) - 1\} \qquad (3.19)$$

Öncelikle bu fonksiyonun *w* ve *k* parametrelerine göre en küçük değerini alacak hale getirilmesi gerekmektedir. Fonksiyonun bu iki parametreye göre alınan türevlerinin sıfıra eşitlenmesiyle,

$$\frac{\partial L}{\partial w} = 0 \;\rightarrow\; w = \sum_{u=1}^{N} a^u c^u x^u \qquad (3.20)$$

$$\frac{\partial L}{\partial k} = 0 \;\rightarrow\; \sum_{u=1}^{N} a^u c^u = 0 \qquad (3.21)$$

koşulları elde edilir. Bu koşullar Denklem 3.19'da yerine konursa

$$\begin{aligned} \hat{L}(a) &= \frac{1}{2}(w^T w) - w^T \sum_{u=1}^{N} a^u c^u x^u - k \sum_{u=1}^{N} a^u c^u + \sum_{u=1}^{N} a^u \\ &= \sum_{u=1}^{N} a^u - \frac{1}{2}(w^T w) \\ &= \sum_{u=1}^{N} a^u - \frac{1}{2} \sum_{u=1}^{N} \sum_{v=1}^{N} a^u a^v c^u c^v (x^u)^T x^v \end{aligned} \qquad (3.22)$$

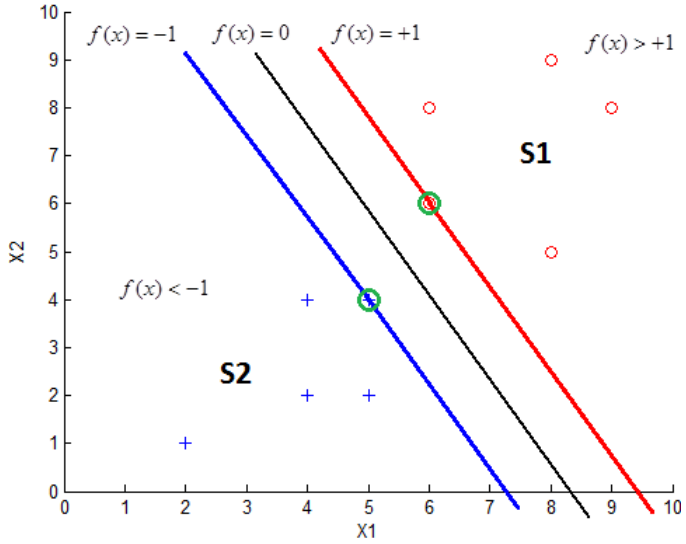
eşlek (dual) problemi elde edilir. Sonrasında eşlek problemin $\sum_{u=1}^{N} a^u c^u = 0$ ve $a^u \geq 0,\ u = 1...N$ kısıtlarına bağlı olarak a^u Lagrange çarpanlarına göre en büyük hale getirilmesi gerekir. Bu da kuadratik bir en iyileme problemidir. Problemin çözümü sonucunda çoğu x^u örneğine karşılık gelen a^u Lagrange çarpanının 0 olduğu görülür. Geriye kalan $a^u \geq 0$ koşulunu sağlayan çarpanlara karşılık gelen x^u örnekleri "Destek Vektörü" olarak adlandırılır (Bkz: Şekil 3.3). Bu vektörler, Denklem 3.20'de yerine konulduğunda verinin izdüşümünün alınması gereken en uygun *w* baz vektörü elde edilir. Söz konusu bu destek vektörleri sınıflandırma doğrusuna tam bir birim mesafede olacaktır ve

$$c^u (w^T x^u + k) = 1 \qquad (3.23)$$

eşitliğini sağlayacaktır. Herhangi bir destek vektörü kullanılarak *k* öteleme miktarı,

$$k = c^u - w^T x^u \qquad (3.24)$$

şeklinde bulunabilir. Ancak en doğrusu tüm *k* parametresini tüm destek vektörleri için hesaplayıp ortalama almaktır.



Şekil 3.3: Destek vektörleri (yeşil daire içinde gösterilen örnekler) yardımıyla bulunan en iyi ayırıcı düzlem

En uygun destek vektörlerini ve buna bağlı olarak *w* ve *k* parametrelerini belirledikten sonra sınıflandırılması istenen yeni örnekleri $f(x) = (w^T x^u + k)$ fonksiyonundan geçirip sonucun işaretine göre S1 ya da S2 sınıfından olduklarına karar verilebilir.

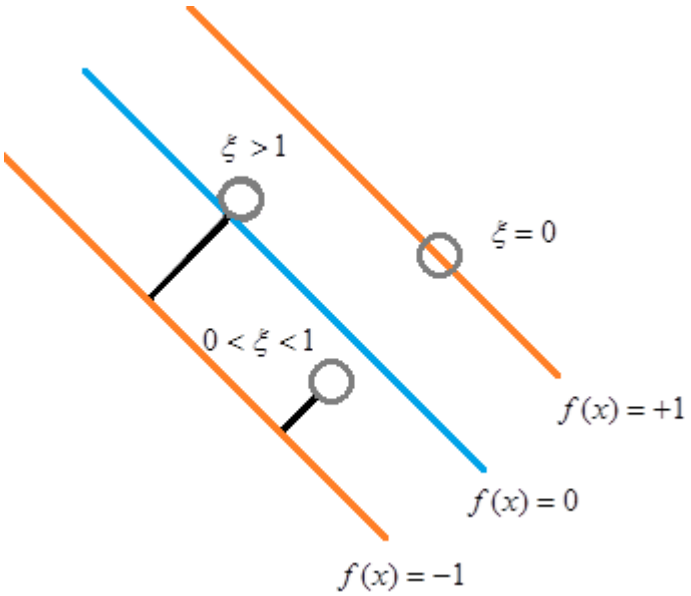
3.2.2 Doğrusal Ayrılamama Durumu

Gerçek dünya problemlerinde her zaman istenen kenar payıyla güvenilir sınıflandırma mümkün olmayabilir. Bazı eğitim örnekleri, sınıflandırma doğrusu ile destek vektörleri arasına, hatta sınıflandırma doğrusunun yanlış tarafına düşebilir. Bu durumu engellemek için eğitim kümesi "çekirdek" (kernel) adı verilen özel fonksiyonlardan geçirilerek doğrusal olarak ayrılabilecekleri daha büyük boyutlu uzaylara taşınabilir ancak bu da bir süre sonra "aşırı öğrenme" (over-learning) problemine yol açar ve verideki gürültü de gerçek bir örnekmiş gibi algılanır. Böyle eğitilen sistemlerde test hatası artmaya başlar. Bu nedenle DVM eğitimi sırasında bir miktar örneğin, kenar payının içine düşmesine veya hatalı sınıflandırılmasına "artık değişkenler" (slack variables) sayesinde izin veren bir yapı öne sürülmüştür (Vapnik, 1998).

Bu yapıda *u*. örneğe karşılık gelen artık değişken $\xi^u \geq 0$ şeklinde tanımlıdır. Bu parametre Denklem 3.16'ya eklendiğinde,

$$c^u (w^T x^u + k) \geq 1 - \xi^u \qquad (3.25)$$

elde edilir. $\xi^u = 0$ ise *u*. örnek sınıflandırma doğrusundan en az bir birim uzaktadır ve bu istenen durumdur. $0 < \xi^u < 1$ ise *u*. örnek doğru sınıflandırılıyordur ama kenar payının içine düşmektedir. $\xi^u > 1$ ise *u*. örnek yanlış sınıflandırılıyordur (Şekil 3.4). Böyle bir tanımlama ile bazı eğitim örneklerinin hatalı sınıflandırılmasına izin verilmekte ve katı kenar payı kavramı yumuşatılmaktadır.



Şekil 3.4: Doğrusal ayrılamama durumu için destek vektörlerinin yerleşimi

Doğrusal problemde amaç sadece *w* baz vektörünün normunu en küçük hale getirmek iken, doğrusal olmayan problemde bunun yanında her örneğe ait ceza parametresini de en küçük kılmak gerekmektedir.

$$\arg\min_{w,k} \frac{1}{2}\|w\|^2 + C\sum_{u=1}^{N} \xi^u \qquad (3.26)$$

Denklem 3.26'da da görüleceği üzere bu ek kısıt denkleme, tüm örnekler için izin verilen toplam ceza miktarının *C* gibi bir katsayı ile ölçeklenmesiyle eklenmektedir. Bu

denklemi eniyilemek için $c^u(w^T x^u + k) \geq 1 - \xi^u$ kısıtının yanı sıra $\xi^u \geq 0$ kısıtını da dikkate almak gerekmektedir. Bu durumda problem çift Lagrange çarpanı ($a^u \geq 0$ ve $b^u \geq 0$) ile tanımlanabilir.

$$L(w,k,a,b) = \frac{1}{2}\|w\|^2 + C\sum_{u=1}^{N}\xi^u - \sum_{u=1}^{N}a^u\{c^u(w^T x^u + k) - 1 + \xi^u\} - \sum_{u=1}^{N}b^u\xi^u \quad (3.27)$$

Tanımlanan bu Lagrange probleminin *w*, *k* ve ξ^u parametrelerine göre türevi alınıp sıfıra eşitlendiğinde Denklem 3.20, 3.21 ve 3.28'deki koşullar elde edilir.

$$\frac{\partial L}{\partial \xi^u} = 0 \;\rightarrow\; C - a^u - b^u = 0 \quad (3.28)$$

$b^u \geq 0$ koşulu gereği $0 \leq a^u \leq C$ kısıtı söz konusudur. Denklem 3.20 ve 3.21'deki eşitlikler Denklem 3.27'de yerine konduğunda,

$$\hat{L}(a) = \sum_{u=1}^{N}a^u - \frac{1}{2}\sum_{u=1}^{N}\sum_{v=1}^{N}a^u a^v c^u c^v (x^u)^T x^v \quad (3.29)$$

bulunur ki doğrusal ayrılabilen durumla birebir aynıdır ancak eniyileme işlemi için kısıt kümesi $\sum_{u=1}^{N}a^u c^u = 0$ ve $0 \leq a^u \leq C$ olarak tanımlıdır. Bu denklem çözüldüğünde, doğru sınıflandırılan ve kenar payından en az bir birim uzakta olan örnekler için $a^u = 0$ olur. $0 < a^u < C$ olan örnekler destek vektörleridir. $a^u \geq C$ olan örnekler hatalı sınıflandırılır.

3.2.3 Çekirdek Makineleri

Bölüm 3.2.2'de de değinildiği gibi sınıflandırma problemi içinde bulunulan uzayda doğrusal olarak çözülemiyorsa daha yüksek boyutlu bir uzaya geçilip doğrusal olarak çözülebilir. Bu işleme "Çekirdek hilesi" (Kernel trick) adı verilir. ϕ, *D* boyutlu uzaydaki *N* adet örneği, *M* boyutlu ($D < M$) yeni bir uzaya taşıyan önceden tanımlanmış bir operatör olsun. Bu durumda *D* boyutlu uzayındaki tüm x^u örnekleri bu operatör aracılığıyla yeni uzaydaki r^u örneklerine dönüştürülebilir.

$$r = \phi(x) \qquad (3.30)$$

Bununla beraber yeni uzayımızdaki sınıflandırma doğrusu,

$$f(x) = w^T \phi(x) \qquad (3.31)$$

şeklinde tanımlıdır. Daha büyük boyutlu yeni uzayımızda *k* öteleme parametresini ihmal edebiliriz. Bu uzayda hâlâ doğrusal sınıflandırma yapamayabiliriz. Bu nedenle artık değişkenleri kullanmaya devam ederiz. Buna göre yeni uzayımızda çözmemiz gereken eniyileme problemi Denklem 3.29'a benzer olarak,

$$\hat{L}(a) = \sum_{u=1}^{N} a^u - \frac{1}{2} \sum_{u=1}^{N} \sum_{v=1}^{N} a^u a^v c^u c^v \phi(x^u)^T \phi(x^v) \qquad (3.32)$$

şeklinde $\sum_{u=1}^{N} a^u c^u = 0$ ve $0 \le a^u \le C$ kısıtlarıyla beraber tanımlıdır. Burada çekirdek fonksiyonumuz,

$$K(x^u, x^v) = \phi(x^u)^T \phi(x^v) \quad u, v = 1...N \qquad (3.33)$$

şeklinde tanımlanabilir. Buna istinaden Denklem 3.32,

$$\hat{L}(a) = \sum_{u=1}^{N} a^u - \frac{1}{2} \sum_{u=1}^{N} \sum_{v=1}^{N} a^u a^v c^u c^v K(x^u, x^v) \qquad (3.34)$$

şeklinde yeniden yazılabilir. $f(x)$ doğru denklemi de *w* parametresi yerine Denklem 3.20 koyulduğunda,

$$f(x) = w^T \phi(x) = \sum_{u=1}^{N} a^u c^u \phi(x^u)^T \phi(x) = \sum_{u=1}^{N} a^u c^u K(x^u, x) \qquad (3.35)$$

şeklinde yazılabilir. Elimizdeki örnekleri mevcut uzayımızda bir *K* çekirdek fonksiyonundan geçirmek, ϕ operatörü ile daha yüksek boyutlu bir uzaya geçirip eleman eleman çarpmaktan daha kolaydır. Bu maksatla tanımlanmış özel çekirdek fonksiyonları vardır.

Dairesel Tabanlı Fonksiyon (Radial Basis Function) Çekirdeği:

$$K(x^u, x) = \exp\left[-\frac{\left\|x^u - x\right\|^2}{2\gamma^2}\right] \quad (3.36)$$

şeklinde tanımlıdır. γ parametresi örneğimizi yerleştirdiğimiz Gauss dağılımının genişliğini belirler. Bu parametrenin çok küçük seçilmesi aşırı öğrenme sorununa, çok büyük seçilmesi de yumuşayan sınıflandırma doğrusu nedeniyle hatalı sınıflandırılan örnek sayısının artmasına sebep olur. Probleme en uygun γ parametresi, artık değişkenler için ölçekleme parametresi olan C ile beraber sezgisel yöntemle elde edilebilir.

Bu çekirdekteki mesafe kriteri olan Öklid mesafesi yerine Manhattan ya da Mahalanobis mesafeleri de kullanılabilmektedir.

Polinom (Polynomial) Türü Çekirdek:

$$K(x^u, x) = (x^T x^u + 1)^p \quad (3.37)$$

şeklinde tanımlı p. dereceden bir fonksiyondur.

3.3 KANONİK BAĞINTI ANALİZİ

Kanonik bağıntı analizi (KBA), aynı nesneye ait iki görü kümesindeki değişkenleri doğrusal bir dönüşümle yeni bir veri uzayına geçirip kümelerin bu uzaydaki izdüşümleri arasındaki bağıntıyı en büyük kılacak izdüşüm vektörlerini bulma problemidir (Hotelling, 1936). Burada kastedilen görü kümesi çifti, bir videoda konuşmakta olan bir insanın dudak hareketleri ile sesi (Sargın, 2007), bir web sitesinde kullanılan resimlere karşılık içerikte kullanılan kelime grupları (Vinokourov, 2003) veya genomda bir virüsün entegre olduğu noktanın her iki yanındaki baz grupları (Gumus, 2012) olabilir. KBA yöntemini incelemeden önce iki dağılım arasındaki bağıntının nasıl tanımlandığını bilmek gerekir.

Bağıntı:

X ve *Y*'nin her biri *N* adet gözlem içeren iki dağılım olduğunu kabul edelim. Bu durumda μ bir dağılımın ortalamasını, σ da bir dağılımın standart sapmasını göstermek üzere bu iki dağılım arasındaki "Pearson bağıntı katsayısı" olarak bilinen r_{XY}, Denklem 3.38'de görüldüğü gibi hesaplanır.

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{(n-1)\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_Y)^2}} \qquad (3.38)$$

Bağıntı katsayısı [-1 : +1] aralığından değerler alabilir. İki dağılım arasındaki bağıntı katsayısı +1'e yaklaştıkça bu dağılımların "pozitif bağımlı", -1'e yaklaştıkça da "negatif bağımlı" olduğu söylenir. Katsayının sıfıra yakın olması bu iki dağılımın "birbirinden bağımsız" olduğu anlamına gelir.

İki dağılım arasındaki Pearson bağıntı katsayısını en büyük kılacak izdüşüm vektörlerini bulmaya yönelik çalışan KBA yöntemi şu şekilde tanımlıdır (Hardoon, 2004):

a ve *b* olarak tanımlı iki sütun vektörünün iç çarpımı olan $a^T b$ işlemi, $< a, b >$ şeklinde gösterilecektir. Elimizdeki nesnenin/gözlemin $S_X = (x_1, x_2, ..., x_n)$ ve $S_Y = (y_1, y_2, ..., y_n)$ şeklinde tanımlı, her biri *N* adet örnek içeren iki görüsü olduğunu kabul edelim. Burada x_i ve y_i, $D \times 1$ boyutlu birer sütun vektörüdür. x_i örneği, $D \times 1$ boyutlu w_X izdüşüm vektörü aracılığıyla yeni bir uzaydaki $< w_X, x_i >$ noktasına taşınabilir. Benzer şekilde S_X ve S_Y kümelerindeki tüm noktalar w_X ve w_Y vektörleri aracılığıyla $< w_X, S_X >$ ve $< w_Y, S_Y >$ nokta kümelerine taşınabilir. Bu durumda bu yeni nokta kümeleri arasındaki bağıntıyı en büyük kılan değer Denklem 3.39'da görüldüğü gibidir.

$$r_{XY} = \max_{w_X, w_Y} \frac{< w_X, S_X >< w_Y, S_Y >^T}{\sqrt{(< w_X, S_X >< w_X, S_X >^T)(< w_Y, S_Y >< w_Y, S_Y >^T)}} \qquad (3.39)$$

Denklem 3.39, 3.40'da görüldüğü şekilde tekrar yazılabilir.

$$r_{XY} = \max_{w_X, w_Y} \frac{w_X{}^T S_X S_Y{}^T w_Y}{\sqrt{(w_X{}^T S_X S_X{}^T w_X)(w_Y{}^T S_Y S_Y{}^T w_Y)}} = \frac{w_X{}^T C_{XY} w_Y}{\sqrt{(w_X{}^T C_{XX} w_X)(w_Y{}^T C_{YY} w_Y)}} \quad (3.40)$$

Burada C_{XX} ve C_{YY} sınıf içi kovaryans matrisleri, $C_{XY} = C_{YX}{}^T$ de sınıflar arası kovaryans matrisleridir. Bu matrislerin sıfır merkezine çekildikleri (bir görüdeki tüm örneklerden o görünün ortalamasının çıkartıldığı) kabul edilmiştir. Görüleceği üzere Denklem 3.40, w_X ve w_Y vektörlerine göre en büyük kılınması gereken bir optimizasyon problemidir ve bu problem Lagrange çarpanları kullanılarak çözülebilir.

Denklem 3.40 incelendiğinde payın, $w_X{}^T C_{XX} w_X = 1$ ve $w_Y{}^T C_{YY} w_Y = 1$ kısıtlarına bağlı olarak en büyük kılınması gerektiği görülmektedir. Bu durumda λ_X ve λ_Y bu kısıtları kontrol eden Lagrange çarpanları olmak üzere ilgili optimizasyon problemi Denklem 3.41'de görüldüğü gibi yazılabilir.

$$L(\lambda, w_X, w_Y) = w_X{}^T C_{XY} w_Y - \frac{\lambda_X}{2}(w_X{}^T C_{XX} w_X - 1) - \frac{\lambda_Y}{2}(w_Y{}^T C_{YY} w_Y - 1) \quad (3.41)$$

Denklem 3.41'in w_X ve w_Y vektörlerine göre türevi alınıp sıfıra eşitlendiğinde,

$$\frac{\partial L}{\partial w_X} = C_{XY} w_Y - \lambda_X C_{XX} w_X = 0 \quad (3.42)$$

$$\frac{\partial L}{\partial w_Y} = C_{YX} w_X - \lambda_Y C_{YY} w_Y = 0 \quad (3.43)$$

Denklem 3.43, $w_Y{}^T$ ile çarpılıp Denklem 3.42'nin $w_X{}^T$ katından çıkartıldığında Denklem 3.44'deki eşitlik elde edilir.

$$\begin{aligned} 0 &= w_X{}^T C_{XY} w_Y - w_X{}^T \lambda_X C_{XX} w_X - w_Y{}^T C_{YX} w_X + w_Y{}^T \lambda_Y C_{YY} w_Y \\ &= \lambda_Y w_Y{}^T C_{YY} w_Y - \lambda_X w_X{}^T C_{XX} w_X \end{aligned} \quad (3.44)$$

Burada $w_X{}^T C_{XX} w_X = 1$ ve $w_Y{}^T C_{YY} w_Y = 1$ kısıtları göz önüne alındığında denklemin bir tek Lagrange çarpanına ($\lambda = \lambda_X = \lambda_Y$) bağlı olduğu görülür. Bu noktadan sonra Denklem 3.43'den,

$$w_Y = \frac{C_{YY}{}^{-1} C_{YX} w_X}{\lambda} \qquad (3.45)$$

elde edilebilir. w_Y, Denklem 3.42'de yerine koyulduğunda,

$$C_{XY} C_{YY}{}^{-1} C_{YX} w_X = \lambda^2 C_{XX} w_X \qquad (3.46)$$

elde edilir. Artık problem, $Ax = \lambda Bx \Rightarrow B^{-1}Ax = \lambda x$ gibi bir öz-problem haline gelmiştir ve $B = C_{XX}$ matrisinin tersi alınabildiği sürece çözülebilir. Çözüm sonucunda elde edilen w_X, Denklem 3.45'de yerine koyularak w_Y de elde edilebilir.

