T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

GENOM ÇAPLI İLİŞKİ ÇALIŞMALARI İÇİN YAPAY ZEKÂ TEKNİKLERİ İLE ETİKET SNP SEÇİMİ

İlhan İLHAN

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Ocak-2013
KONYA

TEZ KABUL VE ONAYI

İlhan İLHAN tarafından hazırlanan "Genom Çaplı İlişki Çalışmaları İçin Yapay Zekâ Teknikleri İle Etiket SNP Seçimi" adlı tez çalışması 21/01/2015 tarihinde aşağıdaki jüri tarafından oy birliği / ~~oy çokluğu~~ ile Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda DOKTORA TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof. Dr. Bekir KARLIK	.....
Danışman Yrd. Doç. Dr. Gülay TEZEL	.....
Üye Doç. Dr. Hasan OĞUL	.....
Üye Doç. Dr. Mehmet ÇUNKAŞ	.....
Üye Yrd. Doç. Dr. Ömer Kaan BAYKAN	.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. Aşır GENÇ
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

İlhan İLHAN

Tarih: 21/01/2013

ÖZET

DOKTORA TEZİ

GENOM ÇAPLI İLİŞKİ ÇALIŞMALARI İÇİN YAPAY ZEKÂ TEKNİKLERİ İLE ETİKET SNP SEÇİMİ

İlhan İLHAN

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Gülay TEZEL

2013, 136 Sayfa

Jüri
Yrd. Doç. Dr. Gülay TEZEL
Prof. Dr. Bekir KARLIK
Doç. Dr. Hasan OĞUL
Doç. Dr. Mehmet ÇUNKAŞ
Yrd. Doç. Dr. Ömer Kaan BAYKAN

Karmaşık hastalıklarla ilişkili genetik değişimlerin araştırılması daha iyi teşhis ve tedaviye hizmet edebilecek önemli bir araştırma konusudur. İnsan genomunda bulunan milyonlarca değişimin çoğunluğunu oluşturan Tek Nükleotid Polimorfizm'leri (Single Nucleotide Polimorphisms - SNPs) hastalık-gen ilişki çalışmaları için oldukça umut vericidir. Bununla beraber, bu çalışmalar çok büyük sayıdaki SNP'leri genotipleme maliyeti ile sınırlıdır. Bu nedenle geriye kalan SNP'leri mümkün olduğu kadar maksimum doğrulukla temsil edecek bütün SNP'lerin bir alt kümesini belirlemek esastır. Bu işlem etiket SNP (tag SNP) seçimi olarak bilinir.

Bu tez çalışmasında, etiket SNP seçim probleminin çözümü için üç farklı metot önerilmiştir. Bu metotlardan ilkinde, etiket SNP seçim metodu olarak Genetik Algoritma (Genetic Algorithm - GA) ve geriye kalan SNP'lerin tahmini için ise Destek Vektör Makinesi (Support Vector Machine - SVM) kullanılmış ve GA-SVM olarak adlandırılmıştır. Geliştirilen ikinci metotta ise GA-SVM metodunda sabit değer olarak kullanılan SVM'nin C ve γ parametrelerinin optimizasyonu için Parçacık Sürü Optimizasyon (Particle Swarm Optimization - PSO) algoritması kullanılmış ve Parametre Optimizasyonlu GA-SVM yöntemi olarak adlandırılmıştır. Son metotta ise Parametre Optimizasyonlu GA-SVM yöntemindeki etiket SNP seçim metodu olarak kullanılan GA yerine Klonal Seçim Algoritması (Clonal Selection Algorithm - CLONALG) kullanılmış ve Parametre Optimizasyonlu CLONTagger yöntemi olarak adlandırılmıştır.

Geliştirilen bu metotlar, farklı veri kümeleri üzerinde farklı sayıdaki etiket SNP'ler için güncel metotlardan daha iyi tahmin doğruluğu sağlamıştır.

Anahtar Kelimeler: Destek Vektör Makinesi, Etiket SNP, Genetik Algoritma, Klonal Seçim Algoritması, Parçacık Sürü Optimizasyonu, Tek Nükleotid Polimorfizm.

ABSTRACT

Ph.D THESIS

TAG SNP SELECTION WITH ARTIFICIAL INTELLIGENCE TECHNIQUES FOR GENOME-WIDE ASSOCIATION STUDIES

İlhan İLHAN

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTER ENGINEERING

Advisor: Asst. Prof. Dr. Gülay TEZEL

2013, 136 Pages

Jury
Asst. Prof. Dr. Gülay TEZEL
Prof. Dr. Bekir KARLIK
Assoc. Prof. Dr. Hasan OĞUL
Assoc. Prof. Dr. Mehmet ÇUNKAŞ
Asst. Prof. Dr. Ömer Kaan BAYKAN

Investigations on genetic variants associated with complex diseases are one of important research subjects to serve the better diagnosis and treatment. SNPs (Single Nucleotide Polymorphisms), which comprise most of the millions of changes in human genome, are promising tools for disease-gene association studies. On the other hand, these studies are limited by cost of genotyping tremendous number of SNPs. Therefore, it is essential to identify a subset of SNPs that represents rest of the SNPs with the accuracy as high as possible. This process is known as the tag SNP selection.

In this study, three different methods are proposed for the solution of the problem of tag SNP selection. In the first of these methods, Genetic Algorithm (GA) was used as tag SNP selection method and Support Vector Machine (SVM) was used for prediction the rest of SNPs. This new method is called as GA-SVM. In the proposed second method, Particle Swarm Optimization (PSO) algorithm was preferred to optimize C and γ parameters of SVM used as a fixed value in the GA-SVM method and it is called as GA-SVM method with parameter optimization. In the last method, instead of GA implemented as tag SNP sellection method in the GA-SVM method with parameter optimization Clonal Selection Algorithm (CLONALG) was used and it is called as CLONTagger method with parameter optimization.

The proposed methods demonstrate considerably higher prediction accuracy than current methods for different datasets at different numbers of tag SNPs.

Keywords: Support Vector Machine, Tag SNP, Genetic Algorithm, Clonal Selection Algorithm, Particle Swarm Optimation, Single Nucleotide Polymorphism.

ÖNSÖZ

Bu çalışma ile genom çaplı ilişki çalışmalarında, karmaşık hastalıklarla ilgili genetik faktörleri çalışırken mümkün olduğu kadar yüksek bir tahmin doğruluğu ile SNP'lerin geri kalanını temsil edebilecek minimum sayıdaki etiket SNP kümesinin seçilmesi amaçlanmıştır. Bunun için yapay zekâ teknikleri yardımıyla mevcut metotların başarımı geliştirilecektir.

Yapılan araştırmalara göre etiket SNP seçimi konusunda Türkiye'de bir ilk olan bu doktora tezimin, genom çaplı ilişki çalışmalarında kullanılması ve lisansüstü öğrencilerinin konuyla ilgili araştırmalarına bir ışık olması en büyük dileğimdir.

Çalışmamda bilgi ve becerilerini paylaşan, her türlü destek ve anlayışını esirgemeyen, yönlendirmeleri ile tezimin tamamlanmasına yardımcı olan danışmanım Selçuk Üniversitesi Mühendislik-Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü öğretim üyesi Yrd. Doç. Dr. Gülay TEZEL hocama en içten teşekkürlerimi sunarım.

Doktora çalışmam boyunca uzun bir süre danışmanlığımı yürüten değerli bilgi, beceri ve yardımlarını hiç esirgemeyen Mevlana Üniversitesi Mühendislik-Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü öğretim üyesi Prof. Dr. Şirzat KAHRAMANLI hocama özellikle şükranlarımı sunarım. Ayrıca doktora öğrenciliğime başladığım ilk dönemlerde danışmanım olan Mevlana Üniversitesi Mühendislik-Mimarlık Fakültesi Elektrik-Elektronik Mühendisliği Bölümü öğretim üyesi Prof. Dr. Mehmet BAYRAK hocama değerli katkı ve önerilerinden dolayı teşekkürlerimi sunarım.

Tez izleme toplantılarında jüri üyesi olarak görev alan Selçuk Üniversitesi Teknik Eğitim Fakültesi Elektronik ve Bilgisayar Eğitimi Bölümü öğretim üyesi Doç. Dr. Mehmet ÇUNKAŞ ve Selçuk Üniversitesi Mühendislik-Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü öğretim üyesi Yrd. Doç. Dr. Ömer Kaan BAYKAN hocalarıma değerli öneri ve yardımlarından dolayı teşekkürlerimi sunarım.

Doktora çalışmam boyunca sürekli ihmal etmeme rağmen bana sabırla destek veren ve anlayışlarını hiç esirgemeyen eşim SEBAHAT, kızım KÜBRA ve oğlum KUTAY'a en içten şükranlarımı sunarım.

İlhan İLHAN
KONYA-2013

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Simgeler

σ	: Sigma
γ	: Gamma
C	: SVM'nin kontrol parametresi
A, T, C, G	: Adenin, Timin, Sitozin, Guanin
r^2	: Korelasyon katsayısı
H	: GA-SVM yöntemindeki giriş veri matrisi
m	: GA-SVM yöntemindeki H matrisindeki satır sayısı
n	: GA-SVM yöntemindeki H matrisindeki sutün sayısı
P	: GA-SVM yöntemindeki GA için popülasyon matrisi
q	: GA-SVM yöntemindeki P matrisindeki satır sayısı
N	: GA-SVM yöntemindeki etiket SNP sayısı
PA	: GA-SVM yöntemindeki etiket SNP kümesinin tahmin doğruluğu
N_c	: GA-SVM yöntemindeki doğru olarak tahmin edilen SNP sayısı
N_a	: GA-SVM yöntemindeki toplam tahmin edilen SNP sayısı
C_R	: GA-SVM yöntemindeki çaprazlama oranı
M_R	: GA-SVM yöntemindeki mutasyon oranı
M	: GA-SVM yöntemindeki P popülasyon matrisindeki her bir kromozomdaki etiket SNP sayısı
N_G	: GA-SVM yöntemindeki GA için jenerasyon sayısı
TS	: GA-SVM yöntemindeki etiket SNP'lerin kümesi
H	: Parametre optimizasyonlu GA-SVM yöntemindeki giriş veri matrisi
m	: Parametre optimizasyonlu GA-SVM yöntemindeki H matrisindeki satır sayısı
n	: Parametre optimizasyonlu GA-SVM yöntemindeki H matrisindeki sutün sayısı
P_{GA}	: Parametre optimizasyonlu GA-SVM yöntemindeki GA için popülasyon matrisi
N_p	: Parametre optimizasyonlu GA-SVM yöntemindeki P_{GA} matrisindeki satır sayısı
N	: Parametre optimizasyonlu GA-SVM yöntemindeki etiket SNP sayısı
PA	: Parametre optimizasyonlu GA-SVM yöntemindeki etiket SNP kümesinin tahmin doğruluğu
N_c	: Parametre optimizasyonlu GA-SVM yöntemindeki doğru olarak tahmin edilen SNP sayısı
N_a	: Parametre optimizasyonlu GA-SVM yöntemindeki toplam tahmin edilen SNP sayısı
C_R	: Parametre optimizasyonlu GA-SVM yöntemindeki çaprazlama oranı
M_R	: Parametre optimizasyonlu GA-SVM yöntemindeki mutasyon oranı
M	: Parametre optimizasyonlu GA-SVM yöntemindeki P_{GA} popülasyon matrisindeki her bir kromozomdaki etiket SNP sayısı
N_G	: Parametre optimizasyonlu GA-SVM yöntemindeki GA için jenerasyon sayısı
TS	: Parametre optimizasyonlu GA-SVM yöntemindeki etiket SNP'lerin kümesi
P_{PSO}	: Parametre optimizasyonlu GA-SVM yöntemindeki PSO için popülasyon matrisi

pbest	: Parametre optimizasyonlu GA-SVM yöntemindeki PSO için bir parçacığın o ona kadar ulaştığı en iyi değer
gbest	: Parametre optimizasyonlu GA-SVM yöntemindeki PSO için popülasyondaki global en iyi değer
c1, c2	: Parametre optimizasyonlu GA-SVM yöntemindeki PSO için öğrenme faktörleri
r1, r2	: Parametre optimizasyonlu GA-SVM yöntemindeki PSO için üretilen rastgele değerler
w	: Parametre optimizasyonlu GA-SVM yöntemindeki PSO için atalet ağırlığı
H	: Parametre optimizasyonlu CLONTagger yöntemindeki giriş veri matrisi
m	: Parametre optimizasyonlu CLONTagger yöntemindeki H matrisindeki satır sayısı
n	: Parametre optimizasyonlu CLONTagger yöntemindeki H matrisindeki sutün sayısı
P_{CSA}	: Parametre optimizasyonlu CLONTagger yöntemindeki CLONALG için popülasyon matrisi
Ab	: Parametre optimizasyonlu CLONTagger yöntemindeki P_{CSA} matrisindeki satır sayısı
N	: Parametre optimizasyonlu CLONTagger yöntemindeki etiket SNP sayısı
PA	: Parametre optimizasyonlu CLONTagger yöntemindeki etiket SNP kümesinin tahmin doğruluğu
N_c	: Parametre optimizasyonlu CLONTagger yöntemindeki doğru olarak tahmin edilen SNP sayısı
N_a	: Parametre optimizasyonlu CLONTagger yöntemindeki toplam tahmin edilen SNP sayısı
β	: Parametre optimizasyonlu CLONTagger yöntemindeki klonlama faktörü
M_R	: Parametre optimizasyonlu CLONTagger yöntemindeki mutasyon oranı
M	: Parametre optimizasyonlu CLONTagger yöntemindeki P_{CSA} popülasyon matrisindeki her bir antikordaki etiket SNP sayısı
N_G	: Parametre optimizasyonlu CLONTagger yöntemindeki CLONALG için jenerasyon sayısı
TS	: Parametre optimizasyonlu CLONTagger yöntemindeki etiket SNP'lerin kümesi
P_{PSO}	: Parametre optimizasyonlu CLONTagger yöntemindeki PSO için popülasyon matrisi
pbest	: Parametre optimizasyonlu CLONTagger yöntemindeki PSO için bir parçacığın o ona kadar ulaştığı en iyi değer
gbest	: Parametre optimizasyonlu CLONTagger yöntemindeki PSO için popülasyondaki global en iyi değer
c1, c2	: Parametre optimizasyonlu CLONTagger yöntemindeki PSO için öğrenme faktörleri
r1, r2	: Parametre optimizasyonlu CLONTagger yöntemindeki PSO için üretilen rastgele değerler
w	: Parametre optimizasyonlu CLONTagger yöntemindeki PSO için atalet ağırlığı

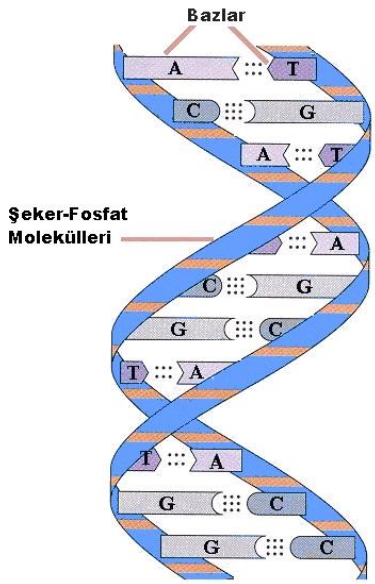
Kısaltmalar

DNA	: Deoxyribonucleic Acid - Deoksiribonükleik Asit
RNA	: Ribonucleic Acid - Ribonükleik Asit
SNP	: Single Nucleotide Polymorphism - Tek Nükleotid Polimorfizmleri
GA	: Genetic Algortihm - Genetik Algoritma
SVM	: Support Vector Machine - Destek Vektör Makinesi
CLONALG	: Clonal Selection Algorithm - Klonal Seçim Algoritması
PSO	: Particle Swarm Optimization - Parçacık Sürü Optimizasyonu
LOOCV	: Leave One-Out Cross Validation - Birini Dışarıda Bırak Çapraz Doğrulama
LD	: Bağlantı Dengesizliği – Linkage Disequilibrium
BN	: Bayes Ağı - Bayesian Network
STSA	: Stepwise Tag Selection Algorithm - Adım Adım Etiket Seçim Algoritması
LMT	: Local-Minimization Tag Selection Algorithm - Yerel Azalan Etiket Seçim Algoritması
BPSO	: Binary Particle Swarm Optimization – İkili Parçacık Sürü Optimizasyonu
RBF	: Radial Basis Function – Radyal Tabanlı Fonksiyon

1. GİRİŞ

Deoksiribonükleik Asit (DNA), bütün hücreli canlıların ve bazı virüslerin biyolojik gelişimleri için gerekli genetik bilgiyi taşıyan bir çeşit nükleik asittir. DNA, canlıların özelliklerinin soydan soya geçmesini sağladığı için bazen "kalıtım molekülü" olarak da adlandırılır (Tekşen, 2006).

DNA aslında tek bir molekül değil, bir çift moleküldür. Bu çift molekül, bir sarmaşığın dalları gibi birbiri çevresinde dönerek bir sarmal oluşturur. Sarmaşık dalına benzeyen her bir molekül, bir DNA ipliğidir (Şekil 1.1). Bu iplikler birbirlerine kimyasal olarak bağlanmış nükleotidlerden oluşur. Nükleotidler ise bir şeker, bir fosfat ve bir de azotlu bazdan oluşur. Bu bazlar dört çeşittir ve *A*, *T*, *C* ve *G* harfleri ile gösterilen Adenin, Timin, Sitozin ve Guanin'dir (Claverie ve Notredame, 2007).