Çekirdek Hilesi:

KBA yönteminin mevcut veri örnekleri arasındaki bağıntıyı doğrusal dönüşümler kullanarak bulamadığı durumlar da söz konusu olabilir. Böyle bir durumda mevcut uzaydaki veri örneklerinin, doğrusal olmayan bir operatör aracılığıyla daha yüksek boyutlu bir uzaya geçirilmesi gerekir. Bu işlemin $\phi(.)$ fonksiyonu aracılığıyla yapıldığını kabul edelim. Yeni uzayımızdaki *x*, *y* örnek çiftleri $K(.)$ çekirdek fonksiyonumuz aracılığıyla işlenmektedir (Hardoon, 2004).

$$K(\bar{x}, \bar{y}) = < \phi(\bar{x}), \phi(\bar{y}) > \qquad (3.47)$$

$\overline{X} = [\bar{x}_1, \bar{x}_2, ..., \bar{x}_m]$, m adet örnekten oluşan bir görü olsun. $m < N$ olmak üzere bu görüdeki örnekleri $\phi(.)$ fonksiyonu aracılığıyla N boyutlu bir uzaya taşıdığımızda $m \times N$ boyutlu yeni X ve Y veri matrislerini elde ederiz. Bu uzaydaki sınıf içi ve sınıflar arası kovaryans matrislerimiz,

$$C_{XX} = X^T X$$
$$C_{YY} = Y^T Y \qquad (3.48)$$
$$C_{XY} = X^T Y$$

olarak tanımlıdır. *N* boyutlu uzaydaki izdüşüm vektörlerimiz olan w_X ve w_Y, veri örneklerinin *m* boyutlu α ve β vektörleri üzerine izdüşümü cinsinden yazılabilir.

$$w_X = X^T \alpha$$
$$w_Y = Y^T \beta \qquad (3.49)$$

Denklem 3.48 ve 3.49'daki eşitlikleri Denklem 3.40'da yerine koyduğumuzda,

$$r_{XY} = \max_{\alpha,\beta} \frac{\alpha^T X X^T Y Y^T \beta}{\sqrt{\alpha^T X X^T X X^T \alpha \beta^T Y Y^T Y Y^T \beta}} \qquad (3.50)$$

$K_X = XX^T$ ve $K_Y = YY^T$ olarak alınırsa Denklem 3.50,

$$r_{XY} = \max_{\alpha,\beta} \frac{\alpha^T K_X K_Y \beta}{\sqrt{\alpha^T K_X K_X \alpha \beta^T K_Y K_Y \beta}} \qquad (3.51)$$

halini alır. Yeni uzayımızdaki optimizasyon problemi α ve β vektörlerine bağlıdır. Problemin çözümü için tanımlı kısıtlar,

$$\alpha^T K_X K_X \alpha = 1$$
$$\beta^T K_Y K_Y \beta = 1 \qquad (3.52)$$

şeklindedir. Bu kısıtlar Lagrange denklemine eklendiğinde,

$$L(\lambda, \alpha, \beta) = \alpha^T K_X K_Y \beta - \frac{\lambda_\alpha}{2}(\alpha^T K_X K_X \alpha - 1) - \frac{\lambda_\beta}{2}(\beta^T K_Y K_Y \beta - 1) \qquad (3.53)$$

bulunur. Lagrange denkleminin α ve β vektörlerine göre türevi alınıp sıfıra eşitlendiğinde,

$$\frac{\partial L}{\partial \alpha} = K_X K_Y \beta - \lambda_\alpha K_X K_X \alpha = 0 \qquad (3.54)$$

$$\frac{\partial L}{\partial \beta} = K_Y K_X \alpha - \lambda_\beta K_Y K_Y \beta = 0 \qquad (3.55)$$

bulunur. Denklem 3.55'in β^T katı, Denklem 3.54'ün α^T katından çıkartılırsa,

$$\begin{aligned} 0 &= \alpha^T K_X K_Y \beta - \alpha^T \lambda_\alpha K_X K_X \alpha - \beta^T K_Y K_X \alpha + \beta^T \lambda_\beta K_Y K_Y \beta \\ &= \lambda_\beta \beta^T K_Y K_Y \beta - \lambda_\alpha \alpha^T K_X K_X \alpha \end{aligned} \qquad (3.56)$$

eşitliği elde edilir. Denklem 3.52'deki kısıtlar göz önüne alındığında $\lambda = \lambda_\alpha = \lambda_\beta$ şeklinde bir tek Lagrange çarpanının olduğu görülür. Denklem 3.55'den β değeri çekildiğinde,

$$\beta = \frac{K_Y{}^{-1} K_Y{}^{-1} K_Y K_X \alpha}{\lambda} = \frac{K_Y{}^{-1} K_X \alpha}{\lambda} \qquad (3.57)$$

bulunur. Bu değer Denklem 3.54'de yerine koyulduğunda,

$$K_X K_X \alpha - \lambda^2 K_X K_X \alpha = 0 \qquad (3.58)$$

elde edilir. Bu noktadan sonra problem $I\alpha = \lambda^2 \alpha$ öz-probleminin çözümüne dayanmaktadır. Bu problemin çözüm kümesindeki tüm özdeğerler 1 olacağından bu özdeğerlere karşılık gelen α özvektörleri *m* elemanlı birim vektörler olarak alınır. Bu durumda,

$$\beta = \frac{K_Y{}^{-1} K_X}{\lambda} \qquad (3.59)$$

olarak bulunur.

3.4 SHANNON DÜZENSİZLİK ÖLÇÜTÜ

Shannon düzensizlik ölçütü (Shannon, 1948), bilgi kuramının öncüsü Claude E. Shannon tarafından veri iletimi sırasında aktarılan veri miktarındaki düzensizliği ölçmek amacıyla geliştirilmiştir. Tabi bu düzensizliği ölçebilmek için öncelikle verinin büyüklüğünü ölçebilmek gereklidir.

Bilindiği üzere bilgisayarlarda verinin simgelenişi 0 veya 1 değerini alabilen bit düzeyindedir. Buna göre söz gelimi bir bilgisayarda X sayısı için gerekli olan bit sayısı (verinin boyutu),

$$R(X) = \log_2 |X| \qquad (3.60)$$

şeklinde bulunur.

Herhangi bir anda karşı tarafa iletmek istediğimiz verinin $Y = \{A, B, C, D\}$ kümesindeki olası dört değişkenden biri olduğunu düşünelim. Bu değişkenlerle karşılaşma olasılığının da eşit ve 0.25 olduğunu kabul edelim. Bu durumda bit bazında bu değişkenlerin kodlaması $A = 00, B = 01, C = 10, D = 11$ şeklinde yapılabilir ve her karakter için iletilecek bit sayısının 2 olduğu görülür. Farklı bir örnekte, aynı Y kümesi elemanları için karşılaşılma olasılığının $P(A) = 0.5, P(B) = 0.25, P(C) = 0.125, P(D) = 0.125$ olarak dağıldığını düşünelim. Bu durumda iletilecek karakter başına düşen bit sayısını en aza indirmek için $A = 0, B = 10, C = 110, D = 111$ şeklinde bir Kanonik Huffman (Huffman, 1952) kodlaması yapılabilir. Böylece karakter başına iletilecek ortalama bit sayısı,

$$H(Y) = -\sum_{i=1}^{N} p_i \log_2 p_i , \qquad N = 4 \qquad (3.61)$$

denkleminden 1.75 olarak bulunur. Bir önceki duruma göre karakter başına 0.25 bit daha az iletim yapılmıştır. Bu değer “bilgi kazanımı” (information gain) olarak da adlandırılır. Aynı zamanda Denklem 3.61 ile hesaplanan iletilecek ortalama bit sayısı da “Shannon Düzensizliği” (Shannon Entropy) olarak adlandırılır.

Yüksek düzensizlik değeri, elimizdeki verideki dağılımın tek düze (uniform) olduğunu gösterir ve bu ilgi çekici bir durum değildir. Buna karşın düşük düzensizlik değeri, veride inişli çıkışlı bir dağılımın hakim olduğunu gösterir ki bu da düzenli dağılıma göre daha ilgi çekicidir.

3.4.1 Bağıl Düzensizlik

Bağıl düzensizlik, iki dağılım arasındaki mesafeyi ölçmek için kullanılan özel bir düzensizlik ölçütüdür ve $D(p \parallel q)$ şeklinde gösterilir. Söz gelimi elimizdeki verinin *p* dağılımından geldiğini kabul edersek bu veri kümesindeki her örneği ortalama $H(p)$ birimlik veri ile simgeleyebilirdik. Ancak bunun yerine verinin *q* dağılımından geldiğini kabul edersek bu durumda verideki her örneği ortalama $H(p) + D(p \parallel q)$ birimlik veri ile simgelememiz gerekirdi (Cover, 1991).

p ve *q* gibi iki olasılık fonksiyonu arasındaki bağıl düzensizlik,

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \qquad (3.62)$$

şeklinde hesaplanır.

3.5 MANN-WHITNEY SIRALAMA TOPLAMI TESTİ

Mann-Whitney sıralama toplamı testi (aynı zamanda Wilcoxon testi olarak da bilinmektedir) iki dağılımın birbirine benzerliğini, dağılımdaki gerçek değerler yerine bu değerlerin sıralandıktan sonraki pozisyonlarını kullanarak ölçmekte kullanılır. Test, başlangıç aşamasında "iki dağılımın benzer olduğu" hipotezi üzerine kurulur. Sıralama toplamı testi sonucunda bu hipotez *p* = 0.05 veya daha düşük bir anlamlılıkla reddedilebilirse dağılımların farklı olduğu kabul edilir. Yöntemin işleyiş basamakları şu şekildedir (Samuels, 2012):

i. $Y_1 = [12 \;\; 5 \;\; 9 \;\; 1 \;\; 20 \;\; 8]$ ve $Y_2 = [3 \;\; 50 \;\; 16 \;\; 7 \;\; 2]$ gibi iki dağılımımız olduğunu kabul edelim. Bu dağılımlar ayrı ayrı küçükten büyüğe doğru sıralanarak $Y_1' = [1 \;\; 5 \;\; 8 \;\; 9 \;\; 12 \;\; 20]$ ve $Y_2' = [2 \;\; 3 \;\; 7 \;\; 16 \;\; 50]$ bulunur. Bu dağılımların eleman sayıları $n_1 = 6$, $n_2 = 5$ şeklindedir.

ii. Y_1' örneğindeki her elemanın Y_2' örneğindeki kaç adet elemandan büyük olduğu bulunur ve elde edilen sayılar toplanarak K_1 değeri bulunur. Buna göre

$K_1 = 2+3+3+3+4 = 15$ elde edilir. $Y_1(a) = Y_2(b)$ olması durumunda söz konusu {a, b} çifti için K_1'e 0.5 eklenir.

iii. Adım ii. aynı şekilde Y_2' örneği için uygulanarak $K_2 = 1+1+2+5+6 = 15$ bulunur.

iv. İşlemin doğrulanması için $K_1 + K_2 = n_1 \times n_2$ eşitliğinin sağlanıp sağlanmadığı kontrol edilir.

v. Testin istatistik parametresi olan U_S, K_1 ve K_2'den büyük olandır. Bu durumda $U_S = \max(K_1, K_2) = 15$'tir.

vi. Testin ürettiği anlamlılık değeri olan *p*-değerini bulmak için önceden tanımlı anlamlılık tablosuna başvurulur. Buna göre $n = n_1 = 6$, $n' = n_2 = 5$ ve $U_S = 15$ için *p* değeri 0.05'den büyük çıkar ve dağılımların benzer olduğu hipotezi reddedilemez.

3.6 DİZİLİM EŞLEŞTİRME YÖNTEMLERİ

Genom derleyicileri, dizilimler arası benzerlikleri bulmak için özel dizilim eşleştirme algoritmaları kullanmaktadırlar. Bu algoritmalar temel olarak incelenen iki dizilim arasındaki benzerliğe göre bir eşleşme skoru (Matching Score) üretmektedir. Eşleştirme işleminin amacına göre benzerlik derecelendirme işlemi değişebilmektedir. Amaçlarına göre dizilim eşleştirme yöntemleri ikiye ayrılır:

Parçalı eşleştirme (Local Alignment): Bu yöntemde karşılaştırılan farklı uzunluklu dizilimlerin sadece bir kısmının eşleşmesine izin verilir. En yüksek eşleşme skorunun elde edildiği yol (path) en uygun eşleşme olarak kabul edilir. Bu işlem sırasında bazı okuma hatalarından kaynaklanmış olabilecek eklenme/silinme durumlarına izin verilir.

Bütünsel Eşleştirme (Global Alignment): Bu yöntemde ise karşılaştırılan farklı uzunluklu dizilimlerin sadece bir kısmının değil baştan sona birbirleriyle eşleştirilmesi hedeflenir. Bunun için eklenme/silinme durumlarının yanı sıra uzun boşluk atamalarına (Gap open ve Gap extension) da izin verilir. Bu yöntemle en yüksek eşleştirme skoru garanti edilmez.

En yaygın bütünsel eşleştirme yöntemi 1970 yılında Saul Needleman ve çalışma arkadaşı Christian Wunsch tarafından öne sürülen Needleman-Wunsch (Needleman, 1970) yöntemidir.

3.6.1 Needleman-Wunsch (NW) Dizilim Eşleştirme Yöntemi

NW eşleştirme yönteminde sorgu ve hedef dizilimleri arasında bir bütünsel eşleştirme işlemi yapılır. Yöntem temel olarak iki dizilimi aynı uzunlukta eşleşecek şekilde boşluklar ile birbiriyle hizalamaktadır. İki örnek dizilim ele alalım:

Sorgu dizilimi: AACAGATTACG
Hedef dizilim: ACGGATACGC

Bu iki dizilimi birbirleriyle eşleştirebilmek için öncelikle bir skor matrisinin belirlenmesi gerekmektedir. Bu örnek için Şekil 3.5'de görülen skor matrisi kullanılabilir: Doğru eşleşmeler +2 puan ile ödüllendirilirken hatalı eşleşmeler -1 puan ile cezalandırılmaktadır.

	A	**G**	**C**	**T**
A	2	-1	-1	-1
G	-1	2	-1	-1
C	-1	-1	2	-1
T	-1	-1	-1	2

Şekil 3.5: NW eşleştirme yöntemi için örnek skor matrisi

Dizilim çıkarma teknolojisindeki hatalara veya mutasyonlara (eklenme/silinme) bağlı olarak dizilimlerde karşılığı olmayan nükleotitler bulunabilir. Bu nükleotitlere karşılık boşluk (gap) ataması yapılmalıdır. Bu örnek için boşluk cezası (gp) -2 olarak belirlenmiştir.

Buna göre aşağıda verilen algoritma kullanılarak *M* eşleştirme matrisi (Şekil 3.6) ve "geri-izleme" (traceback) için kullanılacak *yön* matrisi (Şekil 3.7) oluşturulur.

NW Eşleştirme (A[1..n], B[1..m], skormatrisi[1..4, 1..4], gp)

// A: n elemanlı sorgu dizilimi,

// B: m elemanlı hedef dizilim,

// skormatrisi: nükleotit başına eşleşme/eşleşmeme ödülünü/cezasını gösteren matris,

// gp: boşluk cezası

// olmak üzere A ve B arasında NW eşleştirme yöntemini uygular. Eşleştirme matrisi

// (*M[1..m, 1..n]*) ve geri yayılımda kullanılacak (*yon[1..m, 1..n])* matrisi döndürür.

for $i \leftarrow 0$ *to* n *do* $M(i+1,1) = gp \times i$

for $j \leftarrow 0$ *to* m *do* $M(1,\ j+1) = gp \times j$

for $i \leftarrow 2$ *to* $n+1$ *do*

  for $j \leftarrow 2$ *to* $m+1$ *do*

    $eslesme = M(i-1,\ j-1) + skormatrisi\ (A(i-1), B(j-1))$

    $eklenme = M(i,\ j-1) + gp$

    $silinme = M(i-1,\ j) + gp$

    $[\ M(i,\ j)\ yon\ (i,\ j)\ ] = \max([eslesme, eklenme, silinme])$

    // $yon\ (i,\ j) = 1$ *ise eslesme,* 2 *ise eklenme,* 3 *ise silinme söz konusudur*

return M , yon

SORGU DİZİLİMİ

HEDEF DİZİLİMİ		-	**A**	**A**	**C**	**A**	**G**	**A**	**T**	**T**	**A**	**C**	**G**
	-	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
	A	-2	2	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
	C	-4	0	1	2	0	-2	-4	-6	-8	-10	-12	-14
	G	-6	-2	-1	0	1	2	0	-2	-4	-6	-8	-10
	G	-8	-4	-3	-2	-1	3	1	-1	-3	-5	-7	-6
	A	-10	-6	-2	-4	0	1	5	3	1	-1	-3	-5
	T	-12	-8	-4	-3	-2	-1	3	7	5	3	1	-1
	A	-14	-10	-6	-5	-1	-3	1	5	6	7	5	3
	C	-16	-12	-8	-4	-3	-2	-1	3	4	5	9	7
	G	-18	-14	-10	-6	-5	-1	-3	1	2	3	7	11
	C	-20	-16	-12	-8	-7	-3	-2	-1	0	1	5	9

Şekil 3.6: NW eşleştirme (*M*) matrisi

M matrisindeki her hücre, eşleştirilen iki dizilimin en baştan o ana kadarki eşleşme skorlarını tutar. Bir hücrenin değeri solundaki, üstündeki ve sol üst çaprazındaki üç hücrenin değerinden etkilenir. *M* matrisi bize herhangi uzunlukta iki dizilim parçasının eşleşme skorunu verir. Ancak bu eşleşmenin hangi yolla gerçekleşeceğini *yön* matrisi belirler.

SORGU DİZİLİMİ

HEDEF DİZİLİM		**-**	**A**	**A**	**C**	**A**	**G**	**A**	**T**	**T**	**A**	**C**	**G**
	-	0	0	0	0	0	0	0	0	0	0	0	0
	A	0	1	1	2	1	2	1	2	2	1	2	2
	C	0	3	1	1	2	2	2	2	2	2	1	2
	G	0	3	1	1	1	1	2	2	2	2	2	1
	G	0	3	1	1	1	1	1	1	1	1	1	1
	A	0	1	1	1	1	3	1	2	2	1	2	2
	T	0	3	3	1	3	1	3	1	1	2	2	2
	A	0	1	1	1	1	1	1	3	1	1	2	2
	C	0	3	3	1	3	1	3	3	1	1	1	2
	G	0	3	3	3	1	1	1	3	1	1	3	1
	C	0	3	3	1	1	3	1	3	1	1	1	3

Şekil 3.7: NW *yön* matrisi

yön matrisi oluşturulduktan sonra ideal bütünsel eşleşmenin nasıl gerçekleşeceğini görmek için bir “geri-izleme” (traceback) adımı uygulanır. Geri izleme adımında *yön* matrisinin en sağ altındaki hücre başlangıç noktası olarak kabul edilir. Bu hücreden başlayıp en sol üstteki bitiş hücresine kadar çizilen rotada eğer bu hücrenin değeri:

- 1 ise eşleşme/eşleşmeme (match/mismatch) söz konusudur ve bir sonraki adımda sol-üstteki hücreye geçilir.
- 2 ise bir eklenme (insertion) işlemi söz konusudur. Hedef dizilime bir boşluk eklenir ve soldaki hücreye geçilir.
- 3 ise bir silinme (deletion) söz konusudur. Sorgu dizilimine bir boşluk eklenir ve bir üstteki hücreye geçilir. Geri izleme adımı sonucunda izlenmesi gereken yol Şekil 3.8’de görülmektedir.

SORGU DİZİLİMİ

HEDEF DİZİLİM		**-**	**A**	**A**	**C**	**A**	**G**	**A**	**T**	**T**	**A**	**C**	**G**
	-	0	0	0	0	0	0	0	0	0	0	0	0
	A	0	1	1	2	1	2	1	2	2	1	2	2
	C	0	3	1	1	2	2	2	2	2	2	1	2
	G	0	3	1	1	1	1	2	2	2	2	2	1
	G	0	3	1	1	1	1	1	1	1	1	1	1
	A	0	1	1	1	1	3	1	2	2	1	2	2
	T	0	3	3	1	3	1	3	1	1	2	2	2
	A	0	1	1	1	1	1	1	3	1	1	2	2
	C	0	3	3	1	3	1	3	3	1	1	1	2
	G	0	3	3	3	1	1	1	3	1	1	3	1
	C	0	3	3	1	1	3	1	3	1	1	1	3

Şekil 3.8: *yön* matrisi için geri izleme adımı

Geri izleme adımına göre sorgu ve hedef dizilimler aşağıda gösterildiği gibi bütünsel şekilde eşleştirilebilir.

Sorgu Dizilimi: AACAGATTACG-

Hedef Dizilim: -ACGGA-TACGC

3.6.2 Smith-Waterman (SW) Dizilim Eşleştirme Yöntemi

Bir parçalı eşleştirme yöntemi olan SW yöntemi, Temple Smith ve Michael Waterman tarafından öne sürülmüştür (Smith, 1981). Yöntemin öne sürülen ilk halinin NW eşleştirme yöntemine göre üç önemli farkı vardır:

i. Bu yöntem parçalı eşleştirmeyi hedeflediğinden negatif skorlara izin verilmemektedir. Bu nedenle *M* eşleştirme matrisinin ilk satırı ve ilk sütunu NW eşleştirme yönteminin aksine tamamen 0 olarak atanır.
ii. Eğer bir hücredeki skor negatif çıkarsa (önceki skora boşluk ya da eşleşmeme cezası uygulanmışsa) o hücrenin değeri 0 olarak atanır.
iii. Geri izleme adımında en yüksek skorlu eşleşmenin olduğu hücre başlangıç noktası olarak kabul edilir ve 0 değerli bir hücreyle karşılaşıncaya kadar ideal parçalı eşleşme çıkartılır.

Buna göre Şekil 3.5'de verilen skor matrisi ve -6 değerli boşluk cezası kullanıldığında:

Sorgu Dizilimi: AACAGATTACG
Hedef Dizilim: ACGGATACGC
şeklinde verilen iki dizilim, SW yöntemiyle eşleştirildiğinde 9 eşleşme skoruyla,

Sorgu Dizilimi: ACAGAT
Hedef Dizilim: ACGGAT
şeklinde eşleştirilmektedir.

Biyolojik açıdan bir genomda az sayıda uzun boşlukların bulunması, çok sayıda kısa boşlukların bulunmasına göre daha olasıdır. Hem NW hem de SW eşleştirme yöntemleri, karşılaşılan her boşluk için sabit bir ceza katsayısı uygulamaktadır. Bundan yola çıkarak Osamu Gotoh (Gotoh, 1982), her boşluk için sabit bir ceza uygulanması yerine ardarda gelen boşluklar için özel bir boşluk uzatma (Gap extend) cezasının uygulanmasını önermiştir. Bu yöntem genel kabul görmüş ve bugün kullanılan SW eşleştirme algoritmasına son halini vermiştir.

SW algoritmasının önemli bir dezavantajı *m* ve *n* birim uzunluklu iki dizilim için O(*mn*) gibi bir karmaşıklığa sahip olmasıdır. Bu durum uzun dizilim çiftleri için hesaplama zamanı ve kullanılan bellek açısından büyük bir maliyete yol açmaktadır. Bu maliyeti düşürebilmek amacıyla Altschul ve arkadaşları (Altschul, 1990) "Temel Parçalı Eşleşme Tarama Aracı" (Basic Local Alignment Search Tool -BLAST)'ı geliştirmişlerdir.

3.6.3 Temel Parçalı Eşleşme Tarama Aracı (BLAST)

BLAST, sorgu dizilimlerinin büyük genom veri tabanlarında hızlı ve düşük işlem maliyetiyle aranması için geliştirilmiş bir algoritmadır. Bu algoritmada temel amaç, sorgu ve hedef dizilimler arasında, benzerliği ön tanımlı bir eşik değerini (*S*) geçen "tohum" (Seed) isimli alt-dizilimler bulup listelemek ve daha sonra eşleştirme işlemini sadece bu tohumların bulunduğu bölgelerde gerçekleştirmektir. Bu fikir işlem maliyetini SW parçalı eşleştirme yöntemine göre bir hayli düşürmekte ancak uzun boşluklar içeren sorgu dizilimlerinde eşleşme bulma ihtimalini düşürmektedir.

BLAST algoritması şu adımlardan oluşmaktadır:

i. Tohum dizilim için bir kelime uzunluğu (*w*) belirlenir. İncelenecek dizilimlerin türüne göre bu değer değişmektedir. Örneğin aminoasit dizisi eşleşmelerinde bu değer 3, nükleotit dizilim eşleşmelerinde de 11 olarak ön tanımlıdır.
ii. Sorgu dizilimindeki *w* uzunluklu tüm alt dizilimler çıkartılır. *m* uzunluklu bir dizilim için $m-w+1$ adet alt dizilim tanımlıdır. Söz gelimi *"AAAGCATATGCCAT"* şeklindeki bir nükleotit dizilimi için, her biri 11 nükleotit uzunluklu *"AAAGCATATGC"*, *"AAGCATATGCC"*, *"AGCATATGCCA"* ve *"GCATATGCCAT"* şeklinde 4 adet alt dizilim üretilir.
iii. Üretilen alt dizilimlerin her biri sırasıyla hedef veri tabanındaki *w* uzunluklu alt kelimelerle karşılaştırılır ve her kelime çifti için bir *T* eşleşme skoru belirlenir. Bu eşleşme skoru Şekil 3.5'dekine benzer bir skor matrisi kullanılarak hesaplanır. BLAST algoritması için eşleşme ödülü +5, eşleşmeme cezası da -4 olarak tanımlanmıştır.
iv. $T \geq S$ koşulunu sağlamayan alt dizilimler tohum adayı olamaz. Bu nedenle alt dizilim listesinden çıkartılırlar.
v. Tohumlar belirlendikten sonra her dizilim çiftinin eşleşme skorunu (*T*) artırmak için eşleşme sağdan ve soldan iteratif şekilde birer karakter uzatılır. Bu işlem ile "olası en yüksek eşleşme ölçüsü" (Maximal Segment Pair measure - MSPm) olarak isimlendirilen T' skoru hesaplanır. MSPm'nin azalmaya başladığı noktada iteratif uzatma işlemi durdurulur.