Şekil 1.1. DNA çift sarmalı

DNA'nın özel bir şekilde paketlenmesi sonucu ortaya çıkan kılıflara kromozom denir. İnsan genomu 23 çift kromozomdan oluşur ve bu çiftlerden birisi anneden diğeri ise babadan gelir. Kromozomlar içerisinde çok sayıda gen vardır. Genler, canlı bireylerin kalıtsal karakterlerini taşıyıp ortaya çıkışını sağlayan ve nesilden nesile aktaran kalıtım faktörleridir. Her bir gen diğerinden farklı bir şifre içerir ve farklı bir proteini kodlar. Eğer vücutta bir genin kodladığı proteine gereksinim varsa o gen aktif hale geçerek üzerindeki şifre, haberci ribonükleik asit (RNA) adı verilen bir yapı

şeklinde kopyalanır. Bu yapı hücrenin sitoplâzmasındaki ilgili birimlere gelerek kalıp vazifesi görür ve o proteinin yapımı sağlanır (Claverie ve Notredame, 2007).

Genler içerdikleri şifreler aracılığıyla vücuttaki her türlü olayı uzaktan kumanda sistemi sayılabilecek bir duyarlılıkla kontrol ederler. Bazı genler vücuda gerekli kimyasal yapıların ortaya çıkmasını sağlarken bazı genler ise diğer genler üzerinde düzenleyici olarak şifrelenmiştir. Yönetici moleküller olan genler; insanın fiziksel özelliklerini, vücutta hangi olayların gerçekleştiğini ve hangi hastalıkları geçirmeye eğilimli olduğunu belirler (Chanda ve ark., 2009).

İnsan genomu yaklaşık 3 milyar nükleotidden oluşur. Bir popülasyonu oluşturan insanların genomunun yaklaşık %99'u belli bir konumda aynı nükleotidden oluşur. Buna karşılık, bu insanların genomunun %1'i, farklı nükleotid oluşumları, bir nükleotidin silinmesi veya eklenmesi gibi genetik farklılıklar içerir. İnsanların, fiziksel görünüşlerindeki gibi belirgin veya belli bir hastalığa yatkınlıkları kadar gizli olan bu özellik farklılıkları, insan DNA'sındaki bu genetik farklılıklardan meydana gelir. Günümüze kadar, milyonlarca yaygın DNA farklılığı tanımlanmıştır. Bu farklılıkların çoğunluğunu, tek nükleotid değişimleri oluşturur ve *tek nükleotid polimorfizmleri* (*Single Nucleotide Polymorphism - SNP*) olarak isimlendirilirler (Lee ve Shatkay, 2006). Bugüne kadar insan genomunda çok sayıda SNP tanımlanmış durumdadır ve 11 milyon civarında olduğu tahmin edilmektedir (Kruglyak ve Nickerson, 2001).

Genetik hastalıklar, bir bireyin genetik materyali yani genomundaki bozukluklar sonucu ortaya çıkan hastalıklardır. Bu hastalıklar içerisinde en yaygın olarak karşılaşılan karmaşık hastalıklarla ilişkili genetik değişimlerin araştırılması, insan genomu üzerindeki güncel araştırma konularından bir tanesidir. Birçok genom çaplı ilişki çalışması ile karmaşık hastalıklarla ilişkili olabilecek genetik değişimler belirlenmeye çalışılmaktadır. Bu genetik değişimlerin büyük çoğunluğunu SNP'ler oluşturduğu için bu çalışmalarda öncelikli olarak kullanılmaktadır (Crawford ve Nickerson, 2005; Halldorsson ve ark., 2004a). Bir genom çaplı ilişki çalışmasının istatistiksel önemi, bireylerin ve SNP'lerin sayısı ile doğrudan ilgilidir. Ancak, çok büyük çaplı ilişki çalışmalarında, çok sayıdaki bireyler için aday bölge içindeki bütün SNP'leri genotiplemek hala oldukça maliyetli ve zaman alıcıdır. Bu nedenle küçük bir hata ile SNP'lerin geriye kalanını temsil edecek bütün SNP'lerin uygun bir alt kümesinin seçilmesi gerekir. Seçilen bu SNP'lere *etiket SNP* veya *haplotip etiket SNP* (tag SNP veya htSNP) denir. Etiket SNP seçiminde, çok iyi bir tahmin doğruluğuna

sahip, minimum büyüklükteki etiket SNP kümesinin bulunması esastır (Halperin ve ark., 2005).

Bu tez çalışmasında, etiket SNP seçim uygulamalarında kullanılmak üzere üç farklı metot geliştirilmiştir. Bu metotlardan ilkinde, etiket SNP seçim yöntemi olarak Genetik Algoritma (Genetic Algorithm - GA) ve geriye kalan SNP'lerin (etiket SNP haricindeki SNP'ler) tahmini için ise Destek Vektör Makinesi (Support Vector Machine - SVM) kullanılmış ve bu yöntem GA-SVM olarak adlandırılmıştır. GA-SVM metodunun tahmin doğruluğunun değerlendirilmesi için Birini Dışarıda Bırak Çapraz Doğrulama (Leave One-Out Cross Validation - LOOCV) yöntemi kullanılmıştır. Geliştirilen ikinci metotta ise GA-SVM metodunda sabit değer olarak kullanılan SVM'nin C ve γ parametrelerinin optimizasyonu için Parçacık Sürü Optimizasyon (Particle Swarm Optimization - PSO) algoritması kullanılmış ve bu metot Parametre Optimizasyonlu GA-SVM yöntemi olarak adlandırılmıştır. Bu tez çalışmasında geliştirilen son yöntemde ise Parametre Optimizasyonlu GA-SVM yöntemindeki etiket SNP seçim metodu olarak kullanılan GA yerine, Klonal Seçim Algoritması (Clonal Selection Algorithm - CLONALG) kullanılmış ve Parametre Optimizasyonlu CLONTagger yöntemi olarak adlandırılmıştır. Parametre Optimizasyonlu CLONTagger yöntemindeki mutasyon işleminde, antikorların antijene olan duyarlılıkları ile doğru orantılı olarak çalışan bir mekanizma kullanılmıştır.

Tezin geriye kalan kısmı şu şekilde organize edilmiştir: İkinci bölümde, etiket SNP seçimiyle ilgili olarak literatürde daha önce yapılmış çalışmalar anlatılmıştır. Üçüncü bölümde genetik hastalıklar ve genom çaplı analizlerden bahsedilmiş, etiket SNP seçiminin sayısal haplotip analizi içerisindeki yeri anlatılmış ve konuyla ilgili genel kavramlar tanıtılmıştır. Dördüncü bölümde ise geliştirilen yöntemlerin performanslarını değerlendirmek için kullanılan veri kümeleri incelenmiş ve kullanılan yöntemler hakkında özet bilgiler verilmiştir. Geliştirilen GA-SVM, Parametre Optimizasyonlu GA-SVM ve Parametre Optimizasyonlu CLONTagger yöntemlerinin detayları beşinci bölümde incelenmiştir. Geliştirilen yöntemlerin farklı veri kümeleri üzerinde elde edilen uygulama sonuçları altıncı bölümde sunulmuştur. Tez çalışmasıyla ilgili genel sonuçlar ve öneriler ise yedinci bölümde açıklanmıştır.

2. KAYNAK ARAŞTIRMASI

Etiket SNP seçimi alanında yapılan ilk çalışmalarda SNP'ler arasında mevcut olan bağlantı dengesizliği (Linkage Disequilibrium - LD) ölçüsü dikkate alınmıştır (Goldstein, 2001). Eğer SNP'ler arasındaki LD değeri yüksek ise bu SNP'lerin allel bilgisi hemen hemen aynıdır. Bu nedenle, bu SNP'lerden sadece biri haplotip içindeki diğer SNP'leri temsilen seçilebilir (Weale ve ark., 2003).

Araştırmacılar, haplotip bilgisini temsil etmek için farklı ölçüler önermişler ve bu ölçüleri optimize edecek SNP'lerin alt kümesini belirlemeyi hedeflemişlerdir. Bu ölçüler arasındaki ilişkiler ve etiket SNP'lerin seçimi üzerindeki etkileri, hâlâ devam eden bir araştırma konusudur. Etiket SNP seçimi konusundaki algoritmalar, haplotip bilgisini ölçerken dikkate aldıkları yaklaşıma dayanarak haplotip farklılığına, SNP'ler arasındaki ilişki katsayısına, etiketlenen SNP tahminine ve fenotip ilişkisine olmak üzere dört gruba ayrılır.

İnsan genomunun blok yapısı üzerindeki son gözlemler, genomun farklı bloklara bölünebileceğini ve bir popülasyonun büyük bir kısmının (%80-%90) her bir blok içinde çok küçük sayıda yaygın haplotipleri (3-5 haplotip) paylaştığını göstermiştir (Daly ve ark., 2001; Gabriel ve ark., 2002; Patil ve ark., 2001). Bu varsayıma dayanarak ilk etiket SNP seçim algoritmaları, orijinal veri içerisindeki sınırlı haplotip farklılığının büyük bölümünü yakalayabilen SNP alt kümesini bulmayı hedeflemişlerdir.

Çeşitli haplotip farklılığı ölçüleri önerilmiştir. Bazı araştırmalarda *T'* ile belirlenen haplotip farklılığı ölçüsü olarak, *T'* aday alt kümesi tarafından benzersiz olarak ayırt edilebilen haplotiplerin sayısı kullanılmıştır (Ke ve Cardon, 2003; Patil ve ark., 2001). Johnson ve ark. (2001) ise, *T'* aday alt kümesi ile belirlenemeyen haplotip farklılığını (*T'* kümesinin kalan haplotip farklılığı) tanımlamışlardır. Bu haplotip farklılığı, *T'* kümesine göre aynı grup içindeki her bir haplotip çifti arasındaki allel farklılıklarının sayısı olarak tanımlanmıştır. Eğer *T'* aday alt kümesi bütün farklı haplotipleri farklı gruplara bölerse onun kalan haplotip farklılığı 0 olur. Diğer taraftan, farklı haplotipler aynı grup içerisine yerleştirilirse onun kalan haplotip farklılığı 0'dan büyük olur. Bu nedenle, en küçük kalan haplotip farklılığı değerine sahip *T'* kümesi etiket SNP kümesi olarak seçilir.

Diğer bir popüler haplotip farklılığı ölçüsü ise Shannon Entropisidir (*H*) (Ackerman ve ark., 2003; Avi-Itzhak ve ark., 2003; Hampe ve ark., 2003; Judson ve ark., 2002). *n'*, *D* haplotip veri kümesi içindeki farklı haplotiplerin sayısı ve p_i de *i.*

farklı haplotipin frekansı olarak kabul edilirse *D*'nin haplotip farklılığı onun entropisi (*H*) olarak hesaplanır.

$$H(D) = -\sum_{i=1}^{n'} p_i log_2 p_i \qquad (2.1)$$

Her bir *T'* aday etiket SNP kümesi için haplotipler, gruplara bölünür ve aynı grup içinde olan haplotipler, *T'* kümesinin elemanı olan SNP'lerle aynı allelleri paylaşır. *D* veri kümesinin entropisi, bu bölüme dayanarak ölçülür. Aynı grup içinde yer alan haplotiplerin benzer oldukları düşünülür. Böylece, farklı haplotiplerin sayısı (*n'*) grupların sayısı olur ve *i.* farklı haplotipin frekansı (p_i), *i.* gruptaki haplotiplerin sayısının haplotiplerin toplam sayısına oranıdır. *T'* aday alt kümesi ne kadar çok grup belirlerse gruplamaya dayanan *D* veri kümesinin entropiside o kadar büyük olur. Sonuçta en büyük entropili *T'* kümesi bir çözüm olarak seçilir.

Yukarıda bahsedilen metotlar (Ackerman ve ark., 2003; Avi-Itzhak ve ark., 2003; Hampe ve ark., 2003; Johnson ve ark., 2001; Judson ve ark., 2002; Ke ve Cardon, 2003; Patil ve ark., 2001) S orijinal SNP kümesinin bütün alt kümelerini kapsamlı olarak sınarlar. Bu işlem az sayıdaki SNP'lere bile bu yöntemlerin uygulanabilirliğini sınırlar. Bu problemin üstesinden gelmek için açgözlü (greedy) algoritma (Sheng ve ark., 2004), dal ve sınır (branch and bound) kuralı (Ding ve ark., 2005a), dinamik (dynamic) programlama (Zhang ve ark., 2002a; Zhang ve ark., 2002b; Zhang ve ark., 2003; Zhang ve ark., 2004; Zhang ve ark., 2005) ve temel bileşen analizi (principal component analysis) (Horne ve Camp, 2004; Lin ve Altman, 2004; Meng ve ark., 2003) gibi çeşitli sezgisel ve etkili arama metodları geliştirilmiştir.

İlişki katsayısı tabanlı yaklaşımlar, etiket SNP'lerin kümesinin bir haplotip üzerindeki bir hastalık lokusunu tahmin edebilecek SNP'lerin en küçük alt kümesi olması gerektiği fikrine dayanırlar. Ancak hastalık lokusu, genellikle bizim aradığımız yerdir ve daha önceden bilinmez. Bu nedenle SNP'ler arasındaki ilişki katsayısı tahmin için kullanılır. Prensipte bir haplotip üzerindeki bütün SNP'ler, etiket SNP'lerden en az biri ile yüksek oranda ilişkilidir. Bu nedenle bir hastalık ile ilişkili SNP, etiket SNP olarak seçilmese dahi bu SNP ile hastalık arasındaki ilişki yüksek oranda ilişkili olduğu etiket SNP'den dolaylı olarak çıkarılabilir. Birçok çalışma içinde, SNP'lerin rastgele olmayan bu ilişkisi, ilişki katsayısını tahmin etmek için kullanılmıştır.

Byng ve ark. (2003) ilişki katsayısı tabanlı etiket SNP seçimi için küme analizini kullanmayı önermişlerdir. Bu çalışmada, SNP'lerin orijinal kümesi hiyerarşik kümelere bölünür ve aynı küme içindeki SNP'ler diğer SNP'lerin en az biri ile en az σ ($\sigma > 0.6$-0.8) ön tanımlı seviyesinde LD ilişkisine sahiptir. Kümeleme uygulandıktan sonra genotiplemenin kolaylığı, fiziksel konumunun önemi veya SNP mutasyonunun önemi gibi pratik uygulanabilirliğine dayanarak her bir kümeden bir SNP seçmeyi tavsiye etmişlerdir.

İlişki katsayısı kullanılarak yapılan diğer çalışmalarda (Ao ve ark., 2005; Carlson ve ark., 2004; Wu ve ark., 2003), bir küme içerisinde diğer bütün SNP'lere göre LD ilişkisi σ sabit değerinden daha büyük olan SNP'in etiket SNP olarak seçilmesi gerektiği önerilmiştir. Bu şekilde etiket SNP'leri seçmek için minimax kümeleme (Ao ve ark., 2005) ve greedy binning algoritması (Carlson ve ark., 2004; Wu ve ark., 2003) kullanılmıştır.

Ao ve ark. (2005) tarafından önerilen minimax kümelemede, C_i ve C_j kümeleri arasındaki minimax mesafesi aşağıdaki formül ile hesaplanmıştır.

$$D_{minimax}(C_i, C_j) = \min_{\forall s \in (C_i \cup C_j)} (D_{max}(s)) \qquad (2.2)$$

Burada $(D_{max}(s))$, s SNP'i ile iki küme içerisindeki diğer bütün SNP'ler arasındaki maksimum mesafedir. Başlangıçta her SNP kendi kümesini oluşturur. Daha sonra minimax mesafelerine göre en yakın olan iki küme iteratif olarak birleştirilir. Birleştirme işlemi, iki küme arasındaki en küçük mesafe σ değerinden büyük olduğunda durur. Son olarak her bir birleştirilen kümenin minimax mesafesini tanımlayan SNP, küme temsilcisi olarak seçilir.

Greedy binning algoritmasında ise ilk olarak SNP'ler arasındaki LD ilişki katsayısı hesaplanır ve her bir SNP için σ değerinden daha büyük LD değerine sahip olan SNP'lerin sayısı hesaplanır. Daha sonra en büyük sayıya sahip olan SNP, ilişkili olduğu SNP'ler ile birlikte kümelenir ve bu SNP küme için etiket SNP olur. Bu prosedür, bütün SNP'ler kümeleninceye kadar geriye kalan bütün SNP'ler için tekrarlanır. Diğer SNP'ler ile σ'dan daha büyük LD ilişkisine sahip olmayan SNP'ler bireysel kümeler olarak düşünülür.

Fenotip ilişki tabanlı yaklaşımlar ise fenotip bilgisinin mevcut olduğunu kabul eder ve sağlıklı bireylerden hastalıklı bireyleri ayırt edebilecek SNP kümesini bulmayı

denerler. Bu SNP'ler daha sonra etiket SNP kümesi olarak kullanılır. Bu bakış açısıyla etiket SNP seçimi, küçük bir hata ile iki sınıf (hastalıklı/sağlıklı) arasındaki ayırt edilebilir özellikler kümesini seçmeyi hedefleyen bir çeşit özellik seçimidir. Sonuç olarak sınıf etiketleri ve özellikler arasındaki ilişkiyi ölçen sınıflama teknikleri ve istatistiksel testler bu bağlamda kullanılmıştır.