Son adım, örnekte görülen dizilim çifti için (altı çizili kısım sorgu dizilimdeki *w* = 11 uzunluklu bir tohumdur) açıklanacak olursa:

Hedef Dizilim: G A C T C A A C C A T T T G A C A G C C
Sorgu Dizilimi: C A A G C A T A T G C C A T

Tohum eşleşmesi için *T* skoru, $(8\times 5)+(3\times -4)=28$ olarak hesaplanmış ve bu değer ön tanımlı *S* = 25 eşik değerinden büyük olduğu için *"AGCATATGCCA"* alt dizilimi bir tohum olarak kabul edilmiştir. Bu tohumun sağına bir karakter eklendiğinde yeni MSPm (T') değeri 24'e düşmektedir. Bu nedenle sağ tarafa doğru bir uzatma işlemi

yapılamaz. Ancak tohumun sol tarafına 2 karakter eklendiğinde eşleşme skoru 38'e çıkmaktadır. Böylece bu tohum göze alındığında eşleşen ideal hedef dizilim *"CAACCATTTGACA"* şeklinde bulunmaktadır.

BLAST'ın ilk sürümünde tohumlarda boşlukların kullanılması durumu göz önüne alınmamıştır. Daha sonra geliştirilen sürümlerinde (Zhang, 2000) alt dizilimlere boşluklar eklenerek çok daha fazla tohumun üretilmesi sağlanmıştır.

3.7 SCHRODER'İN VERİ KÜMESİ

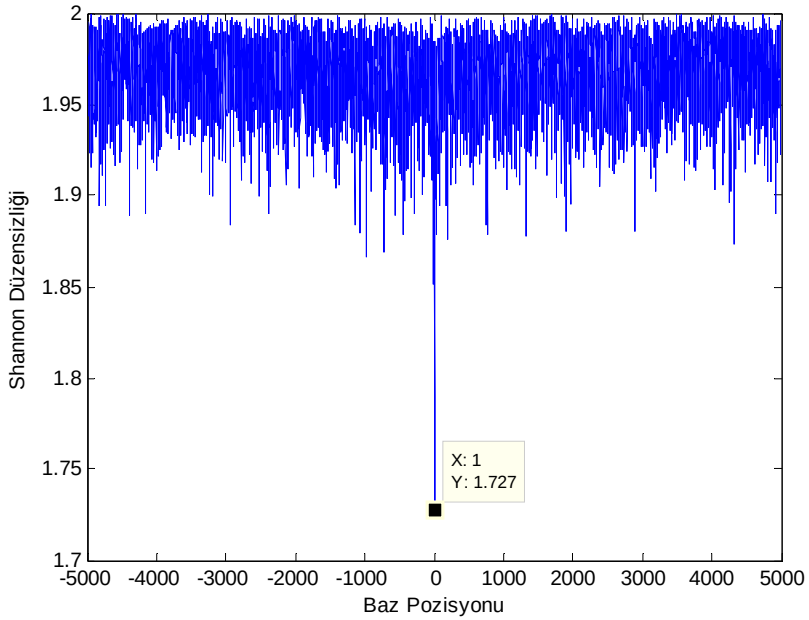
Bölüm 2.2'de Schroder'in (Schroder, 2002), oluşturduğu veri kümesindeki kimerik okumalar kullanılarak tespit edilen insan genomuna ait parçaların, virüsün yerleşim noktasına göre hizalanmasıyla beraber simetrik/palindromik bir değişim gösterdiğine değinilmişti. Bu değişim Tablo 3.1'de görülmektedir (Söz konusu değerler Genbank erişim numarası: BH609398-BH610086 aralığında olan kimerik okumalardan insan genomuna +/+ yönünde uyum sağlayan 231 adedi kullanılarak elde edilmiştir (Gumus, 2012)).

İnsan genomunda nükleotitlerin görülme olasılığı kabaca A: %30, C: %20, G: %20, T: %30 şeklindedir. Tablo 3.1 incelendiğinde bazı pozisyonlardaki nükleotit görülme olasılıklarının beklenen değerin %10 kadar üzerine çıktığı ya da altına indiği görülmektedir. x = 0 noktası ([-1 : +1] aralığı) virüsün yerleşim noktası olmak üzere nükleotit dağılımındaki bozulmanın, x = +3 noktası çevresinde, 5 nükleotit genişliğinde küçük bir alanda yüksek oranda simetrik olarak gerçekleştiği görülmektedir.

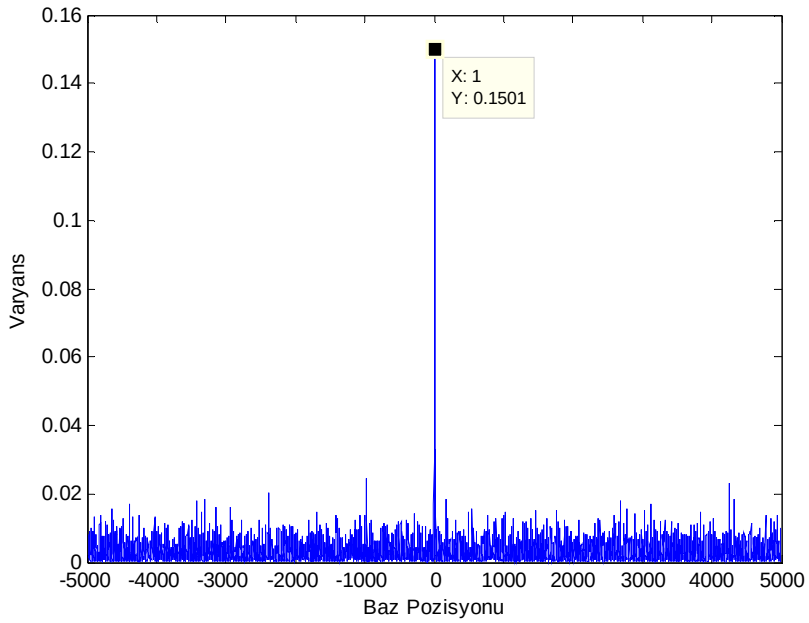
Tablo 3.1: Virüs yerleşim noktası çevresindeki nükleotit görülme olasılıkları
(Kırmızı: Beklenen değerin %10 fazlası, Mavi: Beklenen değerin %10 azı)

	-3	**-2**	**-1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
A	0.21	0.31	0.32	0.25	0.27	0.33	0.43	0.09	0.24	0.36	0.42
C	0.13	0.10	0.18	0.26	0.10	0.14	0.13	0.42	0.34	0.24	0.24
G	0.22	0.20	0.26	0.42	0.13	0.16	0.13	0.25	0.16	0.09	0.13
T	0.43	0.39	0.23	0.06	0.50	0.37	0.32	0.24	0.26	0.31	0.21

Virüs yerleşim noktasına göre + ve – yönde alınan 5000'er nükleotit uzunluğunda, toplamda 10000 nükleotitlik, bir alan incelendiğinde söz konusu simetrik bozulmanın sadece virüs yerleşim noktası civarında gerçekleştiği Şekil 3.9'da görülen Shannon düzensizliği (Bkz: Bölüm 3.4) ve Şekil 3.10'da görülen Temel Bileşen Analizi-Varyans (Bkz: Bölüm 3.1) testleri ile de görülmektedir.



Şekil 3.9: [-5000 : +5000] aralığı için Shannon düzensizliği'ndeki değişim



Şekil 3.10: [-5000 : +5000] aralığı için varyansdaki değişim

Şekil 3.9'da da görüldüğü üzere 231 adet dizilimin tümü dikkate alındığında yerleşim noktası yakınında, nükleotit dağılımındaki rastgelelik azalmaktadır. Benzer bir şekilde her baz pozisyonundaki nükleotit görülme olasılıkları komşu baz pozisyonundaki olasılıklardan, 4 boyutlu (Her nükleotit türü için bir boyut) bir uzayda Temel Bileşen Analizi kullanılarak ayrılmaya çalışıldığında x = +1 ve x = +2 pozisyonlarındaki varyansın 0.15, x = +4 ve x = +5 pozisyonlarındaki varyansın da 0.11 olduğu görülmüştür (Bkz: Şekil 3.10). Bu durum x = +1 pozisyonundan x = +2'ye geçişteki Timin (T) artışını ve x = +4 pozisyonundan x = +5'e geçişteki Adenin (A) azalmasını açıklamaktadır.

3.8 HGDP VERİ KÜMESİ

İnsan genomu çeşitliliği projesi boyunca dünya üzerindeki 51 farklı etnik gruba üye toplam 1064 bireyden kan ve doku örnekleri toplanmıştır (Bu çalışmada ulaşılan veri kümesi, söz konusu bireylerden 1043'ünün genetik bilgisini içermektedir.). Etnik gruplarına göre bireylerin dağılımı Tablo 3.2'de görülmektedir.

Tablo 3.2: HGDP kümesindeki bireylerin etnik gruplara göre dağılımı

Grup Adı	Birey Sayısı	Grup Adı	Birey Sayısı	Grup Adı	Birey Sayısı
Adygei	17	Japanese	29	Papuan	17
Balochi	25	Kalash	25	Pathan	23
Bantu	20	Karitiana	24	Pima	25
Bedouin	48	Lahu	10	Russian	25
Biaka Pygmies	32	Makrani	25	San	6
Brahui	25	Mandenka	24	Sardinian	28
Burusho	25	Maya	25	She	10
Cambodian	11	Mbuti Pygmies	15	Sindhi	25
Colombian	13	Melanesian	19	Surui	21
Dai	10	Miaozu	10	Tu	10
Daur	9	Mongola	10	Tujia	10
Druze	47	Mozabite	30	Tuscan	8
French	29	Naxi	9	Uygur	10
French Basque	24	North Italian	13	Xibo	9
Han	44	Orcadian	16	Yakut	25
Hazara	24	Oroqen	10	Yizu	10
Hezhen	9	Palestinian	51	Yoruba	24

3.8.1 Ham Verinin Sayısallaştırması

HGDP SNP veri kümesi toplamda 1043 bireyin 660918'er adet SNP bilgisini içermektedir. Bu SNP'lerin 163'er tanesi bireylerin mitokondrial DNA'sından elde edilmiş olup bu çalışma kapsamında incelenmemiştir. Kalan 660755'er adet SNP, 24 adet kromozomun (Kromozom 1-22, X ve Y) çeşitli bölgelerinden gelmektedir.

Mevcut 1043 satır - 660755 sütunlu veri kümesinin her sütununda (SNP) iki adet alel bulunmaktadır. Bu alellerden biri bireyin annesinden diğeri de babasından gelmektedir. Bu iki alelin birleşimi ile o SNP bölgesinde bir "Haplogrup" oluşmaktadır. Söz gelimi *N* numaralı SNP pozisyonunda görülen aleller Adenin (A) ve Guanin (G) olsun. Buna göre 1043 birey için o SNP pozisyonunda görülebilecek haplogruplar "AA" (homozigot baskın veya çekinik), "AG" (heterozigot) ve "GG" (homozigot çekinik veya baskın) şeklindedir. Orijinal veri kümesindeki bu haplogruplar ASCII formatında verilmiştir. Bunun yanı sıra bazı bireyler için bazı SNP pozisyonlarında okuma yapılamadığından bu pozisyonlar "--" şeklinde gösterilmiştir.

Sayısallaştırma aşamasında her SNP pozisyonu için heterozigot haplogruplar "0" ile, homozigot baskın (sayıca homozigot çekiniklerden fazla) haplogruplar "1" ile ve homozigot çekinik (sayıca homozigot baskınlardan az) haplogruplar da "-1" ile işaretlenmiştir. Böylece bireyler arası genomik mesafe hesaplanırken herhangi bir homozigot haplogrubun heterozigot başka bir haplogruba mesafesi 1 birim ile ölçülürken, birbirlerine göre uçta bulunan homozigot haplogruplar arası mesafe 2 birim ile ölçülmüştür. Sayısallaştırma işleminin son aşamasında ise "--" ile gösterilen alanlar, o SNP pozisyonunda en çok bulunan haplogrubun değeri ile değiştirilmiştir.

Sayısallaştırma işlemine ek olarak ilerideki çalışmalarda Bağlantı Eşitsizliği'nin (Linkage Disequilibrium) incelenebilmesi maksadıyla ham veride sıralaması rasgele olan SNP'ler, kromozom üzerindeki sıralarına göre tekrar dizilmiştir.

3.8.2 Verinin Alt Kümelere Ayrımı

1043 satır - 660755 sütunlu bir veri matrisi, her hücrenin 1 byte'lık alan kapladığı düşünülürse yaklaşık olarak 700 Megabyte'lık bir bellek alanına ihtiyaç duymaktadır. Kısıtlı bellek alanı ve veri boyutu arttıkça hesaplama hızındaki düşüş göz önüne

alındığında söz konusu verinin anlamlı bir şekilde alt parçalara bölünmesi gerekmektedir. Bu görevi de zaten doğa en anlamlı şekilde üstlenmiş ve yaklaşık üç milyar adet nükleotitten oluşan DNA'mızı 23 kromozom çiftine bölmüştür. Bu nedenle verinin kolay işlenebilmesi maksadıyla mevcut veri kümesi de SNP'lerin ait oldukları kromozom indislerine göre alt parçalara bölünmüştür. Buna göre elde edilen alt parçaların içerdikleri SNP sayıları Tablo 3.3'de verilmiştir.

Tablo 3.3: Kromozomlara göre SNP dağılımı

Kromozom No	**SNP Sayısı**	**Kromozom No**	**SNP Sayısı**
1	49.639	13	25.191
2	53.765	14	21.450
3	44.564	15	19.594
4	39.942	16	19.727
5	40.976	17	16.629
6	43.239	18	20.165
7	35.507	19	10.739
8	37.282	20	16.911
9	31.192	21	9.645
10	34.493	22	9.730
11	32.005	X	16.472
12	31.873	Y	10

3.8.3 Kullanılan Alt Kümede SNP Ön Elemesi

Coğrafi mesafe - genomik mesafe ilişkisinin incelenmesine yönelik olarak söz konusu 1043 bireyin ait oldukları millet gruplarından mümkün olduğunca kapalı (genetik olarak başka ırklarla az karışmış) olan etnik gruplar seçilmiştir. Bu grupların SNP verileri incelendiğinde bazı SNP sütunlarında büyük oranda aynı sayısal değerin bulunduğu görülmüştür. Söz konusu bu SNP kolonları etnik grupların birbirinden ayırt edilebilmesi için gerekli bilgiyi taşımamaktadır. Bu nedenle bu kolonların tespit edilip veri kümesinden silinmesi gerekmektedir. Bu noktada kullanılan tespit yöntemi "Minor Allele Frequency" (MAF) tekniğidir. Buna göre SNP veri kümesinin ham halindeki N numaralı SNP'i ele alalım. Bu SNP'de toplam X adet birey için $2X$ adet alel bulunmaktadır. Eğer en az sayıda bulunan alelin sayısının, tüm alellerin sayısına oranı

(minor allele frequency) %5'ten küçükse söz konusu SNP'in ayırt edici etkisinin olmadığı kabul edilir ve bu SNP, veri kümesinden çıkartılır.

Çalışma süresince SNP sayısı açısından küçük bir küme olan kromozom 21'e ait veri kümesi kullanılmıştır. Kromozom 21, Tablo 3.3'de de görüldüğü üzere 9645 adet SNP içermektedir. MAF tekniğinin uygulanmasından sonra bu sayı 8909'a inmiştir.

3.8.4 Kullanılacak Etnik Grupların Seçimi ve Coğrafi-Genomik Mesafelerin Çıkartılması

Çalışma kapsamında birbirlerinden yeterince uzak ve kendi içlerinde kapalı 12 etnik grup seçilmiştir. Gruplardaki toplam birey sayısı ve dünya üzerindeki coğrafi koordinatları Tablo 3.4'de görülmektedir.

Tablo 3.4: Seçilen etnik grupların özeti

Grup No	Etnik Grup	Birey Sayısı	Coğrafi Konum
1	Mozabite	30	32° Kuzey, 3° Doğu
2	Biaka Pygmies	11	4° Kuzey, 17° Doğu
3	Yoruba	24	6-10° Kuzey, 2-8° Doğu
4	Mandenka	24	12° Kuzey, 12° Batı
5	Cambodia	11	12° Kuzey, 105° Doğu
6	Japanese	29	38° Kuzey, 138° Doğu
7	Balochi	25	30-31° Kuzey, 66-67° Doğu
8	Yakut	25	62-64° Kuzey, 129-130° Doğu
9	Adygei	17	44° Kuzey, 39° Doğu
10	Orcadian	16	59° Kuzey, 3° Batı
11	French Basque	24	43° Kuzey, 0° Doğu
12	Sardinian	28	40° Kuzey, 9° Doğu
	Toplam	264	

Bu grupların yerküre üzerindeki dağılımı Şekil 3.11'de görülmektedir. Seçilen gruplar arasındaki coğrafi mesafe de, dünya gibi küresel düzlemler üzerinde koordinatları verilen iki nokta arasındaki kuş uçuşu mesafeyi hesaplamakta kullanılan "Vincenty"

formülü (Vincenty, 1975) kullanılarak elde edilmiştir. Buna göre seçilen etnik gruplar arası kuş uçuşu coğrafi mesafe kilometre cinsinden Tablo 3.5'de verilmiştir.

Şekil 3.11: Seçilen etnik grupların dünya üzerindeki dağılımı

Tablo 3.5: Seçilen etnik grupların ikili coğrafi mesafeleri

Grup No	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0	3427	2665	2696	10418	10965	5960	8462	3397	3034	1249	1039
2	3427	0	1400	3314	9709	12379	5958	10721	4930	6340	4636	4068
3	2665	1400	0	1915	10911	12971	6839	10850	5173	5699	3909	3567
4	2696	3314	1915	0	12583	13654	8252	11058	6009	5270	3626	3726
5	10418	9709	10911	12583	0	4361	4459	5985	7176	9873	10296	9666
6	10965	12379	12971	13654	4361	0	6478	2840	7824	8668	10085	9930
7	5960	5958	6839	8252	4459	6478	0	5676	2842	5994	5941	5247
8	8462	10721	10850	11058	5985	2840	5676	0	5795	5883	7432	7433
9	3397	4930	5173	6009	7176	7824	2842	5795	0	3283	3127	2511
10	3034	6340	5699	5270	9873	8668	5994	5883	3283	0	1792	2277
11	1249	4636	3909	3626	10296	10085	5941	7432	3127	1792	0	821
12	1039	4068	3567	3726	9666	9930	5247	7433	2511	2277	821	0

Mevcut veri uzayındaki sayısal SNP değerleri üzerinden bireyler arası genomik mesafelerin hesaplanıp, ardından gruplar arası ortalama genomik mesafeler çıkartıldığında Tablo 3.6'daki değerler elde edilmiştir. Burada kullanılan mesafe metriği öklid mesafesidir.

Tablo 3.6: Seçilen etnik grupların ortalama ikili genomik mesafeleri

Grup No	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	76.01	85.03	83.77	83.79	83.81	84.89	82.47	84.38	80.88	80.86	81.00	81.01
2	85.03	66.29	77.28	77.86	89.59	90.97	90.37	91.34	89.52	90.46	90.31	90.70
3	83.77	77.28	71.92	76.41	88.52	89.78	89.17	89.80	88.37	89.19	89.27	89.85
4	83.79	77.86	76.41	72.77	88.46	89.70	89.19	89.63	88.20	89.01	89.09	89.57
5	83.81	89.59	88.52	88.46	67.33	76.09	82.90	77.23	81.95	83.11	83.60	83.95
6	84.89	90.97	89.78	89.70	76.09	71.51	83.23	76.17	82.35	83.54	84.09	84.51
7	82.47	90.37	89.17	89.19	82.90	83.23	77.19	82.58	79.71	80.01	80.65	80.92
8	84.38	91.34	89.80	89.63	77.23	76.17	82.58	70.81	81.38	82.17	82.84	83.65
9	80.88	89.52	88.37	88.20	81.95	82.35	79.71	81.38	72.87	78.08	78.45	78.65
10	80.86	90.46	89.19	89.01	83.11	83.54	80.01	82.17	78.08	71.37	77.60	77.95
11	81.00	90.31	89.27	89.09	83.60	84.09	80.65	82.84	78.45	77.60	73.92	78.04
12	81.01	90.70	89.85	89.57	83.95	84.51	80.92	83.65	78.65	77.95	78.04	74.41

Tablo 3.6'dan da görüleceği ve beklendiği üzere her etnik grubun grup içi ortalama genomik mesafesi bir başka grup ile olan ortalama genomik mesafesinden daima düşüktür. Tablo 3.5 ve 3.6'da görülen veri matrisleri vektör formuna getirilip aralarındaki bağıntı (jeo-genomik bağıntı) ölçüldüğünde %52 düzeyinde olduğu görülmektedir. Bölüm 3.1.1'de değinilen yöntem kullanılarak bu bağıntının düzeyi artırılabilmektedir. Bu durum, Bölüm 4.4'de gösterilmiştir.

3.9 ÖRÜNTÜ TARAMA ARACI

Bu bölümde, çalışma kapsamında hazırlanmış örüntü tarama aracı, kullandığı parametreler ve dizilim özellikleriyle beraber açıklanmıştır.

3.9.1 Dizilim Özellikleri

Kullanılan dizilim özellikleri n-mer (base, dimer ve trimer) frekansları ve moment değişkenleri olmak üzere iki ana başlık altında toplanmaktadır.

3.9.1.1 N-mer Frekansları

α gibi bir alfabeden oluşturulmuş X diziliminin N-mer frekansı, bu dizilim içinde N uzunluklu $\{\alpha\}^N$ adet farklı alt tümcenin toplam geçiş sayısının X diziliminin uzunluğuna oranı olarak tanımlanmaktadır. Burada $\{\alpha\}$, α alfabesindeki eleman sayısını göstermektedir. Alfabemiz işlemekte olduğumuz probleme göre değişmektedir. Buna göre DNA dizilimi için $\alpha = \{A, C, G, T\}$ şeklinde tanımlanırken, protein dizilimleri için alfabemiz $\alpha = \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V, U, O\}$ şeklinde (amino asit isimlerine göre sıralanmışlardır) tanımlıdır.

Geliştirilen örüntü tarama aracı, base (N=1), dimer (N=2) ve trimer (N=3) frekans değerleriyle çalışabilmektedir. $\alpha = \{A, C, G, T\}$ gibi bir alfabede tanımlı her biri L_k uzunluklu K adet dizilimin her birinin base (nükleotit) frekansları olan (b_i^k), Denklem 3.63'de görüldüğü gibi hesaplanmaktadır.

$$b_i^k = \frac{1}{L_k} \sum_{j=1}^{L_k} X_{i,j}^k \quad , \qquad i = 1,2,3,4 \; ve \; k = 1,...,K \tag{3.63}$$

Burada j, her dizilimdeki nükleotit pozisyonunu ve

$$X_{i,j}^k = \begin{Bmatrix} 1 \text{ eğer } \alpha_i \text{ k. dizilimin j. pozisyonunda varsa} \\ 0 \text{ eğer } \alpha_i \text{ k. dizilimin j. pozisyonunda yoksa} \end{Bmatrix}$$

göstermektedir. Bunun yanı sıra, b_1^k, b_2^k, b_3^k ve b_4^k, sırasıyla $\alpha = \{A, C, G, T\}$ alfabesindeki A, C, G, T elemanlarının k. dizilimde görülme sıklıklarını göstermektedir.

Benzer bir şekilde her dizilimin dimer frekansları (d_i^k), $\beta = \{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT\}$ alfabesine göre Denklem 3.64'de görüldüğü gibi hesaplanabilir.

$$d_i^k = \frac{1}{L_k - 1} \sum_{j=1}^{L_k - 1} Y_{i,j:j+1}^k \ , \qquad i = 1,...,16 \ ve \ \ k = 1,..., K \qquad (3.64)$$

Burada j, her dizilimdeki dimer pozisyonunu ve

$$Y_{i,j}^k = \begin{cases} 1 \text{ eğer } \beta_{\text{i}} \text{ k. dizilimin[j : j+1]. pozisyonunda varsa} \\ 0 \text{ eğer } \beta_{\text{i}} \text{ k. dizilimin[j : j+1]. pozisyonunda yoksa} \end{cases}$$

göstermektedir. $d_1^k, d_2^k, ..., d_{16}^k$, sırasıyla β alfabesindeki AA, AC,..., TT eleman çiftlerinin $k.$ dizilimde görülme sıklıklarını göstermektedir.