Tsalenko ve ark. (2003) konuyla ilgili yaptıkları çalışmada en popüler sınıflandırıcılardan biri olan Bayes sınıflandırıcısını kullanmışlardır. Bayes sınıflandırıcısı, bir bireyin fenotipi verildiğinde bir SNP'in allelinin diğerlerinin allelinden koşullu olarak bağımsız olduğunu kabul eder ve bu sınıfların her birine ait olma olasılığına dayanarak hastalıklı veya hastalıksız olarak her bir haplotipi sınıflar. Daha sonra en iyi sınıflama doğruluğuna sahip SNP'lerin *T'* alt kümesi, etiket SNP'lerin alt kümesi olarak seçilir. Bu yaklaşımın ana sınırlaması, Bayes sınıflandırıcısı tarafından kullanılan SNP'ler arasındaki koşullu bağımsızlık varsayımıdır. Gerçekte SNP'ler arasında rastgele olmayan ilişki (LD) mevcuttur (Horne ve Camp, 2004). Shah ve Kusiak (2004), SNP'ler arasındaki korelâsyonu dikkate alan bir özellik seçim metodu kullanarak bu problemi çözmüşlerdir. Onların algoritması, bir özellik eğer henüz seçilmiş olan herhangi bir diğer özellik ile değil bir hedef sınıf etiketi ile ilişkili ise o özelliği seçer.

Yukarıda bahsedilen fenotip ilişki tabanlı metotlar, verilen veriyi hastalıklı ve hastalıksız olmak üzere iki sınıfa doğru olarak bölen SNP'lerin kümesini seçmeye odaklanmışlardır. Hoh ve ark. (2000), verilen veriyi sadece sınıflayan değil aynı zamanda önemli oranda istatistiksel olarak onun performansını garanti eden bir metot önermişlerdir. Onların algoritması bootstrap tekniğine dayanır (Efron, 1982). Bu tekniğe göre orijinal verinin *n* tane haplotip-fenotip çiftinden oluştuğu kabul edilirse ilk olarak orijinal veriden kopyalanarak bir *A* kopya kümesi oluşturulur. Daha sonra, $B_1,\ldots,B_{1000}$ ilave kopya kümeleri oluşturulur. Buradaki her bir küme, *A* kopya kümesinden kopyalanır ve her biri *n* tane haplotip-fenotip çiftini içerir. 1000 kopya küme, haplotip ve fenotipler arasında herhangi bir ilişkinin olmadığı örnekleri temsil ederler. Böylece, onların haplotip ve fenotip etiketleri rastgele olarak değiştirilir. Son olarak, $B_1,\ldots,B_{1000}$ rastgele örneklerinin en az $(1-\alpha)x100$ (genelde $\alpha=0.05$) yüzdesinde ilişki skoru toplamları *A* kümesi içindeki SNP'lerinkinden daha yüksek olanlar seçilir. Bu işlem önceden tanımlanan sayı kadar tekrarlanır ve iterasyonların en az %50'sinde seçilen SNP'ler etiket SNP kümesi olarak belirlenir.

Etiket SNP seçiminde kullanılan diğer bir yaklaşım ise etiketlenen SNP tahmin tabanlı yöntemlerdir. Bu yöntemler etiket SNP seçimini sadece SNP'lerin belli bir kümesini kullanarak orijinal haplotip verisinin yeniden oluşturulması problemi olarak dikkate almışlardır. Böylece bu yöntemler küçük bir hata ile seçilmeyen (etiketlenen) SNP'leri tahmin edebilecek SNP'lerin kümesini seçmeyi amaçlamışlardır. Genellikle, seçilen etiket SNP'ler genotiplendikten sonra etiketlenen SNP'lerin allelleri etiket SNP'lerin allelleri kullanılarak tahmin edilir ve hastalık gen ilişkisi bütünüyle oluşturulan haplotip verisine dayanılarak yönetilir. Bu nedenle bu yöntemler, seçilen etiket SNP kümesi ile birlikte etiketlenen SNP kümesi için bir tahmin kuralı da sunmuşlardır.

Bafna ve ark. (2003) ilk olarak, etiketlenen SNP'leri tahmin etmek için onların tahmin doğruluğuna dayanarak etiket SNP'leri seçmeyi önermişlerdir. $E_{i,j}^{t}$, h_i ve h_j haplotiplerinin t SNP'nde farklı allellere sahip olma ifadesi olduğu kabul edilirse, $S=\{s_1,\ldots,s_k\}$ SNP kümesinin t SNP'ini nasıl tahmin edebileceğini ölçmek için bilgilendiricilik olarak adlandırılan bir ölçü tanımlanmıştır.

$$I(S,t) = Pr_{i\neq j}\left(\bigcup_{l=1}^{k} E_{i,j}^{s_l} \middle| E_{i,j}^{t}\right) \tag{2.3}$$

Önerilen ölçüye dayanarak, geri kalan SNP'leri en iyi tahmin edebilecek en uygun SNP alt kümesi dinamik programlama kullanılarak belirlenmiştir. Bafna ve ark. (2003) her bir etiketlenen SNP'in tahmin edici etiket SNP'lerinin diğerlerine göre w fiziksel yakınlığı içinde olmasını sağlamışlardır.

Halperin ve ark. (2005) etiket SNP seçimi için en iyi çözümü garanti eden polinom zamanlı bir dinamik programlama (STAMPA) ve daha basit ve hızlı olan rastgele örnekleme algoritmasını önermişlerdir. Her iki etiket SNP seçim algoritması da bir alt program olarak aynı tahmin algoritmasını kullanır. Tahmin algoritması, iki SNP'in genotip değerlerinin verildiği ve bu iki SNP arasındaki SNP'lerin değerlerinin bu SNP'ler dikkate alınarak tahmin edilebileceği varsayımına dayanır. Başka bir ifadeyle, bir SNP'in değeri hem sol hem de sağındaki iki en yakın etiket SNP'in değerleri dikkate alınarak tahmin edilir. Bilinmeyen SNP'in değeri ise iki etiket SNP üzerinde yapılan çoğunluk oylaması ile tahmin edilir. Pratikte, SNP'ler arası korelasyon SNP'ler arasındaki mesafe artarken azalır. Bu nedenle komşu etiket SNP'ler arasındaki

mesafenin çok büyük olması zayıf bir tahmin gücü sağlar. Birçok pratik uygulamada, komşu etiket SNP'ler arasındaki SNP sayısı cinsinden maksimum mesafe için bir *c* sınırı kullanılır. STAMPA için etiket SNP'ler arasındaki mesafenin üst sınırı olarak *c=30* kullanılmıştır. Her iki etiket SNP seçim algoritmasının tahmin doğruluğu LOOCV yöntemi kullanılarak değerlendirilmiştir. 5q31 (Daly ve ark., 2001), ENm013, ENr112, ENr113, PP2R4, STEAP ve TRPM8 (International HapMap Consortium, 2003) veri kümeleri üzerinde yapılan deneylerde, STAMPA'nın rastgele örnekleme algoritmasına göre daha iyi sonuçlar verdiği gösterilmiştir. STAMPA metodu ile 2 etiket SNP için 5q31, STEAP ve TRPM8 veri kümeleri üzerinde sırasıyla %80.00, %95.00 ve %82.57 tahmin doğruluğu elde edilmiştir.

Lee ve Shatkay (2006) etiket SNP seçimi için yeni bir metot önermişlerdir. BNTagger ismini verdikleri yeni metot, SNP'ler arasındaki tahmin eden-tahmin edilen ilişkisini belirlemek için Bayes ağları aracılığı ile SNP'ler arasındaki koşullu bağımsızlığı kullanır. Bayes ağları daha önceleri haplotip blok bölme (Greenspan ve Geiger, 2003) ve haplotip fazlama (Xing ve ark., 2004) için kullanılmış olmasına rağmen etiket SNP seçim problemine ilk defa bu çalışmada uygulanmıştır. BNTagger metodu, SNP'ler arasındaki koşullu bağımsızlık ilişkilerinin belirlenmesi, iki sezgisel yöntem kullanılarak etiket SNP'lerin seçilmesi ve yeni genotiplenen örnekler için tam haplotiplerin oluşturulması aşamalarını kullanmıştır. Bu algoritma, etiketlenen SNP'ler için tahmin doğruluğunu maksimize edecek etiket SNP'leri seçmeyi hedeflemiştir. Ayrıca, biallelik olmayan SNP'ler de bu algoritma tarafından kullanılabilir. BNTagger algoritması yeni genotiplenen örnekler için etiket SNP'lerin genotip verisini kullanarak bütün SNP'lerin haplotip verisini doğrudan oluşturabilir. ACE (Reider ve ark., 1999), LPL (Clark ve ark., 1998) ve 5q31 veri kümeleri üzerinde yapılan deneylerde 2 etiket SNP için sırasıyla %89.00, %80.00 ve %89.10 tahmin doğruluğu elde edilmiştir. Yapılan deneylerin tahmin doğrulukları, LOOCV çapraz doğrulama yöntemi kullanılarak değerlendirilmiştir.

He ve Zelikovsky (2006, 2007) etiket SNP seçimi için iki ve SNP tahmini için de iki farklı algoritma önermişlerdir. Etiket SNP seçimi için kullanılan algoritmalardan ilki olan adım adım etiket seçim algoritması (STSA), SNP kümesi içerisinde hatayı minimum yapacak en iyi t_0 etiketini bulur. Daha sonra t_0 etiketinin en iyi uzantısı olan t_1 etiketini bulur ve bu işlem *k* ile gösterilen etiket SNP sayısına ulaşılıncaya kadar devam eder. İkinci etiket SNP seçim algoritması ise yerel azalan etiket seçim algoritmasıdır (LMT). Bu algoritma çok büyük olasılıklar arasında etiketlerin daha iyi kümesini doğru

olarak arar. LMT, STSA tarafından üretilen *k* etiketler ile başlar ve diğerlerini sabit bırakarak her bir tek etiketi iteratif olarak değiştirerek olası en iyi seçimi yapar. Bu yer değiştirme işlemi tahmin doğruluğunda önemli bir gelişme olmayıncaya kadar devam eder. Bu çalışmada önerilen SNP tahmin algoritmalarından ilki destek vektör makinesine dayanan SNP tahmin algoritmasıdır. Bu algoritmada eğitim kümesi olarak verilen haplotiplere göre inşa edilen model, test kümesi olarak verilen haplotipteki bilinmeyen SNP değerlerini tahmin etmekte kullanılır. Diğer SNP tahmin algoritması ise çoklu doğrusal regresyona dayanan algoritmadır. Yazarlar, STSA tabanlı SNP seçiminin hemen hemen aynı sonuçları üretmesine rağmen LMT'den çok daha hızlı olduğunu ve SVM tabanlı SNP tahmininin de MLR'den daha iyi sonuçlar ürettiğini göstermişlerdir. 5q31, STEAP ve TRPM8 veri kümeleri üzerinde iki etiket SNP için SVM/STSA algoritmasında sırasıyla %89.32, %98.18 ve %90.50 tahmin doğruluğu elde edilmiştir.

Yang ve ark. (2008) etiket SNP seçimi için ikili parçacık sürü optimizasyon (BPSO) algoritmasının özellik seçim fikrini kullanmışlardır. Her bir parçacığın pozisyonu bir ikili dizi formunda verilir. Bu ikili formlarda, 1'ler etiket SNP'leri, 0'lar ise etiketlenen SNP'leri temsil eder. Yazarlar, SNP tahmin yöntemi olarak STAMPA ile aynı tahmin prosedürünü kullanmışlardır. Her bir parçacığın tahmin doğruluğunun değerlendirilmesi için de LOOCV yöntemi kullanılmıştır. Bu çalışmada, parçacık sürü optimizasyon algoritmasının standart prosedürlerine ilave olarak parçacıkların güncelleştirilmesi işleminden sonra parçacık düzeltme prosedürü kullanılmıştır. Bu prosedür ile güncelleştirilen parçacıklardaki etiket SNP sayılarının algoritmaya giriş olarak verilen etiket SNP sayısına eşitlenmesi sağlanmıştır. Bu eşitleme işlemi sırasında, en iyi etiket SNP kümesini bulmak için yerel arama algoritması kullanılmıştır. Bu algoritma LPL, STEAP ve 10 popülasyon D (Gabriel ve ark., 2002) veri kümesi üzerinde çalıştırılmıştır. İki etiket SNP için LPL ve STEAP veri kümelerinde sırasıyla %84.60 ve %90.67 oranında tahmin doğrulukları elde edilmiştir.

Lin ve Leu (2010), parçacık sürü optimizasyonu ve destek vektör makinesini birleştiren hibrit PSO-SVM yaklaşımını önermişlerdir. Bu algoritmada PSO etiket SNP'leri seçmek için SVM ise etiketlenen SNP'leri tahmin etmek için kullanılmıştır. Her bir parçacık, seçilen etiket SNP'ler (özellikler) ve SVM'nin C ve γ parametrelerinden oluşur. Böylece özellik seçimi ve parametre optimizasyonu aynı anda yapılmış olur. PSO öğrenme işlemi sırasında, Yang ve ark. (2008) tarafından önerilen düzeltme prosedürü bu algoritmada kullanılmamıştır. Burada, PSO öğrenme işleminde,

daha büyük sayılabilecek parçacıkların 1'e doğru, daha küçük sayılabilecek parçacıkların ise 0'a doğru genişleme eğiliminde olduğu kabul edilir. Bu nedenle, PSO'daki güncelleme işleminden sonra, etiket SNP'lerin sayısını algoritmaya giriş olarak verilen sayıya eşitlemek için bu eğilimler dikkate alınır. Parçacıkların tahmin doğruluğunu değerlendirmek için LOOCV yöntemi kullanılır. Yazarlar tarafından geliştirilen hibrit PSO-SVM metodu 5q31, TRPM8 ve LPL veri kümelerine uygulanmış ve deneyler sonucunda iki etiket SNP için sırasıyla %90.07, %90.00 ve %81.00 oranında tahmin doğrulukları elde edilmiştir.

Mahdevar ve ark. (2010) GTagger olarak adlandırılan yeni bir metot önermişlerdir. Bu yöntemde etiket SNP seçimi için genetik algoritma kullanılmıştır. Uygunluk değerlendirme fonksiyonu olarak Shannon entropi fonksiyonu ile birlikte tanımladıkları özel bir fonksiyonu kullanmışlardır. Tanımlanan özel fonksiyon, mümkün olan en az etiket SNP ile en fazla etiketlenen SNP'in tahmin edilmesi prensibine dayanır.

$$f(I) = \frac{M(I)}{Log\ |I| + 1} \tag{2.4}$$

Burada *M(I)* ve $\left|I\right|$, bir *I* bireyi için sırasıyla tahmin edilen SNP'lerin sayısı ve etiket SNP'lerin sayısıdır. Uygunluk fonksiyonu, Shannon entropi fonksiyonu ile *f(I)* fonksiyonunun ortalaması alınarak hesaplanır. GTagger algoritması yapay ve Avrupa popülasyonundan alınan kromozom 21 (Patil ve ark., 2001) veri kümelerine uygulanmış ve bütün durumlarda çalışma süresi açısından karşılaştırılan diğer metotlardan daha yüksek performans göstermiştir.

İlhan ve ark. (2011a) önerdikleri yöntemde, etiket SNP seçimi için klonal seçim algoritmasını SNP tahmini için ise SVM'yi kullanmışlardır. CLONTagger ismi verilen yöntemde farklı bir mutasyon işlemi kullanılmıştır. Mutasyon oranı belirlenirken, antikorların antijenle olan benzerlikleri ile ters orantılı olarak mutasyon işlemi uygulanmıştır. Etiket SNP'lerin sayısını algoritmaya giriş olarak verilen sayıya eşitlemek için bir düzeltme prosedürü kullanılmıştır. Bu prosedürde en iyi antikoru seçmek için 10 kat çapraz doğrulama prosedürü kullanılmıştır. Algoritmanın uygunluk değerlendirmesi LOOCV metodu ile yapılmıştır. CLONTagger metodu iki etiket SNP için ACE, ABCB1 (Kroetz ve ark., 2003) ve LPL veri kümeleri üzerinde sırasıyla %90.40, %97.00 ve %92.90 oranında tahmin doğruluğu sağlamıştır.

İlhan ve ark. (2011b) önerdikleri diğer bir yöntemde, SVM'yi SNP tahmini için ve genetik algoritmayı da etiket SNP seçimi için kullanmışlardır. GA-SVM ismini verdikleri yöntemde, 10 kat çapraz doğrulama metodunu içeren düzeltme prosedürü kullanılmıştır. Bu prosedür ile en iyi bireyin seçilmesi sağlanmıştır. LOOCV metodu ile yapılan değerlendirmelerde, LPL ve 10 popülasyon D veri kümeleri için karşılaştırılan diğer yöntemlerden daha yüksek tahmin doğrulukları elde edilmiştir. Örneğin iki etiket SNP için LPL veri kümesi üzerinde %92.90, popülasyon D'nin D9 veri kümesi üzerinde ise %84.40 oranında tahmin doğruluğu sağlanmıştır.

3. SAYISAL HAPLOTİP ANALİZİ

3.1. Genetik Hastalıklar

Yönetici moleküller olan genler, insanın fiziksel özelliklerini, vücutta hangi olayların gerçekleştiğini ve hangi hastalıkları geçirmeye eğilimli olduğunu belirlerler. Herhangi bir hastalığın oluşumunda hem çevresel hem de genetik etkenler rol oynar. Genetik hastalıklar, bir bireyin genetik materyali yani genomundaki bozukluklar sonucu ortaya çıkan hastalıklardır ve dört gruba ayrılırlar.

3.1.1. Tek gen (monogenik) hastalıkları

Bu tip hastalıklar, bir genin DNA dizisindeki değişikliği veya mutasyonu sonucu oluşurlar. Böylece bu genin kodladığı proteinin fonksiyonu bozulur ve bir hastalık ortaya çıkar. 6000'den fazla bilinen tek gen hastalığı vardır. Renk körlüğü, hemofili A, Marfan sendromu gibi hastalıklar tek gen hastalıklarına örnek olarak verilebilirler.

3.1.2. Çok etkenli (poligenik) hastalıklar

Bu tip hastalıklar, çevresel faktörler ve birçok gende meydana gelen bozuklukların birlikte etkileşimiyle oluşurlar. Örneğin, meme kanserine yatkınlık sağlayan genler değişik kromozomların üzerinde bulunmaktadır. Karmaşık yapıları nedeniyle bu tip hastalıkların analizi, tek gen ve kromozomal hastalıklara göre çok daha zordur. Kalp hastalıkları, yüksek tansiyon, Alzheimer hastalığı, artrit, diabet, kanser ve obezite bu grubun önemli hastalıkları arasındadır.