Son olarak trimer frekansları $(t_i^k), \gamma = \{AAA, AAC, AAG, ..., TTG, TTT\}$ alfabesine göre Denklem 3.65'de görüldüğü gibi hesaplanabilir.

$$t_i^k = \frac{1}{L_k - 2} \sum_{j=1}^{L_k - 2} Z_{i,j:j+2}^k \ , \qquad i = 1,...,64 \ ve \ \ k = 1,..., K \qquad (3.65)$$

Burada j, her dizilimdeki trimer pozisyonunu ve

$$Z_{i,j}^k = \begin{cases} 1 \text{ eğer } \gamma_{\text{i}} \text{ k. dizilimin[j : j+2]. pozisyonunda varsa} \\ 0 \text{ eğer } \gamma_{\text{i}} \text{ k. dizilimin[j : j+2]. pozisyonunda yoksa} \end{cases}$$

göstermektedir. $t_1^k, t_2^k, ..., t_{64}^k$, sırasıyla γ alfabesindeki AAA, AAC,..., TTT eleman üçlülerinin $k.$ dizilimde görülme sıklıklarını göstermektedir.

3.9.1.2 Moment Değişkenleri

Dizilim kümelerinde kullanılabilecek bir diğer özellik türü de moment değişkenleridir. Bu değişkenler "Ortalama Nükleotit Pozisyonları" ve "Ortalama Nükleotit Pozisyonlarının Varyansları" (Shi, 2006) olmak üzere iki tanedir. $\alpha = \{A, C, G, T\}$ gibi bir alfabede tanımlı her biri L_k uzunluklu K adet dizilimin her birinin (X^k) ortalama nükleotit pozisyonları olan (m_i^k), Denklem 3.66'da görüldüğü gibi hesaplanabilir.

$$m_i^k = \frac{1}{S_i^k} \sum_{j=1}^{L_k} X_{i,j}^k \, j \ , \qquad i = 1,2,3,4 \ ve \ \ k = 1,..., K \qquad (3.66)$$

Benzer şekilde aynı dizilimler için ortalama nükleotit pozisyonlarının varyansları olan (v_i^k) da Denklem 3.67'de görüldüğü gibi hesaplanabilir.

$$v_i^k = \frac{1}{S_i^k} \sum_{j=1}^{L_k} X_{i,j}^k (j - m_i^k)^2 \quad , \qquad i = 1,2,3,4 \; ve \; k = 1,...,K \tag{3.67}$$

Her iki denklem için de j nükleotit pozisyonunu, S_i^k da *i.* türden nükleotitin *k.* dizilimde toplam kaç adet bulunduğunu $\left(S_i^k = \sum_{j=1}^{L_k} X_{i,j}^k \right)$ göstermektedir.

3.9.2 Özellik Vektörlerinin Kullanımı

N-mer frekanslarına bağlı özellik vektörlerinin oluşturulması sırasında, özellikler arasında bulunan doğrusal bağımlılık nedeniyle (frekansların toplamı 1'e eşittir) bir özellik dışarıda bırakılmaktadır. Buna göre nükleotit, dimer ve trimer frekansları için sırasıyla $b^k = [b_1^k b_2^k b_3^k]$, $d^k = [d_1^k d_2^k ... d_{15}^k]$ ve $t^k = [t_1^k t_2^k ... t_{63}^k]$ vektörleri kullanılmaktadır.

Kanonik bağıntı analizi, iki görü arasındaki bağıntıyı en büyük kılmak için en uygun ağırlık katsayılarını bulmak için kullanılmaktadır. Burada kullanılan iki görüden kasıt, virüs yerleşim noktasının solundan ve sağından seçilen L_{sol} ve $L_{sağ}$ uzunluklu bölgelerden çıkartılan özelliklerdir (n-mer ya da moment özellikleri). İki görüde de n-mer özelliklerinin kullanıldığını kabul edelim. KBA, sol görüdeki $D = [d_1 d_2 ... d_{P-1}]^T$ özelliklerine karşılık gelen $A = [a_1 a_2 ... a_{P-1}]^T$ ağırlık vektörünü ve sağ görüdeki $E = [e_1 e_2 ... e_{P-1}]^T$ özelliklerine karşılık gelen $B = [b_1 b_2 ... b_{P-1}]^T$ ağırlık vektörünü,

$$A^T D \approx B^T E \tag{3.68}$$

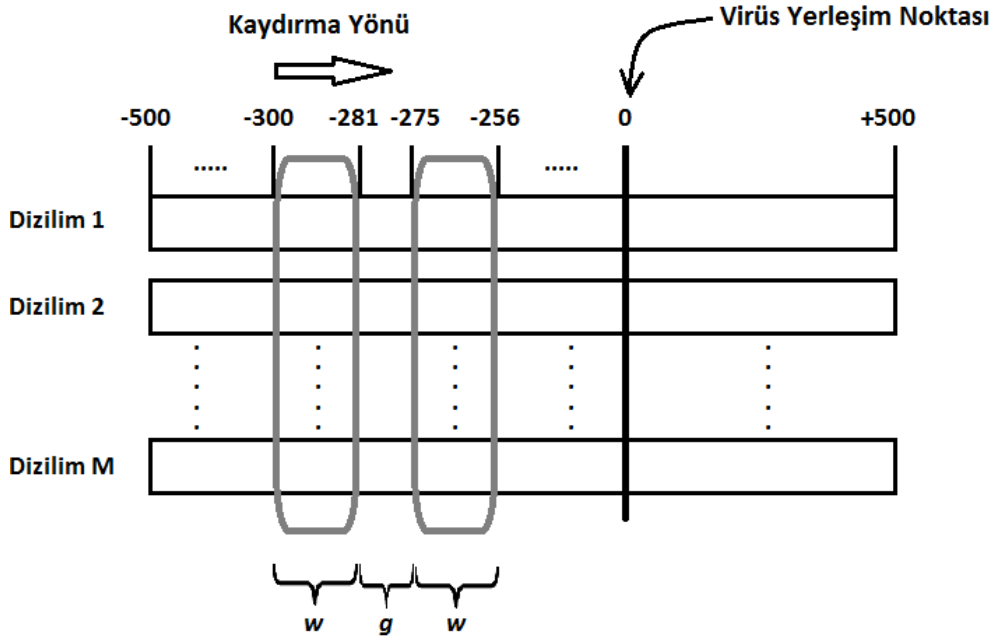
sağlayacak şekilde seçmektedir. Burada dikkat edilmesi gereken nokta $P-1$ adet özellik için $P-1$ adet ağırlık katsayısının elde edilmesidir. $\overline{D} = [d_1 d_2 ... d_{P-1} d_P]^T$, $\overline{E} = [e_1 e_2 ... e_{P-1} e_P]^T$ olmak üzere sol ve sağ görü için P adet özelliğe karşılık gelen gerçek ağırlık vektörleri $\overline{A} = [a_1 a_2 ... a_{P-1} a_P]^T$ ve $\overline{B} = [b_1 b_2 ... b_{P-1} b_P]^T$'dir. Herhangi bir

görü için (örneğin sol görü), gerçek ağırlık vektörü ($\overline{A}$) Denklem 3.69'da görüldüğü gibi hesaplanabilir.

$$A^T D = \overline{A}^T \overline{D}$$
$$(A^T D\overline{D}^T (\overline{D}\overline{D}^T)^{-1})^T = \overline{A} \qquad (3.69)$$

3.9.3 Önerilen Yöntemin Kullanımı

KBA yönteminin viral yerleşim noktasına özgü özelliklerin çıkarımında kullanılması için Şekil 3.12'de görüldüğü gibi virüs yerleşim noktasına göre hizalanmış (üst üste getirilmiş) bir grup eğitim dizilimi üzerinde bir pencere çiftinin yerleştirilmesi gerekmektedir. Burada kullanılan dizilimler, virüsün yerleşim noktasını kapsayan (- ve + yönde) ve tamamı insan genomuna ait olan dizilimlerdir.



Şekil 3.12: KBA yönteminin M adet dizilim üzerinde kullanımı

Sonrasında pencere gruplarının (görülerin) içine düşen alandan Bölüm 3.9.1.1 ve 3.9.1.2'de değinilen dizilim özellikleri çıkartılmakta ve görülerden öğrenilen fonksiyonların bağıntısını en büyük kılan ağırlık katsayılarını bulmak amacıyla KBA kullanılmaktadır. Bu noktada iki parametre öne çıkmaktadır: pencere genişliği (w) ve pencereler arası boşluk (g).

Problemde aranan özelliğe (örüntü) uygun pencere genişliğini (w) ve pencereler arası boşluğu (g) seçmek deneysel bir çaba gerektirmektedir. Örneğin, w değerini küçük tutmak bir örüntüyü tespit etmeye yetecek kadar dizilim özelliğinin çıkartılmasına engel olabilir. Öte yandan, büyük bir w değeri de küçük bir örüntünün gözden kaçmasına neden olabilir. Benzer bir şekilde pencereler arasında bulunabilecek olası tekrarlayan bölgelerin (repetitive region) bağıntıya etkisini azaltmak için tanımlanan boşluk (g) parametresinin seçimi de problemin doğasına göre değişmektedir. Her dizilimde tekrarlayan bir örüntü her zaman aranan (ilgilenilen) bir özellik olmayabilir. Böyle bir örüntünün söz konusu olduğu bir veri kümesinde g parametresinin küçük seçilmesi görüler arasında yüksek ama sahte bağıntıya yol açabilir. Bunun nedeni g parametresinin tekrarlayan örüntüyü görmezden gelecek kadar büyük seçilmemesidir. Bunun aksine küçük örüntülerin tespit edilebilmesi amacıyla da g parametresinin küçük, gerekirse negatif dahi seçilmesi gerekebilir.

4. BULGULAR

Bu bölümde, tez çalışması kapsamında hazırlanan örüntü tarama aracıyla alınan sonuçlar ve diğer uygulamalara ait bulgular sunulmaktadır.

4.1 VİRÜS YERLEŞİM BÖLGELERİNDEKİ SİMETRİK/PALİNDROMİK DAVRANIŞIN TESPİTİ

Bölüm 2.2'de, Schroder'in çalışmasından elde edilen kimerik dizilimlerin insan genomuna eşlenmesi sonucunda söz konusu çalışmada kullanılan HIV türü virüsün konak genoma yerleştiği bölgenin +3 nükleotit sağı merkez olmak üzere, bu merkezin çevresindeki küçük bir alanda nükleotit görülme olasılıkları açısından simetrik/palindromik bir davranış gözlemlendiğinden bahsedilmişti. Bu davranışın ya da bir benzerinin, yerleşim alanını çevreleyen bölgelerde tekrarlanıp tekrarlanmadığını görebilmek için bu çalışma kapsamında geliştirilen "Kanonik Bağıntı Analizi" temelli örüntü tarama aracı değişik parametreler ve dizilim özellikleriyle beraber kullanılmıştır.

KBA yöntemi, Şekil 3.12'de görüldüğü üzere her biri 1000 adet nükleotit içeren M adet dizilimin üzerine yerleştirilen ve aralarında g birim (nükleotit) boşluk bulunan w genişlikli iki pencerenin (görünün) x = -500 noktasından x = +500 noktasına doğru kaydırılması ile uygulanmaktadır (x = 0 noktası kullanılmamıştır). x = +1 noktası , çalışmada kullanılan virüsün insan genomuna yerleştiği noktanın başlangıcını (ilk nükleotidini) göstermekte olup kullanılan M adet dizilimin hepsi için virüsün yerleşimi $x > 0$ pozisyonunda (yerleşim noktasının sağında) gerçekleşmiştir. Kayan pencere yapısıyla çalışmanın gerekçesi, herhangi bir x pozisyonunda bulunan bağıntının seçilen aralıktaki başka bir pozisyonda daha görülüp görülmediğinin tespit edilmesidir. Her bir nükleotit pozisyonundaki bağıntının ne kadar anlamlı ya da anlamsız olduğu, alınan tüm ölçümleri kullanan bir istatistik testi (z-test) ile ortaya çıkmaktadır.

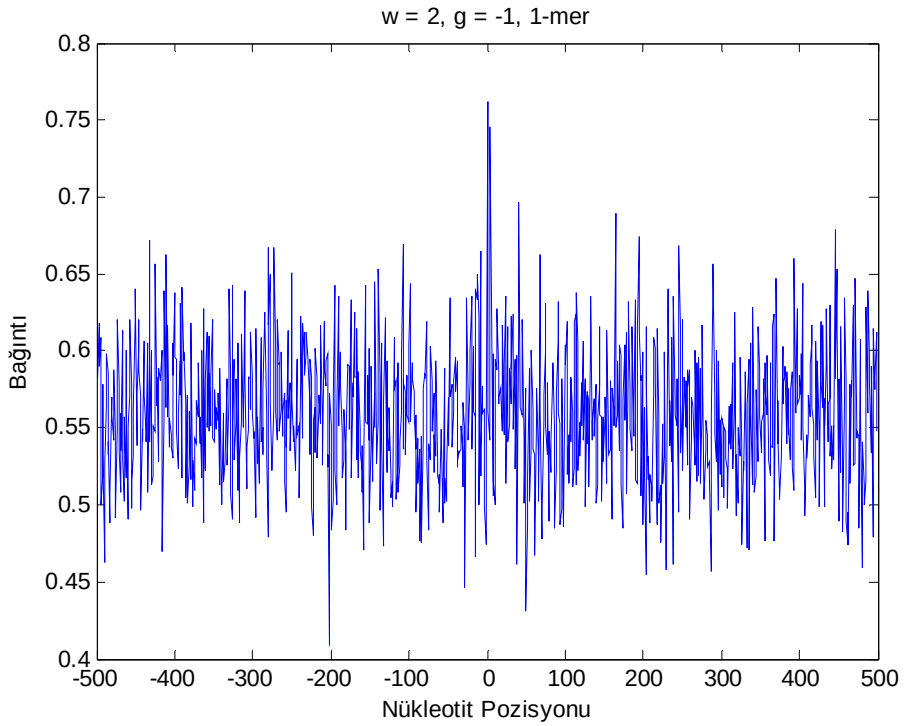
Uygulama aşamasında, Bölüm 2.2'de de değinildiği gibi Schroder'in çalışmasından (Schroder, 2002) elde edilen kimerik dizilimler kullanılmıştır. Bu dizilimlerin insan genomuna ait olan kısımları BLAST algoritması (Bkz: Bölüm 3.6.3) ile taranarak virüsün üzerine yerleştiği, insan genomuna ait 10000 nükleotit uzunluklu dizilim

parçaları da elde edilmiş ve tüm dizilimler virüs yerleşim noktasına göre hizalanmıştır. Mevcut 629 adet dizilimden insan genomuyla +/+ yönünde eşleşen 231 adedi KBA yöntemi ile taranmıştır. Denemeler sırasında söz konusu dizilimler 5×2 çapraz geçerleme tekniği ile eğitim ve test kümelerine ayrılmıştır. Her denemede eğitim kümesindeki dizilimleri kapsayan pencere çiftlerinden Bölüm 3.9.1.1 ve 3.9.1.2'de değinilen dizilim özellikleri çıkartılmış ve görülerin çıktıları arasındaki bağıntıyı (eğitim bağıntısı) en büyük kılan ağırlık vektörleri elde edilmiştir. Sonrasında bu vektörler, test kümesindeki aynı pencere pozisyonundan çıkartılan dizilim özellikleri ile çarpılmış, test görü çıktıları ve bu çıktıların bağıntısı (test bağıntısı) hesaplanmıştır. Burada amaç, çapraz geçerleme ile yapılan 10 denemede daha önceden görülmemiş test dizilimleri için ortalama en yüksek test bağıntısını veren bölgeyi bulmak ve bu bölgedeki bağıntının rasgele olmadığı hipotezini istatistik testi ile doğrulayabilmektir.

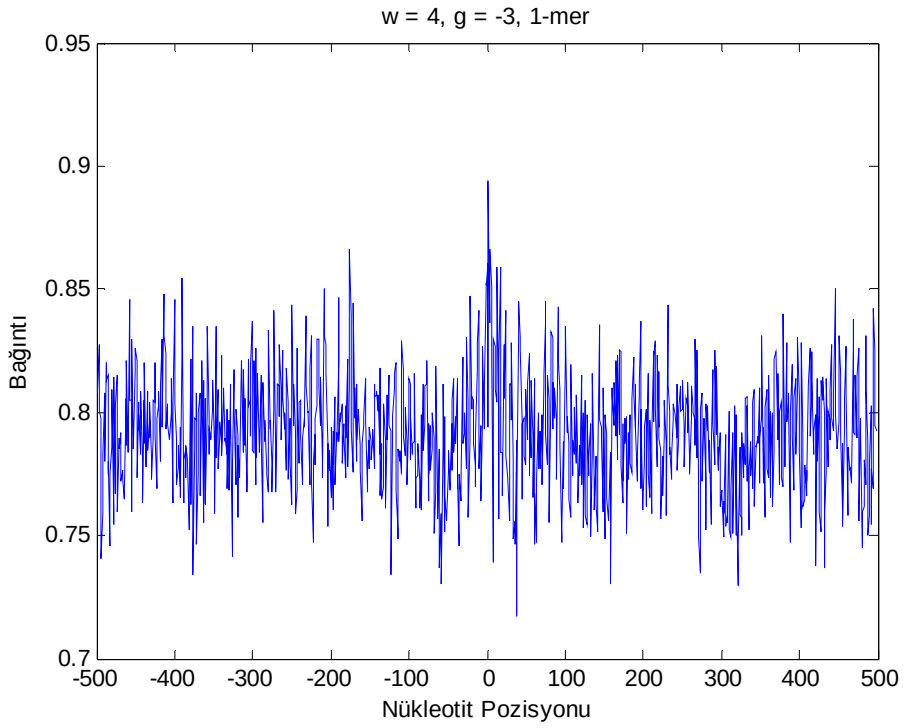
Bu denemeler esnasında probleme uygun seçilmesi gereken üç parametre vardır:

i. Pencere genişliği (w)
ii. Pencereler arası boşluk (g)
iii. Kullanılacak dizilim özelliği

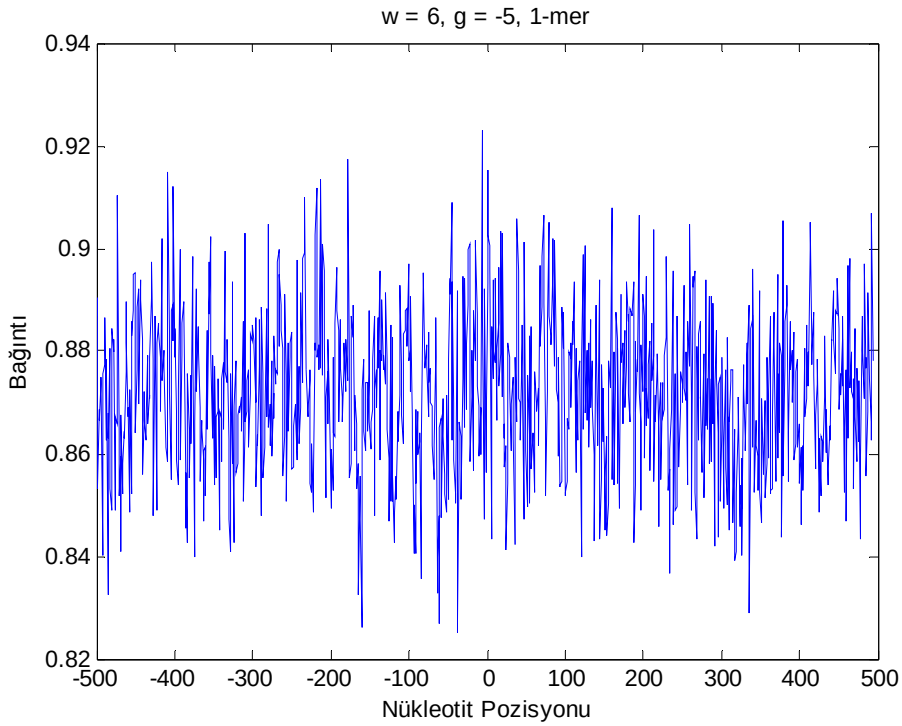
Dizilim özelliği olarak öncelikle nükleotit frekansları (1-mer) kullanılmıştır. Bölüm 2.2'de de değinildiği üzere Schroder'in çalışmasında kullanılan HIV türevi virüsün insan konağına yerleşim noktasında simetrik/palindromik bir davranış gözlemlenmektedir (Holman, 2005 – Wu, 2005). Bu davranışın, yerleşim noktasını çevreleyen küçük bir alanda gerçekleştiği bilinmektedir. Bu nedenle pencere genişliği parametresi (w) sırasıyla w = [2, 4, 6, 8] gibi değerlerle denenmiştir. Bunun yanı sıra, yerleşim noktasında rapor edilen bulgunun simetrik/palindromik bir karakteristik gösterdiği bilindiğinden pencereler arası boşluk parametresi olan g negatif olarak (g = [-1, -3, -5, -7]) seçilmiştir. Bir başka deyişle pencerelerin kısmi olarak üst üste binmesi hedeflenmiştir. Burada amaç değişen pencere boyutları için her seferinde pencere uçlarından birer nükleotiti analizin dışında tutmak ve özellikle dışarıda tutulan bu nükleotitler için iyi bir KBA eğitimi&testi yapılıp yapılamadığını görmektir. Şekil 4.1 - 4.4'de bu denemelerin x = [-500 : +500] aralığı için verdiği ortalama (5×2 çapraz geçerleme ortalaması) test bağıntısı değerleri görülmektedir.



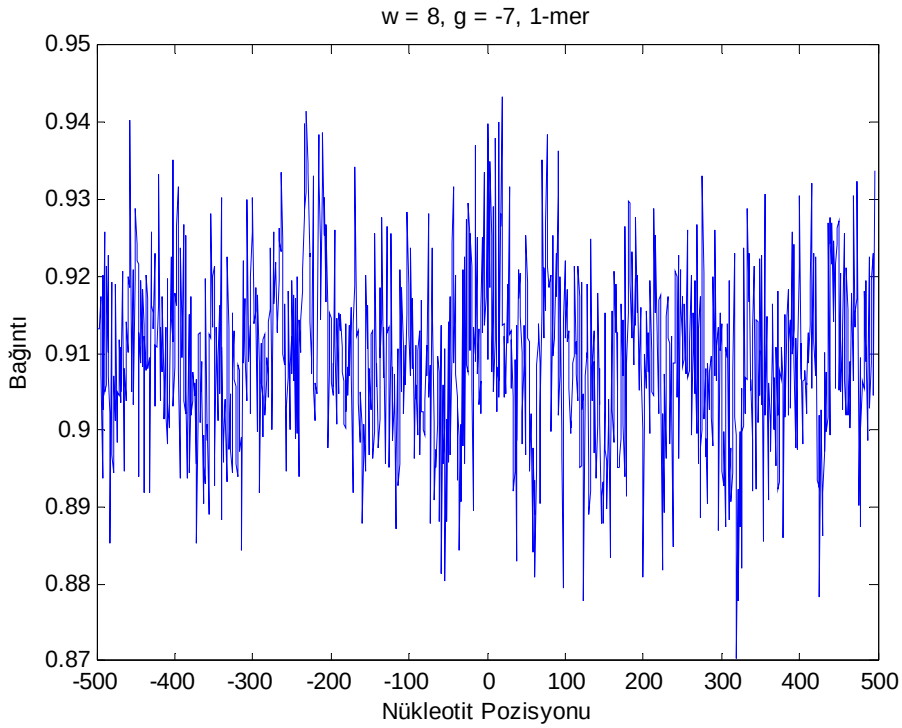
Şekil 4.1: Nükleotit frekansları ile elde edilen test bağıntıları (w = 2, g = -1)



Şekil 4.2: Nükleotit frekansları ile elde edilen test bağıntıları (w = 4, g = -3)



Şekil 4.3: Nükleotit frekansları ile elde edilen test bağıntıları (w = 6, g = -5)



Şekil 4.4: Nükleotit frekansları ile elde edilen test bağıntıları (w = 8, g = -7)

Şekil 4.1 ve 4.2'den görüleceği üzere $w = 2$ ve $w = 4$ gibi küçük pencere genişlikleri için $x = +1$ ve $x = +3$ pozisyonlarında sırasıyla 0.76 ve 0.89 gibi (Her çalıştırmada rasgele ve farklı bir eğitim-test kümesi çifti oluşturulduğundan değerlerde küçük değişiklikler olabilmektedir.) yüksek test bağıntıları elde edilmektedir. x pozisyonu, palindromik örüntünün merkez noktasını göstermektedir. Buna göre sol pencere $\left\lfloor x - w + 1 - \frac{g}{2} \right\rfloor$ ve $\left\lfloor x - \frac{g}{2} \right\rfloor$ aralığını kapsarken, sağ pencere $\left\lfloor x + 1 + \frac{g}{2} \right\rfloor$ ve $\left\lfloor x + w + \frac{g}{2} \right\rfloor$ aralığını kapsamaktadır.