3.1.3. Kromozomal hastalıklar

Kromozomlar, her hücrenin çekirdeğinde birbirinden ayrı olarak yerleşmiş DNA ve proteinden oluşan yapılardır ve genetik materyalin taşıyıcıları oldukları için, bu yapılarda ortaya çıkan kayıplar, fazlalıklar, kırıklar ve yeniden birleşmeler hastalıkla sonuçlanabilir. Bazı önemli kromozomal düzensizlikler mikroskopik inceleme ile belirlenebilir. Down sendromu veya trizomi 21, bir insanda üç kopya kromozom 21 olduğu zaman oluşan önemli bir hastalıktır.

3.1.4. Mitokondriyal hastalıklar

Nispeten seyrek görülen bu tip genetik hastalıklar, mitokondrinin kromozomal olmayan DNA'sındaki mutasyonlar (değişiklikler) sonucu ortaya çıkarlar. Bazı nörolojik hastalıklar, kas ve göz hastalıkları bu gruba girerler.

3.2. Genom Çaplı Analizler

Genom çaplı yaklaşımlar ileri analiz için potansiyel genetik ilişkilerin büyük bir kümesini üretir. Birçok karmaşık hastalık üzerinde devam eden bu birlikteliğin, tek fonksiyon bozukluklu genlerden ziyade, bir hastalık fenotipinin temelini oluşturduğuna inanılır. Genom çaplı analizler bağlantı analizi (linkage analysis) ve ilişki çalışmaları (association studies) olmak üzere iki gruba ayrılır.

3.2.1. Bağlantı analizi

Bağlantı analizi, monogenik hastalıkların tanımlanması için kullanılır. Bu analizde fenotipik örüntüler (hastalık durumundaki gibi) ile genetik işaretler arasındaki ilişki tespit edilir.

Karmaşık genetik hastalıklar birkaç gen içerisindeki polimorfizmlerin birleştirilmiş etkisi ile oluşur ve bağlantı analizi ile bu genlerin tanımlanması büyük oranda başarısızdır. Çünkü her gen hastalık hassasiyetine küçük de olsa belli bir katkı sağlar.

3.2.2. İlişki çalışmaları

İlişki çalışmaları, bir hastalık popülasyonu içindeki allel frekansı ile bir kontrol popülasyonu içindeki allel frekansının karşılaştırılması ile çalışır. Bu iki popülasyon arasındaki önemli farklar, potansiyel olarak hastalık fenotipi ile ilişkili üzerinde düşünülmesi gereken lokusu gösterir. Bu ilişki doğrudan veya dolaylı olabilir. Doğrudan ilişki durumunda, polimorfizm hastalığa sebep olan veya etkileyen bazı fonksiyonlara sahiptir. Dolaylı ilişki durumunda ise polimorfizm hastalık için fonksiyonel olmayabilir. Fakat hastalığa sebep olan polimorfizme yakın olabilir.

Allellerin bu şekilde çok sık olarak birlikte görülmesi bağlantı dengesizliği (linkage disequilibrium) veya allelik ilişki (allelic association) olarak adlandırılır.

İlişki çalışmaları bağlantı analizi ile karşılaştırıldığında, küçük genetik etkileri keşfetme gücünün daha yüksek olması ve LD aralığının 10 kb civarında olması gibi bazı avantajlara sahiptir.

3.3. Sayısal Genetik Analizindeki Temel Kavramlar

Popülasyon genetiği, hem türler arasındaki hem de türler içindeki genetik değişimlerin evrimsel önemini anlamak için popülâsyonlardaki genetik değişimi çalışır (Hedrick, 2004). Böylece, yaygın ve karmaşık hastalık-gen ilişkisi için bir temel sağlar. Başka bir ifadeyle, bir insan popülasyonundaki yaygın DNA değişimlerinin bir kümesini tanımlamak, bir karmaşık hastalıkla ilgili nedensel bir bağlantıya sahiptir (Zhao ve ark., 2003). Sayısal haplotip analizinin esas amacı hastalık-gen ilişkisi olduğu için popülasyon genetiklerindeki bazı temel kavramların tanımlanması gerekir.

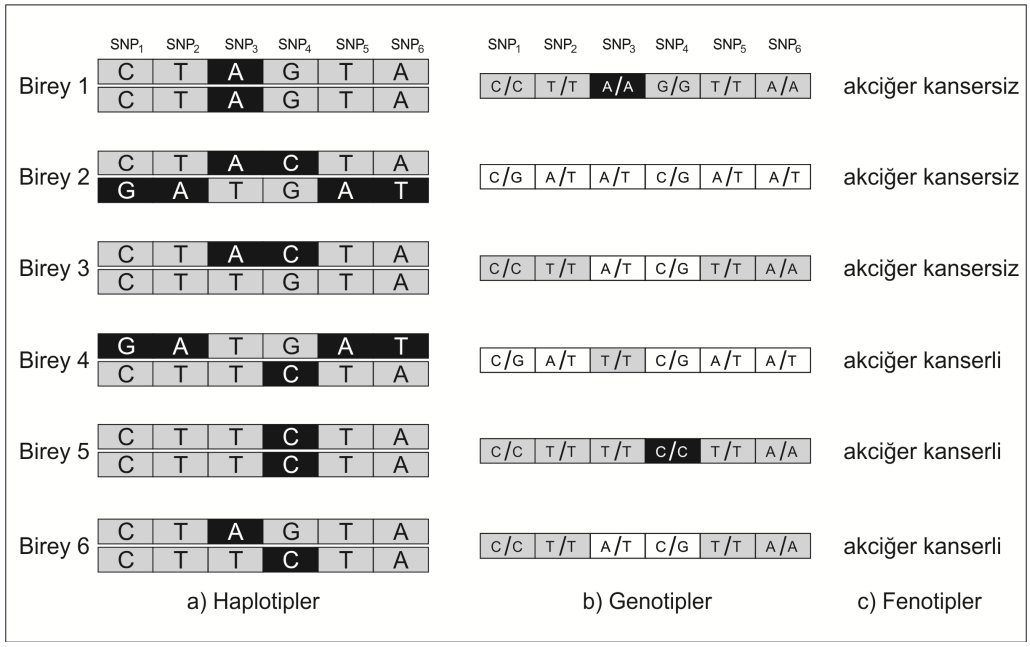
3.3.1. Haplotip, genotip ve fenotip

Şekil 3.1, üçü akciğer kanserli diğer üçü ise akciğer kansersiz olan altı bireyin kromozom örneklerini göstermektedir. Burada amaç, bu kromozom örneklerini kullanarak akciğer kanseriyle ilişkili DNA değişimlerinin kümesini tanımlamaktır. Maliyet ve zaman kısıtlaması nedeniyle, diğer moleküler deneyler tarafından akciğer kanseri ile ilişkili olduğu öncelikle ileri sürülen bir kromozomun sadece sınırlı bir bölgesi incelenir. Hedef bölgenin kromozomsal konumuna *lokus* adı verilir. Bir lokus, bütün bir kromozom kadar büyük veya bir genin bir parçası kadar küçük olabilir (Lee ve Shatkay, 2009).

Cinsel yolla çoğalan bütün türler, biri anneden diğeri ise babadan kalıtımla kazanılan bir çift kromozoma sahiptir. Bu nedenle, örnekteki her bir birey her bir SNP için biri anneden diğeri ise babadan gelen iki allele sahiptir. Bu iki kromozomdaki alleller ya aynıdır ya da farklıdır. Eğer onlar aynı ise ilgili SNP *homozigos*, farklı ise *hetorozigos* olarak adlandırılır (Lee ve Shatkay, 2009).

Şekil 3.1 için, hedef lokusun altı SNP'e sahip olduğu ve her bir SNP'in de sadece iki farklı allele sahip olduğu (*biallelik* SNP'ler) kabul edilsin. Allel bilgisi Şekil 3.1.a'da gösterilmiştir. Gri renge boyanmış alleller SNP'in majör allelini, siyah renge

boyanmış olanlar ise SNP'in minör allelini göstermektedir. Şekilden de görüldüğü gibi her bir birey, anne ve babasının iki kromozomlarından üretilen altı SNP'li iki kümeye sahiptir. Bir kromozom üzerinde bulunan SNP kümesine *haplotip* (haplotype) denir (Crawford ve Nickerson, 2005). Şekil 3.1.a'da, altı çift kromozom örneğinden oluşan 12 haplotip vardır ve her bir haplotip çifti bir bireye aittir (Lee ve Shatkay, 2009).