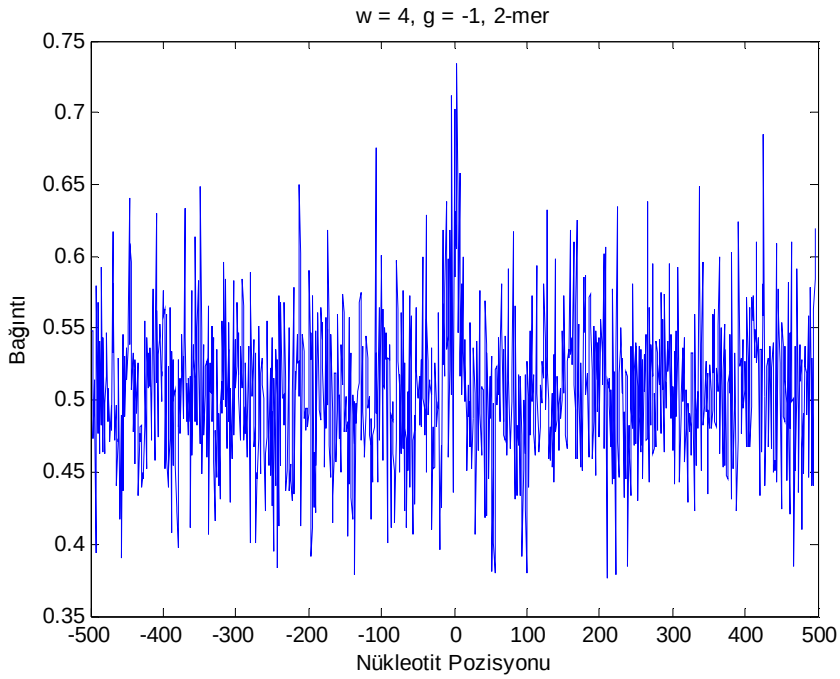
Pencere genişliği arttıkça, test bağıntısı ortalamasının arttığı görülmektedir. Bunun nedeni pencerelerin kapsadığı ortak alanın büyümesi ve dışarıda kalan tek nükleotitlik pozisyonun toplam fonksiyona olan etkisinin azalmasıdır. Bunun yanı sıra, Şekil 4.3 ve 4.4'den de görüleceği üzere yerleşim noktasındaki yüksek bağıntı değerleri belirginliğini kaybetmektedir. Bu durum, palindromik davranışın nükleotit genişliği açısından küçük bir alanda belirgin olduğunu doğrulamaktadır.

Yerleşim noktasında elde edilen yüksek bağıntı değerlerinin rasgele olmadığını test edebilmek için z-test istatistik testi (Sprinthall, 2003) kullanılmıştır. Tablo 4.1'de $x = [-5 : +6]$ aralığı için değişen w ve g parametreleri ile yapılan testlerde elde edilen bağıntı değerleri ve bu değerlere karşılık gelen anlamlılık değerleri (p-value) görülmektedir. İstatistiksel anlamlılık açısından $p < 0.05$ değerine sahip bağıntıların rasgele olmadığı değerlendirilmektedir. Buna göre, farklı w parametreleri ile yapılan testlerde $x = +5$ pozisyonunu merkez olarak kabul eden pencere çiftlerinin her seferinde anlamlı ve yüksek bağıntı değerleri ürettiği görülmektedir.

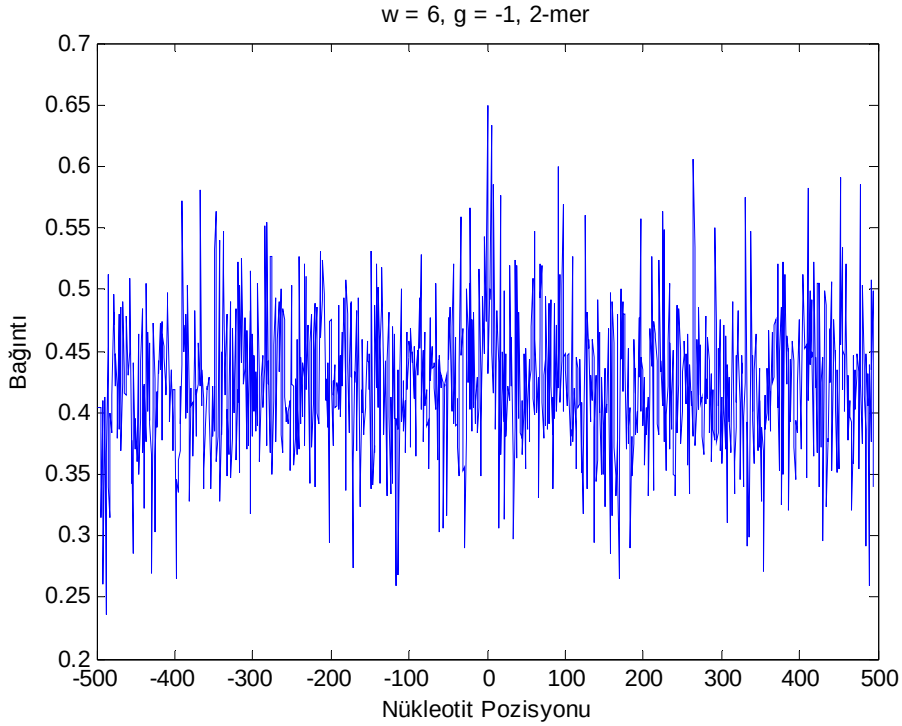
Tablo 4.1: Yerleşim noktası çevresinde görülen bağıntı ve z-test anlamlılık değerleri

		***x* pozisyonu**										
w	**g**	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6
2	-1	0.56 (0.99)	0.56 (0.94)	0.50 (0.17)	0.47 (0.05)	0.56 (0.97)	0.76 (0.00)	0.61 (0.25)	0.56 (0.96)	0.54 (0.70)	0.75 (0.00)	0.62 (0.19)
4	-3	0.83 (0.14)	0.79 (0.95)	0.81 (0.44)	0.85 (0.01)	0.85 (0.01)	0.86 (0.00)	0.79 (0.93)	0.89 (0.00)	0.84 (0.06)	0.87 (0.00)	0.85 (0.01)
6	-5	0.86 (0.45)	0.87 (0.73)	0.92 (0.00)	0.89 (0.35)	0.85 (0.12)	0.89 (0.19)	0.86 (0.33)	0.87 (0.77)	0.89 (0.17)	0.92 (0.01)	0.90 (0.05)
8	-7	0.91 (0.76)	0.93 (0.03)	0.91 (0.71)	0.92 (0.47)	0.92 (0.46)	0.94 (0.01)	0.91 (0.99)	0.94 (0.01)	0.92 (0.41)	0.93 (0.02)	0.93 (0.16)

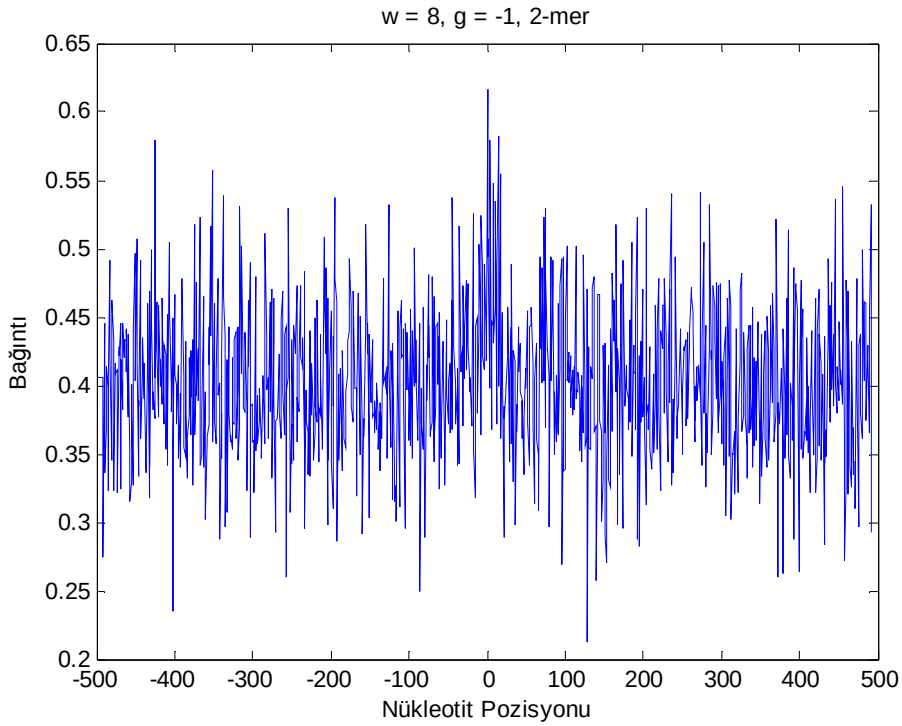
KBA testlerinde dizilim özelliği olarak dimer (2-mer) frekansları kullanıldığında simetrik/palindromik davranışın en iyi w = [4 6 8] ve g = -1 parametreleri ile gözlemlendiği tespit edilmiştir. Ancak burada nükleotit frekansları ile yapılan testlerden farklı olarak g değeri sürekli -1 olarak alındığından artan pencere genişlikleri için ortalama test bağıntıları düşmektedir. Bunun nedeni KBA'nın w arttıkça birbirinden farklılaşan pencere içerikleri arasında eğitimde iyi bir bağıntı fonksiyonu üretememesi ve dolayısıyla öğrenilen ağırlık katsayılarının test kümesinde daha düşük değerler üretmesidir. Bu durum Şekil 4.5 – 4.7'de görülmektedir.



Şekil 4.5: Dimer frekansları ile elde edilen test bağıntıları (w = 4, g = -1)



Şekil 4.6: Dimer frekansları ile elde edilen test bağıntıları (w = 6, g = -1)



Şekil 4.7: Dimer frekansları ile elde edilen test bağıntıları (w = 8, g = -1)

Moment değişkenleri ile yapılan denemelerde söz konusu aralık için istatistiksel olarak anlamlı bir hedef bölge bulunamamıştır. Bunun nedeni, moment değişkenlerinin bu veri kümesindeki örüntünün aksine daha uzun dizilimlerle kullanılmaya uygun olmasıdır.

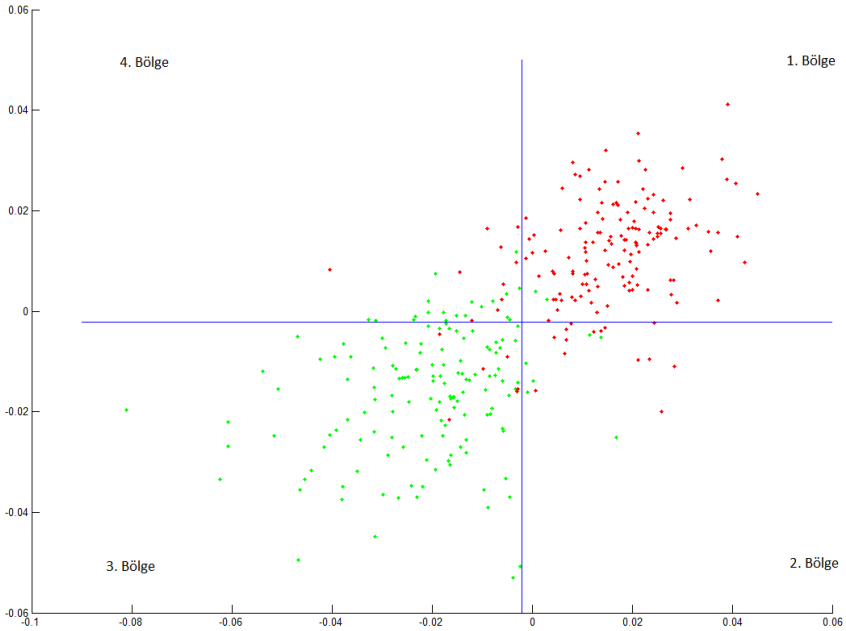
4.1.1 KBA İle Danışmasız Öğrenme

Önceki bölümde bahsedilen palindromik davranış, genomla +/+ (298 adet) ve -/+ (331 adet) yönünde eşleşen dizilimler beraber kullanıldığında da (toplam 629 dizilim) görülmüştür. Dizilim özelliği olarak dimer frekanslarının kullanıldığı bu denemelerin sonunda, yapılan istatistik testine göre diğerleri ile karşılaştırıldığında istatistiksel olarak belirgin bir şekilde öne çıkan bir ilişki görülmemiştir. Ancak bu testler sırasında, KBA yönteminin iki görü arasındaki bağıntıyı en büyük kılacak izdüşüm vektörlerini bulmanın yanı sıra örneklerin sınıf bilgisini (genomda -/+ ya da +/+ yönünde yerleşim) kullanmadan (danışmasız) öbeklenmesinde de kullanılabileceği görülmüştür.

Buna göre sol ve sağ görülerden elde edilen izdüşüm değerlerinin (fonksiyon çıktılarının) ortalaması hesaplanmış ve dizilimler sırasıyla:

i. Hem sol hem de sağ görüde izdüşüm ortalamalarından büyük çıktılara sahip dizilimler (1. Bölge),
ii. Sol görüdeki izdüşümlerin ortalamasından büyük ancak sağ görüdeki izdüşümlerin ortalamasından küçük çıktılara sahip dizilimler (2. Bölge),
iii. Sağ görüdeki izdüşümlerin ortalamasından büyük ancak sol görüdeki izdüşümlerin ortalamasından küçük çıktılara sahip dizilimler (3. Bölge),
iv. Hem sol hem de sağ görüde izdüşüm ortalamalarından küçük çıktılara sahip dizilimler (4. Bölge),

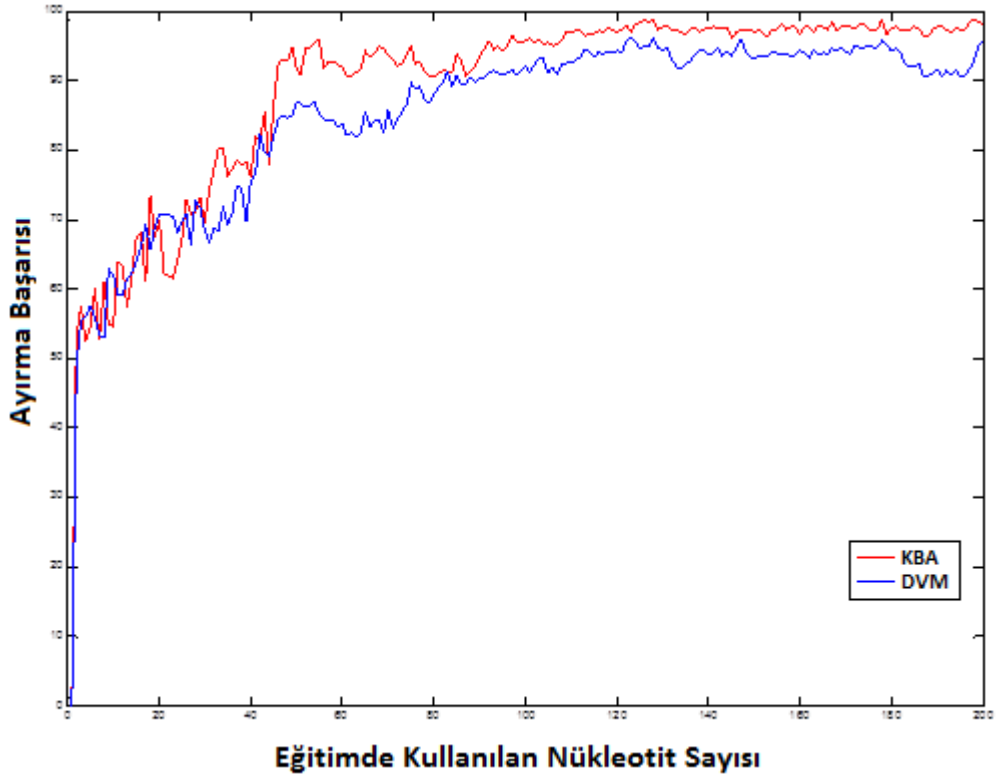
şeklinde dörde ayrılmıştır. w = 150 genişlikli iki komşu görü arasında yapılan bir denemede Şekil 4.8'deki test dizilimleri dağılımı gözlemlenmiştir. Bu denemede 629 adet dizilimin yarısı eğitim, kalan yarısı da test kümesi olarak kullanılmıştır.



Şekil 4.8: Dizilimlerin izdüşümlerinin ortalamaya göre dört bölgeye dağılımı

Şekil 4.8'de farklı renkler ile gösterilen örnekler (dizilimler) gerçekte farklı sınıflara (genomda -/+ yönünde yerleşim gösteren dizilimler ve +/+ yönünde yerleşim gösteren dizilimler) aittirler. Buna göre 1. ve 3. bölgelere düşen örnekler (265 adet) birbirlerinden az hata ile ayrılabilirken (%96.98 ayrım başarısı), 2. ve 4. bölgelere düşen örnekler için aynı ayrım başarısı elde edilememiştir. Bunun nedeni, 2. ve 4. bölgelere düşen örneklerden elde edilen dimer frekansları için KBA yönteminin iyi ağırlık vektörleri üretememesidir.

Bir sonraki aşamada KBA yöntemi ile danışmasız olarak yapılan ayrımın başarısı danışmalı bir sınıflandırıcı olan "Destek Vektör Makineleri" yöntemi (Bkz: Bölüm 3.2) ile karşılaştırılmıştır. Bu amaçla, $w = [2 : 200]$ aralığında artan pencere boyutları için sadece 1. ve 3. bölgelere düşen fragmanların KBA yöntemi ile ayrılma başarısı, pencere çiftini kapsayan alanda ($2w = [4 : 400]$) eğitilen ve "Dairesel Tabanlı Fonksiyon" (RBF) çekirdek yöntemi kullanan DVM sınıflandırıcısı ile karşılaştırılmıştır. Bu denemelerin her birinde önceden olduğu gibi 629 adet dizilimin rasgele seçilen yarısı eğitimde, kalan diğer yarısı da testte kullanılmıştır. Buna göre artan pencere genişliği için danışmalı ve danışmasız öğrenme yöntemlerinin örnekleri ayırma başarısı Şekil 4.9'da görüldüğü gibidir. Burada, kırmızı renkli eğri KBA, mavi renkli eğri DVM yöntemi ile alınan sonuçları göstermektedir.

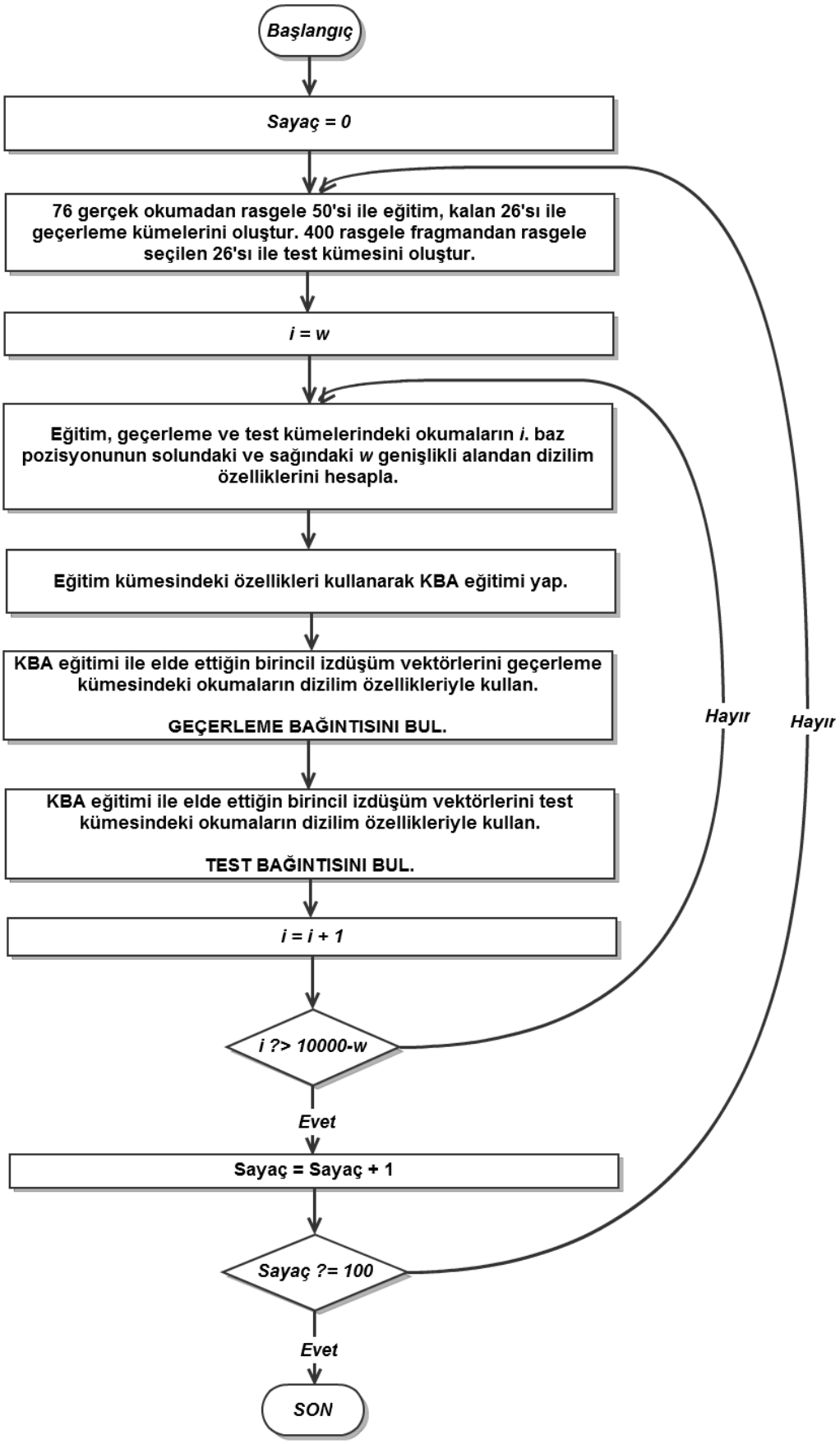


Şekil 4.9: KBA ve DVM'nin aynı nükleotit aralığı için ayırma başarısı

Şekil 4.9'da da görüleceği üzere KBA yöntemi iki sınıflı bir sınıflandırma probleminde DVM yönteminden daha iyi ayrım başarısı gösterebilmektedir. Bu da yöntemin genomdaki ilişki çıkarımı çalışmalarının yanında öbekleme amacıyla da kullanılabileceğini göstermektedir (Gumus, 2013b).

4.2 VİRÜS YERLEŞİM BÖLGELERİNDEKİ DİĞER ÖZELLİKLER

Bölüm 2.2'de, İstanbul Üniversitesi Deneysel Tıp Araştırmaları Enstitüsü'nde (DETAE) Schroder'in çalışmasına benzer bir retrovirüs çalışmasının yapıldığı ve bu çalışmada kullanılan virüsün, insan genomunun 76 farklı noktasına yerleştiğine değinilmiştir. Önerilen örüntü tarama aracının farklı dizilim özellikleri ve parametrelerle kullanıldığı analizlerde söz konusu okumalara ait yerleşim noktalarında Holman ve Wu'nun (Holman, 2005 – Wu, 2005) rapor ettiği palindromik davranış gözlemlenmemiştir.

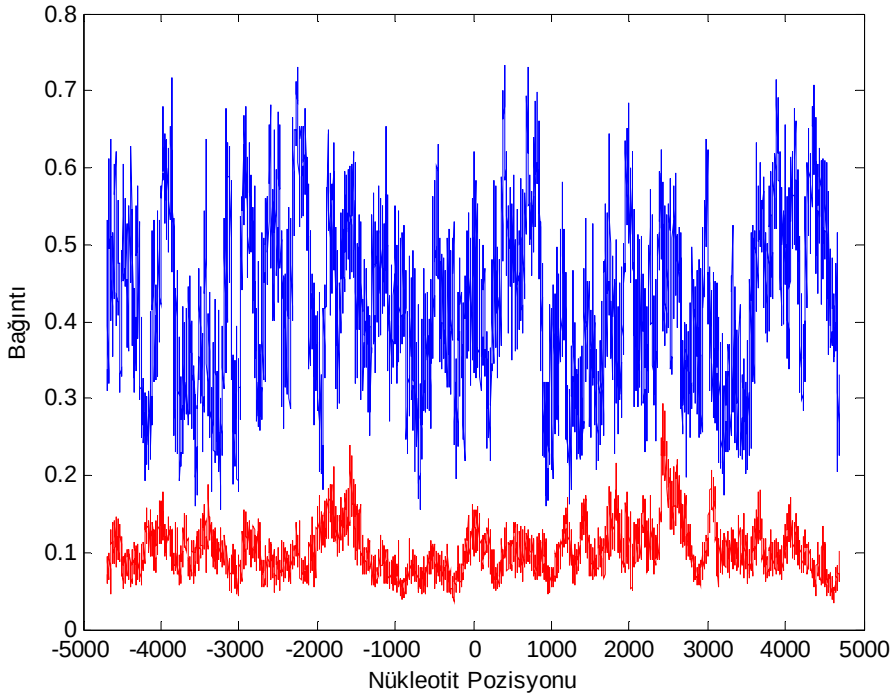


Şekil 4.10: Eğitim – Geçerleme – Test İşleyişi

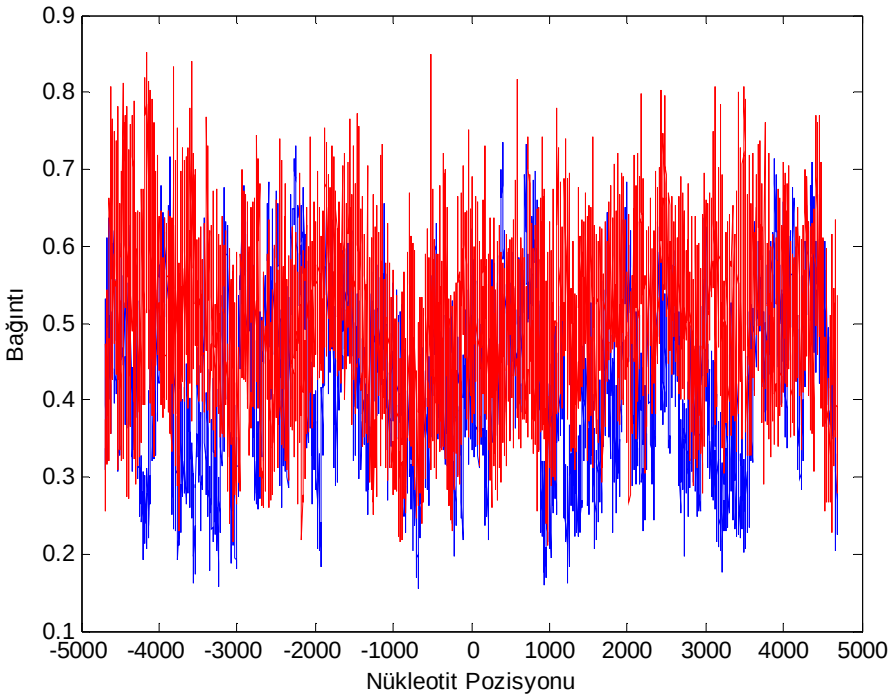
Bir sonraki aşamada, viral yerleşimin gerçekleştiği bölgelerde görüldüğü halde genomun rasgele seçilen bölgelerinde görülmeyen, bir diğer deyişle sadece viral yerleşimin görüldüğü bölgeye özgü başka bir karakteristiğin var olup olmadığı araştırılmıştır. Bu amaçla 76 okumadan tespit edilen virüs yerleşim pozisyonları $x = 0$ noktası olarak kabul edilmiş ve her okuma için bu noktanın solundan ve sağından 5000'er nükleotitlik, toplam 10000 nükleotit genişliğinde bir alan alınmıştır. Bu 76 okuma, eğitim ve geçerleme işlemleri için kullanılırken buna ek olarak virüsün insan genomunda yerleşmediği bilinen rasgele bölgelerden de yine 10000 nükleotit uzunluklu 400 adet rasgele okuma alınmıştır. Bu okumalar da test işlemi için kullanılmıştır.

Oluşturulan 100 denemelik analizin her basamağında Şekil 4.10'da gösterildiği gibi her seferinde 76 okumanın 50'sinden eğitim, kalan 26'sından geçerleme ve 400 rasgele fragmanın rasgele seçilen 26 tanesinden test kümeleri oluşturulmuştur. i, o sırada işlenmekte olan nükleotit pozisyonunu ($i \in$ [w : 10000 - w]), w da sol ve sağ görüler için pencere genişliğini göstermektedir. Her i pozisyonu için KBA yöntemi kullanılarak geçerleme ve test bağıntıları hesaplanmaktadır. Bu metodoloji, dizilim özelliği olarak dimer frekansları, pencereler arası boşluk için $g = 0$ ve bir dizi pencere genişliği ($w \in [25 : 400]$) kullanılarak uygulanmış ve ortalama geçerleme bağıntısının ortalama test bağıntısından sürekli olarak büyük olduğu görülmüştür (Bkz: Şekil 4.11, $w = 300$). Bu da virüs yerleşim noktasına göre hizalanan okumaların her nükleotit pozisyonu için sol ve sağ görüler arasındaki ilişkinin sadece virüsün yerleşim gösterdiği bölgelere özgü olduğunu göstermektedir.

Analizin 100 denemelik olmasının amacı, tüm denemeler sonucu elde edilen ortalama geçerleme bağıntısının 100 adet test bağıntısının hepsinden yüksek olduğu bir i merkezli bölgenin olup olmadığını belirlemektir. Eğitim ve geçerleme kümeleri, virüsün yerleşmeyi tercih ettiği noktalardan alınan okumalarla, test kümesi de rasgele okumalardan oluşturulduğu için böyle bir bölgenin bulunması virüsün yerleşim karakteristiğini anlamaya yardımcı olabilir. Şekil 4.12'de bu özelliğe uyan bölgeler kırmızı eğrinin üstünde kalan mavi alanlardır.



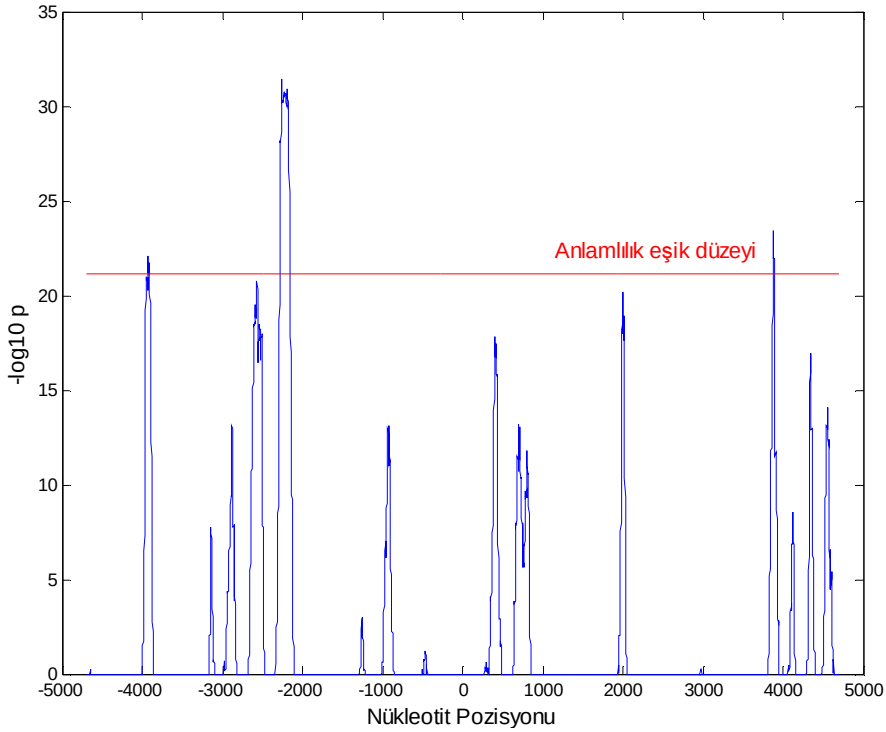
Şekil 4.11: Ortalama geçerleme (Mavi) ve test (Kırmızı) bağıntıları (w = 300, g = 0)



Şekil 4.12: Ortalama geçerleme (Mavi) ve en yüksek test (Kırmızı) Bağıntıları (w = 300, g = 0)

Şekil 4.12'de görüldüğü üzere, ortalama geçerleme bağıntısının 100 denemeden elde edilen en yüksek test bağıntısını dahi geçebildiği kısıtlı sayıda bölge vardır.

Geçerleme bağıntısının, en yüksek test bağıntısını belli bir baz pozisyonu boyunca geçebildiği bu bölgelerdeki bağıntıların farklı medyanlara sahip dağılımlardan olup olmadığı Bölüm 3.5'de değinilen Mann-Whitney sıralama toplamı testi ile incelenmiştir. Bu test, *X* ve *Y* gibi iki dağılımın aynı medyana sahip (bir diğer deyişle benzer) olduğu hipotezinin geçerliliğini sınar. Buna göre her nükleotit pozisyonu, x_n, için $[x_n - 50 : x_n + 50]$ aralığındaki ortalama geçerleme bağıntıları, yine aynı aralıktaki en büyük test bağıntıları ile medyan testine tabi tutulmuş ve bu aralıktaki iki dağılımın aynı medyana sahip olma hipotezinin ne kadarlık bir anlamlılık (*p*) değeri ile reddedilebildiği bulunmuştur. Bu değerlerin $-\log_{10} p$ türünden ifadesi Şekil 4.13'de görülmektedir.



Şekil 4.13: Ortalama geçerleme bağıntısının en büyük test bağıntısını geçtiği anlamlı bölgeler

Burada görüleceği üzere, viral yerleşim noktasını çevreleyen alanda uzun nükleotit pozisyonları boyunca ortalama geçerleme bağıntısının, en büyük test bağıntısından daha büyük ve hipotezi çürütecek kadar farklı bir medyana sahip olduğu pek çok bölge vardır. Bu bölgelerin hangilerinin dikkate alınacak düzeyde olduğunun tespiti için bir anlamlılık eşik düzeyi gerekmektedir. Bu eşik değerini bulmak için her x_n pozisyonunu çevreleyen $[x_n - 50 : x_n + 50]$ aralığındaki 100 adet test bağıntısı kullanılmıştır. Yapılan 100 iterasyonluk bir analizde her seferinde test bağıntılarından biri geçerleme bağıntısı gibi, kalan 99 bağıntının her nükleotit pozisyonu için en büyük değerleri de bir tek test

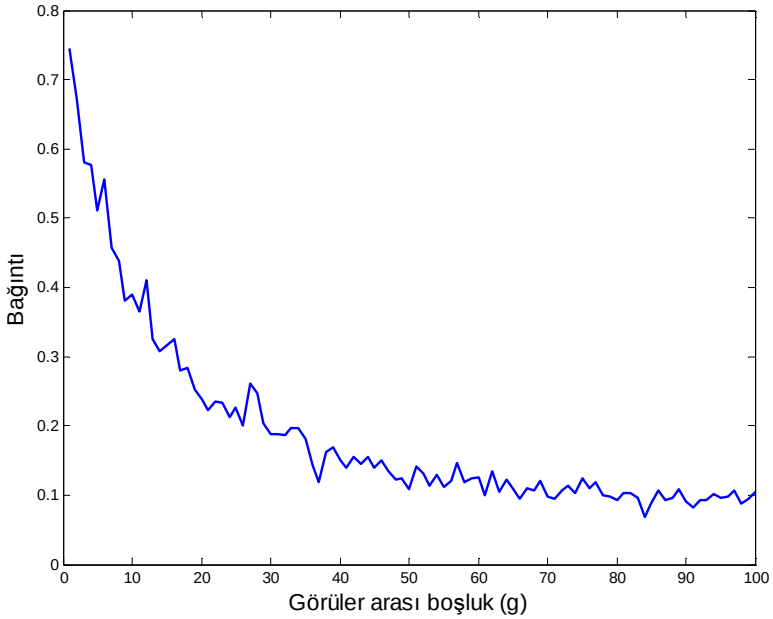
bağıntısı gibi kullanılmıştır. Bu iki yeni bağıntının medyan testi sonucunda tüm nükleotit pozisyonları değerlendirildiğinde bulunan en küçük p değeri 6.94×10^{-22} ve $-\log_{10} p$ değeri 21.15 olmuştur. Bu eşik değeri Şekil 4.13'de kırmızı çizgi ile gösterilmiştir. Bu eşiği aşabilen bölgelerdeki geçerleme ve test bağıntıları arası medyan farkının en belirgin düzeyde olduğu değerlendirilmiştir. Bu bölgeler içinden x_n = [-2283 : -2153] aralığındaki alanda bu fark 131 nükleotit boyunca sürdüğünden söz konusu alan virüsün yerleşim tercihini incelemek için ilgi çekici bulunmuştur. w = 300 parametresi ile çalıştırılan örüntü tarama aracının bu bölge için ürettiği dimer ağırlık katsayıları (KBA sol ve sağ görü izdüşüm vektörleri) kullanılarak görüler arasında bulunan bağıntının rasgele bir şekilde elde edilemeyeceği görülmüştür.

4.3 GENOM KAPSAMINDA İLİŞKİ ÇIKARIMI İLE HASTALIK ETMENİ MUTASYONLARIN TESPİTİ

Behçet hastalığına neden olan faktörlerin tespiti amacıyla yapılmış bir çalışmada (Remmers, 2010), Türkiye'deki 1215 Behçet hastası ve 1278 sağlıklı bireyden alınan örneklerle genom haritaları düzenlenmiş ve 311459 adet SNP (Single Nucleotide Polymorphism) üzerinden genom kapsamında ilişki çıkarımı çalışması (GWAS) gerçekleştirilmiştir. SNP bazında yapılan ki-kare anlamlılık testleri ile 6 numaralı kromozomda bulunan MHC (Major Histocompatibility Complex) bölgesinin (29691117 – 33253403 numaralı nükleotitler ve aynı zamanda 4305 – 6718 nolu SNP'ler arasında kalan alan) hastalıkla ilişkili SNP'ler bulundurduğu gözlemlenmiştir. Bunu takiben bu çalışma kapsamında, aynı veri kümesi için, SNP'lerin tek başına değerlendirilmesinden ziyade grup olarak ele alındıklarında hastalığa olan etkileri incelenmiştir. Bu noktada "Bağlantı Eşitsizliği" kavramından faydalanılmıştır.

Bağlantı eşitsizliği, önceden de tanımlandığı gibi, bir türe ait bireylerin genomlarındaki bir grup SNP'nin beraber değişim göstermesi ya da diğer bir deyişle birbirinden bağımsız olmaması durumudur. Bağlantı eşitsizliği, genellikle fiziksel konumları itibariyle birbirine yakın SNP'ler arasında yüksek bir değere sahipken bazı durumlarda konum olarak birbirine uzak olan SNP'ler arasında da yüksek bir değere sahip olabilmektedir. İki SNP arasındaki bu ilişkiyi ölçmek için kullanılan ölçütlere Bölüm 2.3.1.1'de değinilmiştir. Ancak bahsedilen ölçütler sadece iki SNP arasındaki ilişkiyi

ölçmek için kullanılmakta olup iki farklı SNP grubu arasındaki ilişkiyi ölçmek için kullanılamaz. Bu noktada, birden fazla değişken içeren görüler arasındaki olası en yüksek bağıntıyı sağlayan izdüşüm vektörlerini (ağırlık katsayılarını) bulmayı hedefleyen Kanonik Bağıntı Analizi yönteminden faydalanılabilir. Şekil 4.14'de KBA yönteminin, her biri $w = 2$ adet SNP içeren iki SNP görü grubu için, artan görüler arası boşluk (g) parametresi ile çalıştırılması sonucu elde edilen ortalama eğitim bağıntısı görülmektedir. Her g değeri için bağıntı değeri, 6. kromozomun rasgele seçilen 100 farklı bölgesinde KBA eğitimi yapılarak elde edilen eğitim bağıntılarının ortalaması alınarak hesaplanmıştır.



Şekil 4.14: SNP görü grupları arası mesafe – KBA eğitim bağıntısı ilişkisi

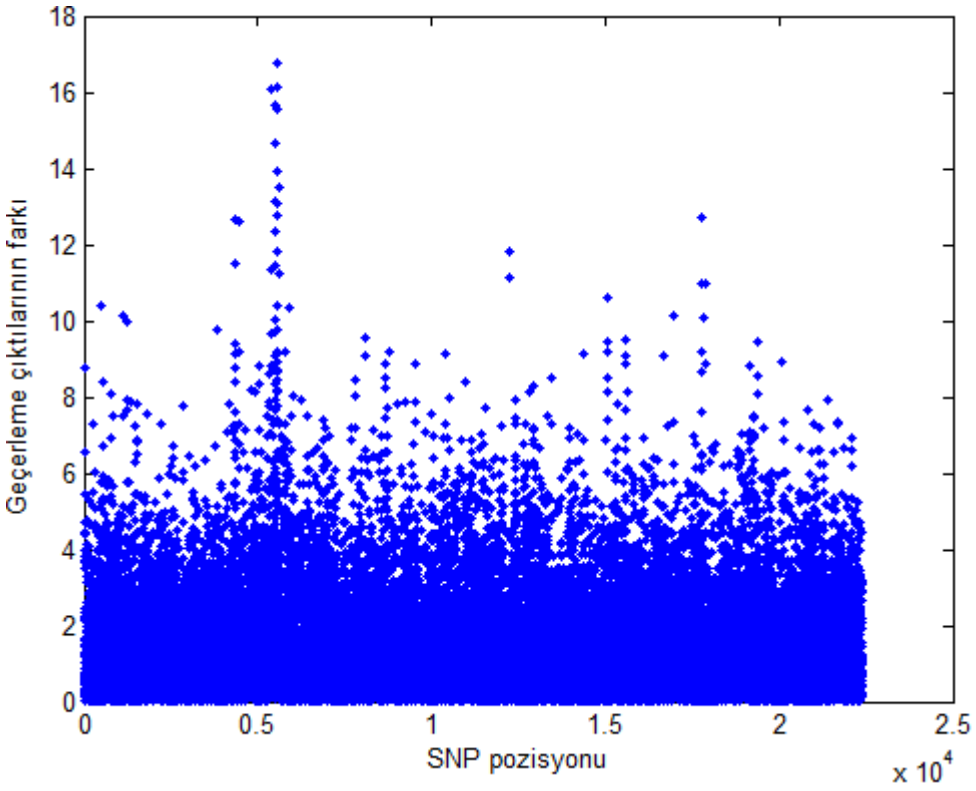
Şekil 4.14'de de görüldüğü gibi, SNP grupları arası fiziksel mesafe arttıkça KBA eğitiminde görüler arasında sağlanabilen en yüksek eğitim bağıntısında keskin bir düşüş olmaktadır. Bunun nedeni, birbirlerine uzak SNP'ler arasındaki Bağlantı Eşitsizliği'nin düşük seviyede olmasıdır.

Behçet hastalığı gibi genetik etmenli hastalıklar, genomdaki bağlantı eşitsizliğini bozacak düzeyde mutasyonlara sahip olabilmektedir. Bu bölümde, mutasyonların tek başlarına ya da bir grup halinde bağlantı eşitsizliği üzerine olan etkilerinin genomdaki ne kadarlık bir alanda sürdüğü KBA yöntemi ile tespit edilmeye çalışılmıştır. Buradaki temel amaç, sadece sağlıklı bireylerden oluşan bir eğitim kümesinde öğrenilen KBA

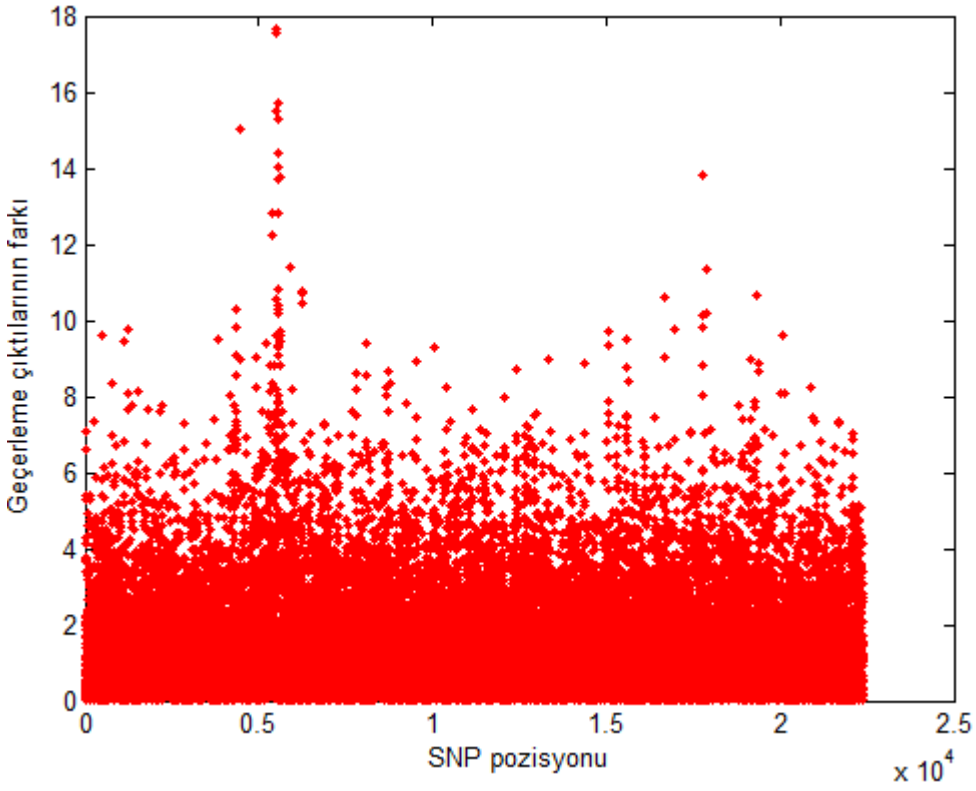
ağırlık katsayıları ile elde edilen eğitim kümesi çıktılarının iç çarpımının, aynı katsayılar sadece hastalardan oluşan bir geçerleme kümesinde kullanıldığında elde edilemeyeceği alanları (komşu SNP'lerden oluşan grupları) bulmaktır. Bu durum, hastalık etmeni mutasyonların sağlıklı bireylerde görülen bağlantı eşitsizliğini bozmasından ve dolayısıyla sağlıklı bireylerde yüksek oranda bağlantı eşitsizliği sağlayan KBA ağırlık katsayılarının hasta bireylerden alınan veride iyi çalışmamasından ileri gelir. KBA yöntemi bunun aksini de destekleyebilir. Eğer hastalığa neden olan mutasyonlar kendi aralarında yüksek bir bağlantı eşitsizliğine sahipse, hasta bireylerden oluşan bir eğitim kümesinden öğrenilen KBA ağırlık katsayıları, aynı bölge için sağlıklı bireylerin veri kümesinde iyi çalışmayacaktır.

Denemelerde, Behçet veri kümesindeki 1200 hasta ve 1200 sağlıklı bireyin 6. kromozoma ait genomik verisinin 800'er adedi KBA eğitim ve geçerleme kümesi, kalan 400'er adedi de KBA test kümesi örnekleri olarak ayrılmıştır. Sağlıklı ve hasta bireylerden alınan 800'er adet örnek, 10 kat çapraz geçerleme tekniği ile ayrı ayrı KBA eğitimine tabi tutulmuştur. Her iki grup için 10 geçerleme katından elde edilen görü çıktılarının iç çarpımlarının ortalaması alınarak nihai sağlıklı ve hasta geçerleme çıktıları bulunmuştur. Zira, sağlıklı bireylerin kendi içlerinde dahi 10 kat çapraz geçerlemenin her katında farklı geçerleme çıktıları elde edilebildiği görülmüştür. Bu nedenle 10 denemeden ortalama sağlıklı ve hasta geçerleme çıktıları hesaplanmıştır.

KBA yönteminin amacı, *w* genişlikli iki görü arasında en yüksek bağıntıyı sağlayacak ağırlık katsayılarını bulmak olduğundan sağlıklı ve hasta bireylerden ayrı ayrı elde edilen ortalama geçerleme çıktılarını karşılaştırmak anlamsız olacaktır. Önemli olan 10 kat çapraz geçerlemenin her katında, bir gruptaki (sağlıklı veya hasta) eğitim kümesinden öğrenilen KBA ağırlık katsayılarını kullanarak diğer gruptaki (hasta veya sağlıklı) geçerleme kümesinden bir gruplar arası geçerleme çıktısı elde etmek, diğer bir deyişle sağlıklılardan (veya hastalardan) oluşan bir eğitim gurubundaki bağlantı eşitsizliğini en üst düzeyde belirleyen ağırlık katsayılarının hastalarda (veya sağlıklılarda) da aynı düzeyde bağlantı eşitsizliğine neden olup olmadığını bulmaktır. Bir gruptaki grup içi ortalama geçerleme çıktısının, gruplar arası ortalama geçerleme çıktısından farklı olduğu noktalar, hastalar ve sağlıklıların farklılaştığı yani Behçet hastalığının olası kaynağı olan bölgelerdir.



Şekil 4.15: Sağlıklı-sağlıklı geçerleme çıktısının sağlıklı-hasta geçerleme çıktısından farkı



Şekil 4.16: Hasta-hasta geçerleme çıktısının hasta-sağlıklı geçerleme çıktısından farkı

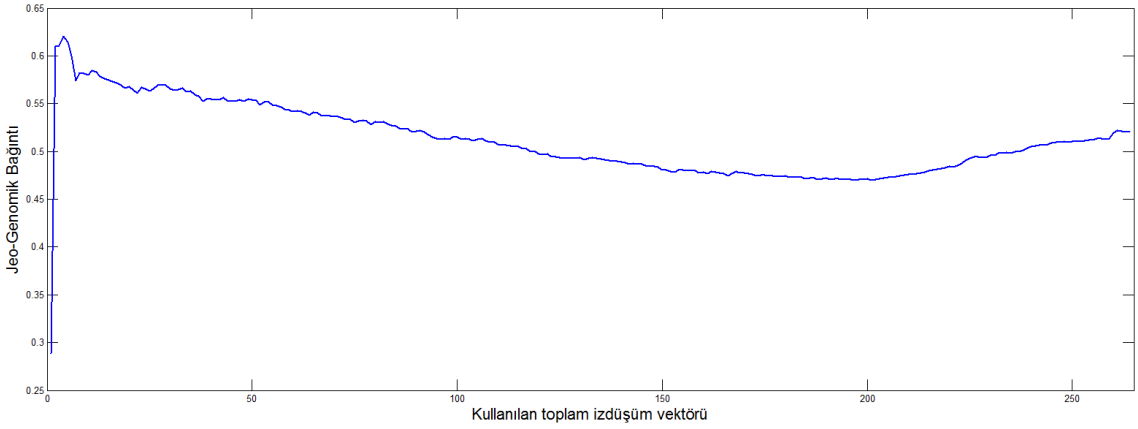
Şekil 4.15 ve 4.16, $w = 4$, $g = 0$ parametreleriyle yapılan KBA denemeleri sonucunda elde edilen geçerleme çıktıları arasındaki farkı göstermektedir. Her iki açıdan bakıldığında da farkın en yüksek olduğu bölge 6. kromozomun kabaca 5500. SNP'inin çevresidir ki bu da Remmers'ın çalışmasında (Remmers, 2010) da değinildiği gibi MHC (Major Histocompatibility Complex) bölgesidir. Aynı çalışmada 6. kromozomdaki her SNP pozisyonu için (toplamda 22393 adet SNP pozisyonu) hasta ve sağlıklı tüm bireyleri içeren bir ki-kare anlamlılık testi uygulanmış ve bu bölgeden bazı SNP'ler hastalığın olası etmenleri olarak işaretlenmiştir. Ancak ki-kare gibi istatistik testleri bir dağılımı sadece kendi içerisinde değerlendirdiğinden, diğer bir deyişle bir dağılımın komşuluğundaki dağılımlarla beraber olan toplam etkisini incelemediğinden yetersiz kalabilmektedir. Bu noktada KBA yöntemi, ki-kare testi ile tespit edilen tekil SNP'lere ek olarak, iki küme arasındaki farklı bağlantı eşitsizliği değerlerine neden olan SNP gruplarını bulabilir ve her bir SNP'nin toplam etkideki payını ağırlıklandırabilir. Söz konusu SNP'ler, tek başlarına belirgin bir fark yaratmasa da beraber incelendiklerinde hastalığa etkilerinin derecesi daha iyi anlaşılabilir.

Tespit edilen SNP gruplarından, birarada incelendiklerinde ilgi çekici olanlar için bir kabulde bulunulmuştur. Buna göre, hastalığa neden olabilecek bir SNP grubunun içindeki SNP'lerden en az $w / 2$ adedinin (bir tek görü için) KBA ağırlık katsayısının mutlak değerinin en az 0.5 olması gerektiği kabul edilmiştir. Bu değer, bir SNP'in görüler arası bağıntıya etki edebilmesi için yeterince büyüktür. $w = 4$, $g = 0$ parametreleriyle yapılan denemeler sonucunda bu kriteri sağlayan ve geçerleme çıktılarının farkı açısından t istatistik testindeki anlamlılık değerleri sıralandığında normal dağılımdan sapan ilk %1'lik dilime giren 75 adet $2\times w$ genişliğinde SNP bölgesi bulunmuştur. Bu bölgelerden istatistiksel anlamlılık açısından ilk sırada olanı $-\log_{10} p$ değeri 3.6922 olan ve [5567 : 5574] SNP aralığını kapsayan bölge olarak değerlendirilmiştir.

4.4 JEO – GENOMİK İLİŞKİNİN ANALİZİ

Bölüm 2.3.2'de Temel Bileşen Analizi yöntemiyle büyük boyutlu veri kümelerinde boyut indirgeme ve dolayısıyla verinin görselleştirilmesi işleminin yapılabileceğine değinilmiştir. Bu amaçla kullanılan HGDP SNP veri kümesi Bölüm 3.8'de tanıtılmış ve

çalışmaya konu olan etnik gruplar için orijinal veri uzayındaki jeo-genomik bağıntının %52 düzeyinde olduğu gözlemlenmiştir. Bir sonraki aşamada Bölüm 3.1.1'de değinilen boyut hilesi yöntemi kullanılarak seçilen 12 etnik gruba ait bireylerin 21. kromozom'a ait SNP verileri örnekler arası varyansı en büyük kılacak veri uzayına taşınmıştır. Bu işlem sonucunda eğitim kümesindeki örnek sayısı kadar izdüşüm vektörü elde edilmiş ve bu vektörler önem sıralarına göre (her özvektöre karşılık gelen özdeğerin büyüklüğüne göre) sıralanmıştır. 21. kromozom'a ait SNP verisinin, sırasıyla bu vektörler kullanılarak yeni veri uzayına izdüşürülmesiyle hesaplanan jeo-genomik bağıntının ilk dört adet izdüşüm vektörünün kullanıldığı dört boyutlu bir uzayda %62 düzeyine kadar çıktığı görülmüştür. Bu durum, Şekil 4.17'de görülmektedir.

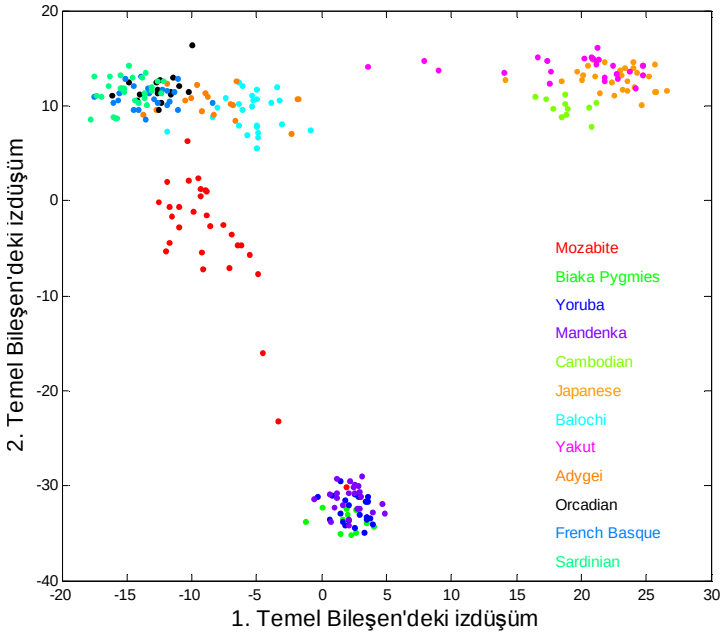


Şekil 4.17: İzdüşüm vektörü sayısına göre jeo-genomik bağıntıdaki değişim

Örnekler arası varyansı açıklamada etkisi düşük olan boyutların genomik mesafe hesabına katılmasıyla beraber jeo-genomik bağıntı beklendiği üzere düşmektedir.

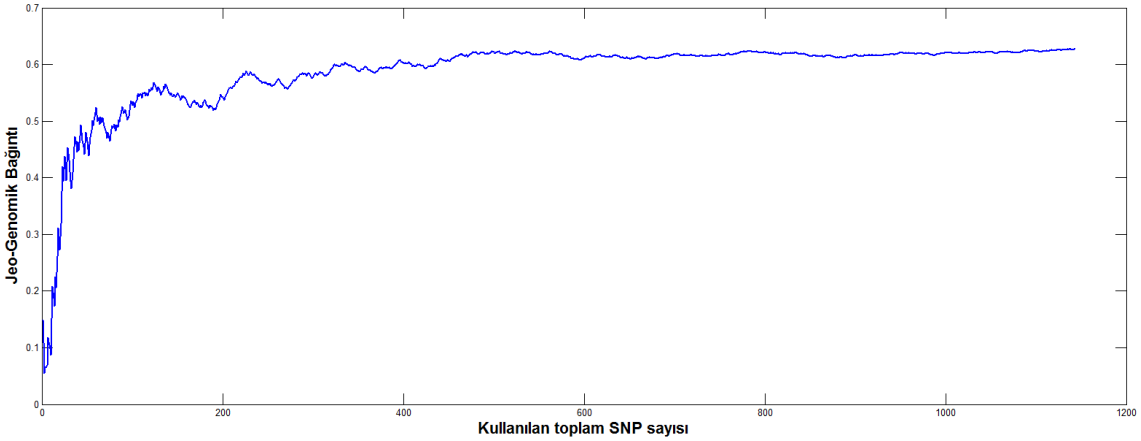
Bölüm 3.1'de değinildiği üzere Temel Bileşen Analizi yöntemi yüksek boyutlu verinin daha düşük boyutlu veri uzayına indirgenerek görselleştirilmesinde de kullanılabilmektedir. Bu amaçla seçilen bireylerin 21. kromozom'a ait olan tüm SNP bilgileri sadece ilk iki temel bileşen kullanılarak temel bileşen uzayına izdüşürüldüğünde Şekil 4.18 elde edilmektedir. Bu şeklin, uç noktalardaki bazı örnekler ihmal edildiğinde Şekil 3.11'de görülen dünya haritasını andırdığı ve bu nedenle anlamlı bir jeo-genomik bağıntı sağlandığı görülmektedir. Avrupa kıtasından seçilen Sardinian, Orcadian ve French_Basque örnekleri ile Asya kıtasından seçilen Yakut, Japanese ve Cambodian örnekleri ayrı iki küme oluşturmuş, Adygei ve Balochi grupları

coğrafi yerleşimlerine paralel olarak bu iki küme arasında yer almıştır. Bununla beraber Afrika kıtasından seçilen Yoruba, Mandenka ve Biaka_Pygmies grupları bir Afrika kümesi oluşturmuş ve bugünkü Fas bölgesinde yaşayan Mozabite grubu yine coğrafi konumuna paralel olarak Avrupa ve Afrika kümeleri arasında yer almıştır.



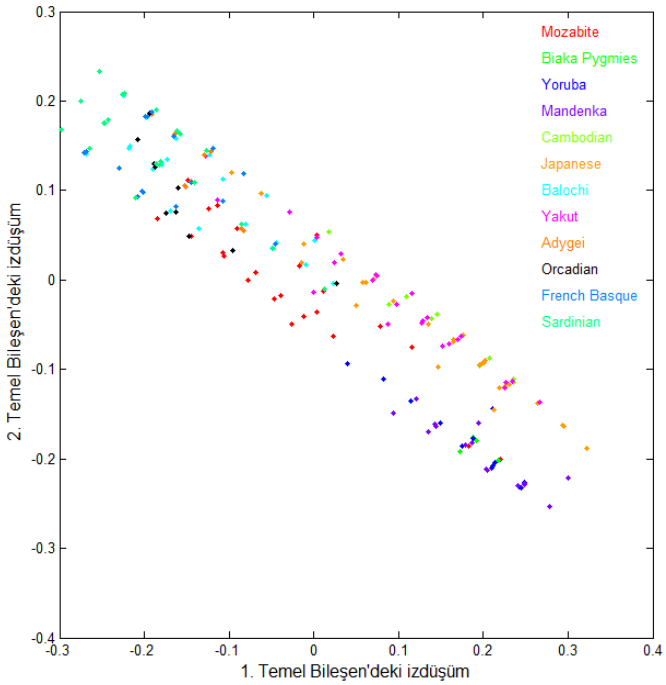
Şekil 4.18: Seçilen örneklerin ilk iki temel bileşene izdüşümü

Her özvektör, 8909 adet SNP için aynı sayıda ağırlık katsayısı içermektedir. Bir SNP'e karşılık gelen ağırlık katsayısının mutlak değeri ne kadar yüksekse bu, o SNP'in söz konusu boyuttaki örnekler arası varyansı açıklamada o kadar büyük bir öneme sahip olduğunu gösterir. Burdan yola çıkarak örnekler arası varyansı en iyi açıklayan ilk *D* adet SNP seçilerek jeo-genomik bağıntıyı koruyan ve aynı zamanda grupların sınıflandırılmasını sağlayan daha küçük boyutlu modeller de oluşturulabilir. Bu işlem için birden fazla kriteri (bu durum için birden fazla özvektördeki SNP ağırlıkları) göz önünde bulundurabilen sıralama teknikleri kullanılabilir (Gumus, 2013a – Gormez, 2013). En basitinden, her SNP'nin birinci ve ikinci özvektörlerdeki ağırlıklarının mutlak değerlerinin çarpımı kullanılarak SNP'ler gruplar arası varyansı sağlamadaki önemlerine göre sıralanabilir. Bu şekilde sıralanmış SNP'lerin kümülatif şekilde kullanılmasıyla elde edilen jeo-genomik bağıntı Şekil 4.19'da görülmektedir.

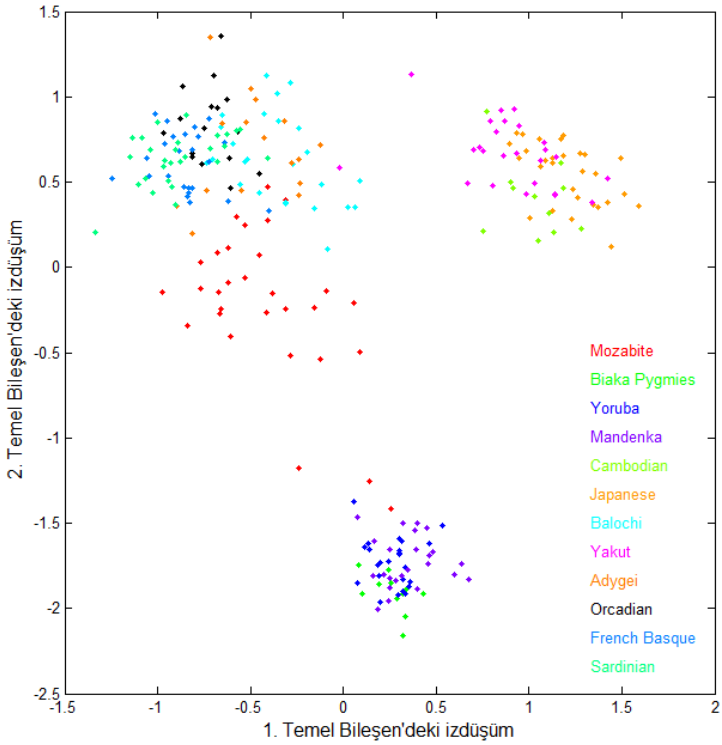


Şekil 4.19: SNP sayısına göre jeo-genomik bağıntıdaki değişim

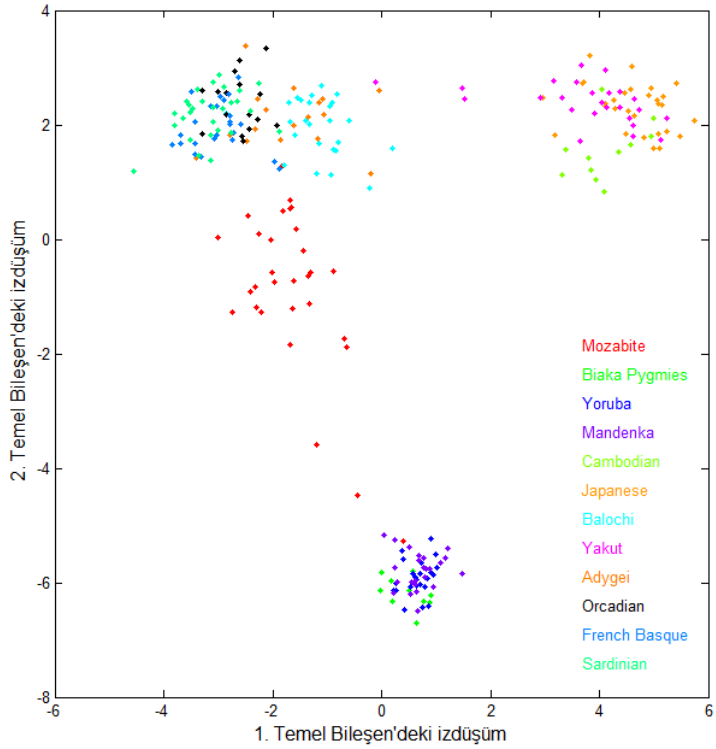
Şekil 4.19 incelendiğinde kullanılan toplam SNP sayısındaki artışın jeo-genomik bağıntı noktasında düzenli olmasa da genel bir artışa yol açtığı görülmektedir. Buna göre sadece en yüksek ağırlık çarpımlı ilk 10 SNP grubu kullanıldığında sadece %8 gibi bir bağıntı elde edilirken, ilk 100 SNP kullanıldığında bu sayı %52'ye çıkmaktadır. Denemelerde en yüksek %62'lik bir jeo-genomik bağıntıya yaklaşık 500 SNP ile ulaşılmış ve bundan sonra belirgin bir artış görülmemiştir. Örneklerin sadece en yüksek ağırlık çarpımına sahip 10, 100 ve 500 SNP ile 1. ve 2. özvektörlerce gerilen bir izdüşüm uzayına düşürülmesi sonucu etnik grupların dağılımı Şekil 4.20 – 4.22'de verilmiştir.



Şekil 4.20: Seçilen örneklerin 10 SNP kullanarak alınan izdüşümü



Şekil 4.21: Seçilen örneklerin 100 SNP kullanarak alınan izdüşümü



Şekil 4.22: Seçilen örneklerin 500 SNP kullanarak alınan izdüşümü

Şekil 4.20 – 4.22 incelendiğinde kullanılan SNP sayısı arttıkça örneklerin izdüşüm uzayındaki dağılımının Şekil 4.18'e yaklaştığı görülmektedir.

5. TARTIŞMA SONUÇ

Son yıllarda gelişen yeni nesil gen tedavisi ve kişiye özel ilaç tasarımı çalışmalarında, viral yerleşim bölgelerindeki karakteristik özelliklerin tespiti büyük önem kazanmıştır. Bu çalışma kapsamında, söz konusu problemin çözümüne yönelik olarak, Kanonik Bağıntı Analizi (KBA) yöntemine dayanan bir örüntü tarama aracı geliştirilmiştir. Bu araç iki farklı veri kümesi üzerinde, (i) kullanılan dizilim özelliği (*N*-mer frekansları veya moment değişkenleri), (ii) görü genişliği ve (iii) görüler arası boşluk gibi üç parametre ile beraber kullanılmış, alınan sonuçların anlamlılığı istatistik testleri ile sınanmıştır. Aracın kullandığı parametreler, aranan örüntünün büyüklüğüne, ayrıklığına ya da birleşikliğine bağlı olarak değişebilmektedir.

Çalışmada öncelikli olarak Schroder'in çalışmasından (Schroder, 2002) elde edilen kimerik veri kümesi kullanılmıştır. Bu küme üzerinde yapılan önceki çalışmalarda (Holman, 2005 – Wu, 2005), virüsün insan genomunda yerleştiği bölgelerin tümü ele alındığında yerleşim noktasını çevreleyen küçük bir alanda her nükleotit pozisyonu için, nükleotitlerin görülme olasılıklarının simetrik/palindromik bir değişime sahip olduğu gösterilmiştir. Bu durum Shannon düzensizliği ve varyans testleri ile de görülmektedir. Sadece bir grup örnek beraber incelendiğinde ortaya çıkan bu özellik, uygun parametreler ile kullanılan örüntü tarama aracıyla yakalanabilir. Burada, bahsedilen özelliğin yakalanmasından kasıt, eğitim kümesinden KBA ile öğrenilen ağırlık katsayılarının test kümesi örneklerinin aynı nükleotit pozisyonunu çevreleyen özellik kümesi ile beraber kullanıldığında yüksek bir test bağıntısı verebilmesidir. Schroder veri kümesi için tanımlanan karakteristik özellik küçük boyutlu bir alanda gözlemlendiğinden (i) dizilim özelliği olarak moment değişkenlerinin kullanımı uygun değildir. Bu nedenle küçük boyutlu görülerde (ii) $w \in \{2, 4, 6, 8\}$ için nükleotit frekansları ve $w \in \{4, 6, 8\}$ için dimer frekansları kullanılmıştır. Söz konusu aranan örüntü, simetrik bir davranış gösterdiğinden nükleotit frekansları kullanıldığında (iii) $1 - w$ gibi bir görüler arası boşluk (g) parametresi kullanılması daha uygundur. Bu şekilde iç içe geçen iki görünün sadece örtüşmeyen, sağ/sol uçlarında kalan (palindromik davranışın gözlemlenebileceği) pozisyonlar arasında bir ilişkinin varlığı araştırılmıştır. Durum dimer frekansları açısından ele alındığında da söz konusu palindromik davranış gereği, görülerin yine iç içe geçmesi/örtüşmesi gerektiği ancak

nükleotit frekanslarının aksine bu örtüşmenin daha az bir oranda olması gerektiği görülmüştür. Bunun nedeni, görülerin örtüşen alanlarının dışında kalan bölgeden yeterli miktarda dimer özelliğinin çıkartılması gereksinimidir. Bu nedenle dimer özelliklerinin kullanıldığı denemelerde g = -1 olarak alınmış ve olası en iyi sonuçların bu parametre ile elde edildiği görülmüştür. Z istatistik testi kullanılarak yapılan analizlerin sonunda, söz konusu simetrik/palindromik davranışın x = +5 noktasını çevreleyen küçük bir alanda görüldüğü doğrulanmıştır.

Bir canlının genomundan alınan okumaların yönü (+/+ veya -/+), söz konusu okumaların içindeki 16 dimerin frekansları ile tahmin edilebilir. Bu tahmin, çok sayıda örnek içeren veri kümelerinin referans genoma eşlenmesi işleminin maliyetini düşürebilir. Zira -/+ yönünde hizalı her okumanın referans genomla eşleşme maliyeti iki sorgudur. Öne sürülen örüntü tarama aracının, Schroder'in veri kümesinin sadece +/+ yönde hizalı bir alt grubundan ziyade, -/+ örnekleri de içeren tüm örnekleri üzerinde kullanılmasıyla bu tahminin de yapılabileceği görülmüştür. Pozitif ve negatif yönde hizalı örneklerden oluşan bir veri kümesinin yarısı eğitim, kalan yarısı test kümesi olarak ayrılmış ve eğitim örneklerinden oluşan kümede w = 150, g = 0 ve dimer dizilim özellikleri parametreleri kullanılarak her nükleotit pozisyonu için bağıntı analizi yapılmıştır. Sonrasında her pozisyon için öğrenilen ağırlık katsayıları test kümesinin aynı pozisyonundaki görülerin dimer frekansları ile çarpılarak test örneklerinin KBA uzayındaki izdüşümleri elde edilmiştir. Söz konusu iki boyutlu (iki görülü) uzaydaki örnekler, izdüşümlerin ortalamasına göre dört bölgeye ayrıldığında birinci ve üçüncü bölgelere (test bağıntısının +1 olabileceği doğrultu) düşen örneklerin sınıf bilgilerinin (dizilim yönlerinin) farklı olduğu görülmüştür. Buradan, KBA yönteminin her dimere verdiği ağırlık katsayılarının, farklı yönlerde hizalı örnekler için farklı aralıklarda görü çıktılarına yol açtığı ve bu farkın, ortalama referans alındığında istenen tahmin işlemi için kullanılabileceği görülmüştür. Bu da yöntemin, dizilimlerin farklı bölgeleri arasında saklı olan özel bir ilişkiyi yakalamanın yanı sıra bir danışmasız öğrenme şekli olan öbekleme amacıyla da kullanılabileceğini göstermektedir. Bununla birlikte, yön tahmini amacıyla kullanılabilecek başka makine öğrenme yöntemleri de vardır. Söz gelimi, orijinal veri uzayında doğrusal ayrılamayan örnekleri daha yüksek boyutlu bir uzaya taşıyarak doğrusal ayırabilmesi nedeniyle sıkça başvurulan Destek Vektör Makineleri (DVM) yöntemi de bu amaçla kullanılabilir. Her iki yöntem aynı nükleotit

pozisyonunda ve eşit sayıda nükleotit üzerinde kullanıldığında, KBA yöntemi DVM yöntemine göre daha iyi bir ayırma başarısına sahip olmuştur (Bkz: Şekil 4.9). Ayrıca, DVM yönteminin yön tahminini sağlıklı olarak (%90 üzeri doğrulukla) yapabilmesi için en az kabaca $2w = 100$ nükleotit genişlikte bir alana ihtiyaç duyarken, KBA yönteminin aynı doğruluk oranını en az kabaca $2w = 50$ genişlikte bir alanda sağlayabildiği görülmüştür. Bu da KBA yönteminin bu amaç için daha tercih edilebilir olduğunu göstermektedir.

Schroder'in çalışmasına benzer olarak DETAE bünyesinde yapılan başka bir çalışmada insan konağına ait hücreler HIV'den türetilmiş bir lentivirüs ile enfekte edilmiş ve bu hücrelere ait okumalar GS FLX platformu (Bkz: Bölüm 2.1.2) ile elde edilmiştir. Smith-Waterman parçalı eşleştirme tekniği (Bkz: Bölüm 3.6.2) ile, içinde bu lentivirüse ait parçalar bulunduran okumalar (kimerik okumalar) tespit edilerek ön eleme yapılmış ve BLAST (Bkz: Bölüm 3.6.3) aracı kullanılarak bu okumalardaki insan genomuna ait kısımlar kromozomal pozisyonları ile beraber bulunmuştur. Tekrarlı okumaların elenmesiyle beraber söz konusu virüsün insan genomunda yerleşmeyi tercih ettiği 76 farklı bölgeye ait 10000 nükleotit uzunluklu okumalar virüs yerleşim noktasına göre hizalanarak bir araya getirilmiştir. Bu okumalar eğitim ve geçerleme kümeleri olarak kabul edilmiş ve örüntü tarama aracı, değişen görü (pencere) genişlikleri ve dimer dizilim özellikleri kullanılarak söz konusu 10000 nükleotit uzunluklu viral yerleşim alanında sadece viral yerleşime özgü bir örüntünün bulunup bulunmadığını tespit etmek için kullanılmıştır. Burada $w \geq 300$ gibi geniş bir alanda yapılan KBA eğitimi ile öğrenilen ağırlık katsayılarının, viral yerleşimin olmadığı bilinen ve insan genomunun 400 farklı bölgesinden çekilen okumalardan (test kümesi) elde edilen dimer dizilim özellikleri ile beraber kullanıldığında, en yüksek test bağıntılarının ortalama geçerleme bağıntılarını hiçbir şekilde geçemediği bölgeler bulunmuştur. Bu da, söz konusu bölgelerde yüksek geçerleme bağıntısına neden olan karakteristiğin (örüntünün) rastgele olamayacağını göstermektedir. Ancak bu karakteristiğin bu denli geniş görüler için kayan pencere yöntemi ile beraber düşünüldüğünde sadece ardışık birkaç pozisyonu çevreleyen alanlarda görülmesi gerçekçi olmayacaktır. Söz konusu karakteristiğin ardışık bir grup alanda devam etmesi gerekmektedir. Bu amaçla her x_n nükleotit pozisyonu için $[x_n - 50 : x_n + 50]$ aralığındaki ortalama geçerleme bağıntılarının yine aynı aralıktaki en yüksek test bağıntılarından ne kadar farklı bir medyana sahip olduğu

(iki dağılımın ne kadar farklı olduğu) Mann-Whitney-Wilcoxon sıralama toplam testi (Bkz: Bölüm 3.5) ile hesaplanmıştır. Buna göre en az 100 nükleotit pozisyonu boyunca ($x_n + 50 - (x_n - 50)$) istatistiksel medyan testine göre birbirlerinden farklı geçerleme-test bağıntılarına sahip olan ve bu farkın, test bağıntılarının kendi içlerinde birbirlerine göre sınanması ile elde edilen eşik değerinden yüksek olduğu bölgelerdeki dimer frekansı karakteristiğinin sadece viral okumalara yani virüsün yerleşmeyi tercih ettiği bölgelere özgü olduğu değerlendirilmiştir. Bu bölgeler, $x_n = 0$ virüs yerleşim noktası olmak üzere, $x_n \in$ [-2283 : -2153] aralığındaki tüm x_n nükleotit pozisyonlarını merkez kabul eden $[x_n - w : x_n + w]$ ($w = 300$) genişliğindeki alanlardır. Yeni bir çalışma konusu olarak, viral yerleşime özgü, rasgele olmayan bir dimer frekansı karakteristiğinin bulunduğu bu bölgelerde üç boyutlu dizilim özelliklerini de kullanan örüntü tarama aracıyla analizler yapılması planlanmaktadır. Ancak mevcut haliyle dahi, istatistik testlerinin desteğiyle söz konusu aracın, sadece okumalar toplu olarak değerlendirildiğinde ortaya çıkan karakteristikleri yakalayabildiği görülmüştür.

Bu çalışma kapsamında önerilen örüntü tarama aracına dayanak olan KBA yönteminin, farklı amaçlarla kullanımı da yine DETAE'den elde edilen veri kümeleri ile araştırılmıştır. Burada, kurumdan elde edilen Behçet veri kümesindeki her SNP ayrık sayısal değerlerle kodlanmış ve ait oldukları kromozomlara göre farklı kümelere tasnif edilmiştir. KBA yöntemi, SNP'ler üzerinde kayan pencere yaklaşımıyla kullanıldığında, birbirine komşu SNP grupları ($g = 0$ ile yapışık pencereler) arasında yüksek bir KBA eğitim bağıntısının elde edildiği ve buna karşın $g > 0$ için bu bağıntının azaldığı görülmüştür. Bu durum, biyolojik bir etmen olan "bağlantı eşitsizliği" kavramına dayanmaktadır. Literatürde geçen bağlantı eşitsizliği ölçütleri (Bkz: Bölüm 2.3.1.1) sadece iki SNP arasındaki ilişkiyi ölçebilirken KBA yöntemi SNP grupları arasındaki ilişkiyi ölçebilmektedir. Bu özellik, Behçet hastalığı gibi genetik kökenli hastalıkların genomdaki kaynağını belirlemek için kullanılabilir. Hasta ve sağlıklı bireylerin aynı bölgedeki SNP'leri içeren görüleri için yapılan KBA testlerinde, eğitim kümesinden elde edilen ağırlık katsayılarının geçerleme kümesinde daha düşük çıktılara neden olduğu ve ortalama eğitim – ortalama geçerleme çıktıları arasındaki farkın yüksek olduğu bölgelerin farklı bağlantı eşitsizliği değerlerine sahip olduğu tespit edilmiştir. Aynı ırkın bireylerini içeren bu veri kümesi için ortalama bağıntılar arasındaki bu farkın hastalık etmeni SNP'lerden kaynaklandığı açıktır. Yöntemin, literatürde çokça

kullanılan ki-kare testine göre avantajı, bir SNP'i yalnız başına kendi içindeki dağılıma bakarak değerlendirmesinden ziyade komşularındaki dağılımları da dikkate alması ve hastalarla sağlıklılar arasında görülen bu farka hangi SNP'in ne kadar ağırlıkla etkili olduğunu belirleyebilmesidir. Bununla beraber yöntem, beklendiği üzere (Remmers, 2010) bu farkın yüksek oranda görüldüğü bölgenin 6. kromozomun MHC bölgesi olduğunu doğrulamıştır.

Çalışma kapsamında değinilen son konu, dünyanın farklı bölgelerindeki milletlerin birbirlerine göre olan ortalama genomik mesafeleri (genetik olarak ortalama benzemezlik miktarları) ile coğrafi mesafeleri arasındaki ilişkinin (jeo-genomik bağıntı) analizi ve bu ilişkinin daha güçlü ifade edilmesi amacıyla neler yapılabileceği olmuştur. Bu amaç için, Temel Bileşen Analizi (TBA) yönteminden faydalanılmış ve örneklerin varyansını daha iyi ifade eden yeni bir TBA izdüşüm uzayının kullanılmasıyla jeo-genomik bağıntı düzeyinin arttığı görülmüştür. Zira 21. kromozom'un tüm SNP'leri kullanılarak (8909 boyutlu orijinal veri uzayı) elde edilen jeo-genomik bağıntı %52 düzeyinde gerçekleşirken, sadece 4 boyutlu bir TBA uzayına aktarılan örneklerden elde edilen jeo-genomik bağıntı %62 düzeyine çıkmıştır. Böylece mevcut jeo-genomik ilişki, daha düşük boyutlu veri ile daha iyi bir şekilde ifade edilebilir hale gelmiştir. Bunun yanı sıra, aktarım işleminde kullanılan en iyi N adet izdüşüm vektörünün içerdiği 8909 adet ağırlık katsayısının, orijinal uzaydaki her SNP'nin jeo-genomik bağıntıya olan etkisinin büyüklüğünü gösterdiği bilinmektedir. Bu noktadan yola çıkarak, jeo-genomik bağıntıyı en iyi ifade eden SNP alt grupları da elde edilebilir. Yapılan analizlerde, 8909 SNP'den kabaca sadece 500 adedi (ilk iki izdüşüm vektöründeki mutlak ağırlık katsayılarının çarpımı en yüksek olan ilk 500) kullanıldığında da yine %62 oranında bir jeo-genomik bağıntı elde edilebileceği görülmüştür. SNP ya da öznitelik seçimi olarak isimlendirebileceğimiz bu işlem, sadece jeo-genomik bağıntının en iyileştirilmesinden ziyade hem bu bağıntının hem de sınıflandırma başarısının beraber en iyi olduğu özniteliklerin tespiti şeklinde de gerçekleştirilebilir (Gumus, 2013a). Bu işlem için literatürde geçen pek çok, çok-kriterli en iyileme/sıralama yöntemi mevcuttur. Bu yöntemlerin birbirlerine göre performansları karşılaştırılmış olup (Gormez, 2013), ileride bu konuda yeni analizler yapılması planlanmaktadır.

KAYNAKLAR

ADDGENE, 2012, *Plasmid 12247: pLVTHM* [online], http://www.addgene.org/12247/, [Ziyaret Tarihi: 24 Mayıs 2012].

ALPAYDIN E., 2007, *Yapay Öğrenme*, Boğaziçi Üniversitesi Yayınevi, İstanbul, ISBN:9786054238491.

ALTSCHUL S., GISH W., MILLER W., MYERS E., LIPMAN D., 1990, Basic Local Alignment Search Tool, *Journal of Molecular Biology*, Vol.215, PP.403-410.

AVERY O.T., MACLEOD C.M., MCCARTY M., 1944, Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types, *The Journal of Experimental Medicine*, Vol.79(2), PP.137-158.

BALDI P., HATFIELD G.W., 2002, *DNA Microarrays and Gene Expression*, Cambridge, ISBN: 9780521800228.

BISHOP C.M., 2006, *Pattern Recognition and Machine Learning*, Springer, New York, ISBN-10:0387310738.

BOUTTE D., LIU J., 2010, Sparse Canonical Correlation Analysis Applied to fMRI and Genetic Data Fusion, *IEEE International Conference on Bioinformatics and Biomedicine*, 18-21 Aralık 2010, PP.422-426.

COVER T.M., THOMAS J.A., 1991, *Elements of Information Theory*, John Wiley&Sons Inc., New York, ISBN:0-471-06259-6.

DAVIES D., 1999, The Bombe: a Remarkable Logic Machine, *Cryptologia*, Vol.23(2), PP.108-138.

GORANIA M., SEKER H., HARIS P.I., 2010, Predicting a Protein's Melting Temperature From Its Amino Acid Sequence, *Conference of the IEEE Engineering in Medicine and Biology Society*, Vol.1, PP. 1820-1823.

GORMEZ Z., GUMUS E., KURSUN O. SERTBAS A., 2013, Comparison of Aggregators for Multi-Objective SNP Selection, *The 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13)*, 3-7 Temmuz 2013, Osaka (kabul edildi).

GOTOH O., 1982, An Improved Algorithm for Matching Biological Sequences, *Journal of Molecular Biology*, Vol.162, PP.705-708.

GUMUS E., KURSUN O., SERTBAS A. USTEK D., 2012, Application of Canonical Correlation Analysis for Identifying Viral Integration Preferences, *Bioinformatics*, Vol.28(5), PP.651-655.

GUMUS E., GORMEZ Z., KURSUN O., 2013a, Multi Objective SNP Selection Using Pareto Optimality, *Computational Biology and Chemistry*, Vol.43, PP.23-28.

GUMUS E., KURSUN O. SERTBAS A., 2013b, Viral Yerleşim Bölgelerindeki Karakteristiğin Sınıflandırma Başarımına Etkisi, *21. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU2013)*, 23-26 Nisan 2013, KKTC.

HARDOON D.R., SZEDMAK S., TAYLOR J.S., 2004, Canonical Correlation Analysis: An Overview with Application to Learning Methods, *Neural Computation*, Vol.16(12), PP.2639-2664.

HOLMAN A.G., COFFIN J.M., 2005, Symmetrical Base Preferences Surrounding HIV-1, Avian Sarcoma/Leukosis Virus and Murine Leukemia Virus Integration Sites, *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, Vol.102(17), PP.6103-6107.

HOTELLING H., 1936, Relations Between Two Sets of Variates, *Biometrika*, Vol.28, PP.321-377.

HUFFMAN D.A., 1952, A Method for the Construction of Minimum-Redundancy Codes, *Proceedings of the I.R.E.*, PP.1098-1102.

KITTLES R.A., WEISS K.M., 2003, Race, ancestry and genes: implications for defining disease risk, *Annual Review of Genomics and Human Genetics*, Vol.4, PP.33-67.

MODELS IN HUMAN GENETICS, 2006, *Lecture 03 -- Linkage Disequilibrium* [online], http://www.sph.umich.edu/csg/abecasis/class/666.03.pdf , [Ziyaret Tarihi: 16 Temmuz 2012].

MIZUKI N., OTA M., KATSUYAMA Y., YABUKI K., ANDO H., SHIINA T., NOMURA E., ONARI K., OHNO S., INOKO H., 2001, HLA-B*51 allele analysis by the PCR-SBT method and a strong association of HLA-B*5101 with Japanese patients with Behçet's disease, *Tissue Antigens*, Vol.58(3), PP.181-184.

NAYLOR M.G., LIN X.L., WEISS S.T., RABY B.A., LANGE C., 2010, Using Canonical Correlation Analysis to Discover Genetic Regulatory Variants, *PLOS One*, Vol.5(5), e10395.

NEEDLEMAN S.B., WUNSCH C.D., 1970, A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *Journal of Molecular Biology*, Vol.48(3), PP.443–453.

NOVEMBRE J., JOHNSON T., BRYC K., KUTALIK Z., BOYKO A.R., AUTON A., INDAP A., KING K.S., BERGMANN S., NELSON M.R., STEPHENS M., BUSTAMANTE C.D., 2008, Genes Mirror Geography within Europe, *Nature (Letter)*, Vol.456, PP.98-101.

PARKHOMENKO E., TRITCHLER D., BEYENE J., 2009, Sparse Canonical Correlation Analysis with Application to Genomic Data Integration, *Statistical Applications in Genetics and Molecular Biology*, Vol.8(1), 1st Article.

PEDIATRIC RHEUMATOLOGY, 2003, *Behçet Hastalığı* [online], http://www.printo.it/pediatric-rheumatology/information/Turchia/PDF/10_BEHCET_turkey.pdf , [Ziyaret Tarihi: 16 Temmuz 2012].

PENG Q., ZHAO J., XUE F., 2010, A Gene Based Method for Detecting Gene-Gene Co-Association in a Case-Control Association Study, *European Journal of Human Genetics*, Vol.18, PP.582-587.

REMMERS E.F., COSAN F., KIRINO Y., OMBRELLO M.J., ABACI N., SATORIUS C., LE J.M., YANG B., KORMAN B.D., CAKIRIS A., AGLAR O., EMRENCE Z., AZAKLI H., USTEK D. TUTKUN I., DEMIR G., CHEN W., AMOS C.I., DIZON M.B., KOSE A.A., AZIZLERLI G., ERER B., BRAND O.J., KAKLAMANI V.G., KAKLAMANIS P., CHETRIT E., STANFORD M., FORTUNE F., GHABRA M., OLLIER W.E.R., CHO Y.H., BANG D., OSHEA J., WALLACE G.R., GADINA M., KASTNER D.L., GUL A., 2010, Genome-Wide Association Study Identifies Variants in the MHC Class I, IL10 and IL23R-IL12RB2 Regions Associated with Behçet's Disease, *Nature Genetics*, Vol.42(8), PP.698-702.

SAMUELS M.L., WITMER J.A., SCHAFFNER A., 2012, *Statistics for the Life Sciences*, Prentice Hall, New Jersey, ISBN-10:0-321-65280-0.

SANGER F., NICKLEN S., COULSON A.R., 1977, DNA Sequencing with Chain-Terminating Inhibitors, *Proceedings of the National Academy of Sciences of the United States of America*, Vol.74(12), PP.5463-5467.

SARGIN M.E., YEMEZ Y., TEKALP A.M., 2007, Audio-Visual Synchronization and Fusion Using Canonical Correlation Analysis, *IEEE Transactions on Multimedia*, Vol.9(7), PP.1396-1403.

SCHRODER A.R., SHINN P., CHEN H., BERRY C., ECKER JR, BUSHMAN F., 2002, HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots, *Cell*, Vol.110(4), PP.521-529.

SFORZA L.L, 2005, The Human Genome Diversity Project: Past, Present and Future, *Nature Reviews-Genetics*, Vol. 6, PP. 333-340.

SHANNON C.E., 1948, A Mathematical Theory of Communication, *Bell System Technical Journal*, Vol.27(3), PP.379-423.

SHI J., ZHANG S., LIANG Y., PAN Q., 2006, Prediction of Protein Subcellular Localizations Using Moment Descriptors and Support Vector Machine, *Lecture Notes in Computer Science*, Vol. 4146/2006, PP.105-114.

SIRIUS GENOMICS, 2011, *Technology* [online], http://www.siriusgenomics.com/technology/ , [Ziyaret Tarihi: 24 Mayıs 2012].

SMITH T.F., WATERMAN M.S., 1981, Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, Vol.147, PP.195-197.

SPRINTHALL R.C., 2003, *Basic Statistical Analysis*, 7. Baskı, Pearson, Boston, ISBN: 1428814248.

ŞANLIOĞLU S., ÇAĞLAYAN A.O., 2010, Gen Tedavisi, *Modern Biyoteknoloji ve Uygulamaları*, Erciyes Üniversitesi Yayınları, (Bölüm 19) ISBN: 9789756478639.

TURING A., 1950, Computing Machinery and Intelligence, *Mind*, Vol.59(236), PP.433-460.

TURK M.A., PENTLAND A.P., 1991, Face Recognition Using Eigenfaces, *Proc. IEEE Conference on Computer Vision and Pattern Recognition,* PP.586-591.

USTEK D., SIRMA S., GUMUS E., ARIKAN M., CAKIRIS A., ABACI N., MATHEW J., EMRENCE Z., AZAKLI H., COSAN F., CAKAR A., PARLAK M., KURSUN O. ,2012, A Genome-Wide Analysis of Lentivector Integration Sites Using Targeted Sequence Capture and Next Generation Sequencing Technology, *Infection, Genetics and Evolution*, Vol. 12(7), PP.1349-1354.

VAPNIK V., 1998, *Statistical Learning Theory*, Wiley, New York, ISBN: 9780471030034

VINCENTY T., 1975, Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations, *Survey Review*, Vol.23(176), PP.88–93.

VINOKOUROV A., HARDOON D.R., TAYLOR J.S., 2003, Learning the Semantics of Multimedia Content with Application to Web Image Retrieval and Classification, *Proc. of 4th International Symposium on Independent Component Analysis*, PP.697-701.

WAAIJENBORG S., WITT HAMER P.C.V., ZWINDERMAN A.H., 2008, Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis, *Statistical Applications in Genetics and Molecular Biology*, Vol.7(1), 3rd article.

WAAIJENBORG S., ZWINDERMAN A.H., 2009, Correlating Multiple SNPs and Multiple Disease Phenotypes: Penalized Non-linear Canonical Correlation Analysis, *Bioinformatics*, Vol.25(21), PP.2764-2771.

WAAIJENBORG S., ZWINDERMAN A.H., 2010, Association of Repeatedly Measured Intermediate Risk Factors for Complex Diseases with High Dimensional SNP Data, *Algorithms for Molecular Biology*, Vol.5,17th article.

WEINBERG W., 1908, Über den Nachweis der Vererbung beim Menschen, *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, Vol.64, PP.368–382.

WITTEN D.M., TIBSHIRANI R.J., 2009, Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data, *Statistical Applications in Genetics and Molecular Biology*, Vol.8(1), 28th Article.

WU X., LI Y., CRISE B., BURGESS S.M., MUNROE D.J., 2005, Weak Palindromic Consensus Sequences Are a Common Feature Found at the Integration Target Sites of Many Retroviruses, *Journal of Virology*, Vol.79(8), PP.5211-5214.

ZHANG Z., SCHWARTZ S., WAGNER L., MILLER W., 2000, A Greedy Algorithm for Aligning DNA Sequences, *Journal of Computational Biology*, Vol.7(1-2), PP.203-214.

ÖZGEÇMİŞ

Ergün Gümüş 1984 yılında Almanya'nın Stuttgart şehrinde doğdu. Ortaokul ve lise öğrenimini İstanbul Vefa Anadolu Lisesi'nde gördü. 2002-2006 ve 2006-2008 yılları arasında İstanbul Üniversitesi Bilgisayar Mühendisliği Bölümünde lisans ve yüksek lisans programlarını tamamladı. Kendisi, 2008 yılında doktora programına başladığı aynı bölümde halen araştırma görevlisi olarak çalışmaktadır. Çalışma konuları, makine öğrenmesi, örüntü tanıma ve biyoinformatik alanlarını kapsamaktadır.

Ergün GÜMÜŞ