Şekil 3.1. Haplotipler, genotipler ve fenotipler

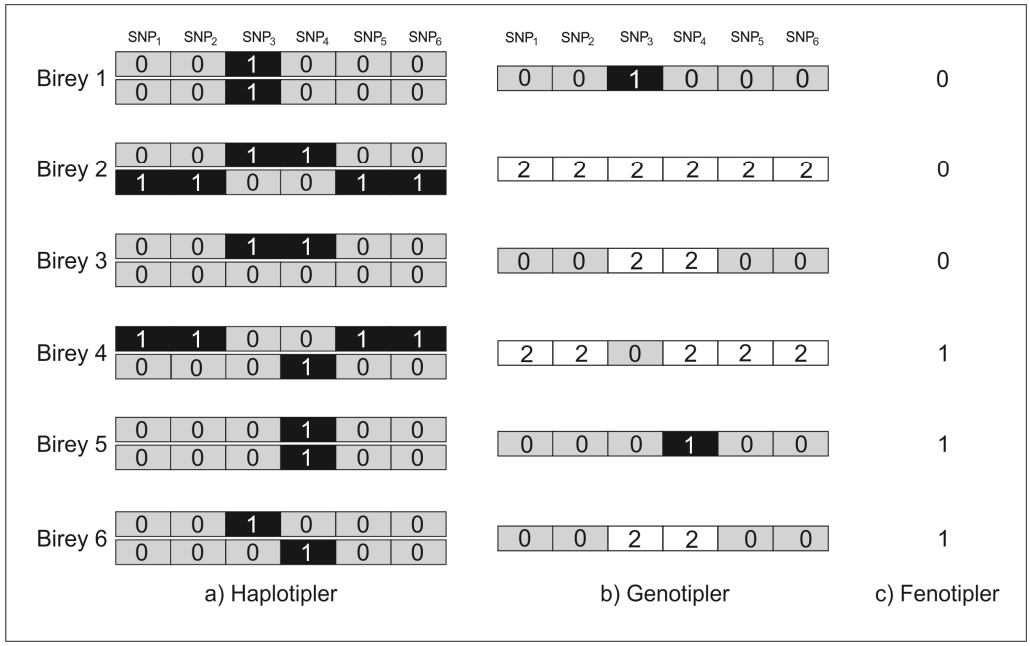
Farklı biyomoleküler metotlar, doğrudan kromozomlardan haplotip bilgisini tanımlayabilir. Ancak, yüksek maliyet ve uzun işlem zamanı nedeniyle bu metodlar yoğun olarak küçük ölçekli örnekler için kullanılır (genellikle birkaç bireyden onlarca bireye kadar) (Crawford ve Nickerson, 2005). Büyük ölçekli bireyler için (genellikle yüzlerce bireyden binlerce bireye kadar) her bir birey için yüksek verili biyomoleküler metodlar hedef lokusun allellerini tanımlamak için kullanılır. Yüksek verili metotların ana eksikliği, her bir allelin kaynak kromozomunu ayırt edebilme yeteneğinin olmamasıdır. Genellikle böyle metotlar sadece bir SNP pozisyonundaki iki allelle ilgilenir. Onların kaynak kromozomlarını tanımlamazlar. Hedef lokustaki bu birleştirilmiş allel bilgisine *genotip* (genotype) denir ve genotip bilgisini elde etmek için yapılan deneysel işleme genotipleme denir (Lee, 2009).

Şekil 3.1.b verilen örnekteki genotip bilgisini gösterir. Gri renkli genotipler bir SNP'in birleştirilmiş allel bilgisinin ikisinin de majör allel, siyah renkli olanlar bir SNP'in birleştirilmiş allel bilgisinin ikisinin de minör allel ve beyaz renkli olanlar ise

bir SNP'in allel bilgisinin birinin majör diğerinin de minör allel olduğunu gösterir. Genotiplerin sayısı bireylerin sayısına eşittir ve verilen örnek için altıdır (Lee ve Shatkay, 2009).

Haplotipler ve genotipler, kromozomlar üzerindeki hedef lokusun allel bilgisini temsil ederken bir *fenotip* (phenotype) fizikseldir. Verilen örnekte bir bireyin fenotipi ya akciğer kanserli ya da akciğer kansersiz olmasıdır. Genellikle hastalıklı bireylere durumlar (cases) hastalıksız bireylere ise kontroller (controls) denir. Şekil 3.1.c, verilen örneğin fenotip bilgisini gösterir (Lee, 2009).

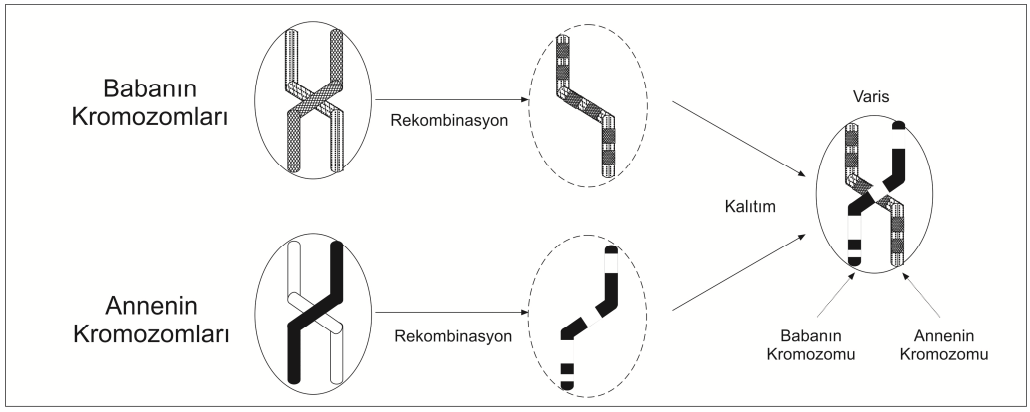
Biallelik SNP'ler için her bir haplotip bir binary string ile temsil edilebilir. Burada 0 majör alleli, 1 minör alleli gösterir (Şekil 3.2.a). Bir genotip içerisinde de SNP'ler ya homozigos ya da heterozigos'dur. Bu nedenle bir genotip de 0, 1 veya 2 rakamları ile temsil edilebilir. Buradaki 0 ilgili SNP'in iki allelinin de major homozigos (0/0), 1 ilgili SNP'in iki allelinin de minor homozigos (1/1) ve 2 ise ilgili SNP'in iki allelinin de heterozigos (0/1, 1/0) olduğunu gösterir (Şekil 3.2.b). Bireylerin fenotip bilgisi de 0 ve 1 rakamları ile temsil edilebilir. 0 rakamı örneğimizdeki akciğer kansersiz bireyleri, 1 rakamı ise akçiğer kanserli bireyleri temsil etmek için kullanılabilir (Şekil 3.2.c) (İlhan ve ark. 2011a).



Şekil 3.2. Haplotip, genotip ve fenotiplerin sayısal gösterimi

3.3.2. İnsan genomunun bağlantı dengesizliği ve blok yapısı

Bir haplotipin ilginç bir özelliği, *bağlantı dengesizliği* (linkage disequilibrium-LD) olarak adlandırılan kendini oluşturan SNP'ler arasındaki rastgele olmayan ilişkidir (Goldstein, 2001). Daha önce bahsedildiği gibi insanlar biri anneden diğeri babadan olmak üzere her bir kromozomdan iki kopyaya sahiptir. Bu kromozomlardan her biri, ebeveynlerin iki kopya kromozomunun rekombinasyonu sonucu üretilir ve kalıtım yoluyla oğullara geçirilir. Şekil 3.3 bu işlemi gösterir (Lee ve Shatkay, 2009).



Şekil 3.3. Rekombinasyon ve kalıtım

Teorik olarak rekombinasyon, iki kromozom boyunca herhangi bir zamanda herhangi bir pozisyonda meydana gelebilir. Böylece bir kromozom üzerindeki bir SNP, eşit olasılıkla ebeveynlerin iki kromozomunun her iki kopyasından da gelebilir ve bir SNP'in kaynağı diğerlerinin kaynağından etkilenmez. SNP'ler arasındaki bu bağımsızlık karakteristiği bağlantı eşitliği (linkage equilibrium) olarak isimlendirilir (Lee ve Shatkay, 2009).

s_1 ve s_2 isminde iki SNP için $|s_1|$ ve $|s_2|$ sırasıyla bu iki SNP'in sahip olduğu allellerin sayısını göstersin. $i=1,..,|s_1|$ ve $j=1,..,|s_2|$ olmak üzere s_{1i}, ilk s_1 SNP'inin *i.* allelini ve s_{2j}, ilk s_2 SNP'inin *j.* allelini göstersin. Bağlantı eşitliği altında s_{1i} ve s_{2j} allellerinin bağlanma olasılığı, s_1 ve s_2 bağımsız olduğu için bu iki allelin bireysel olasılıklarının çarpımına eşit olması beklenir. Bu bağımsızlık varsayımı altında;

$$\forall_{i,j}\ Pr(s_{1i}, s_{2j}) = Pr(s_{1i}).Pr(s_{2j}) \tag{3.1}$$

İki SNP'in allelleri bağımsız olmadığı için Denklem 3.1 sağlanmaz. Bu nedenle, iki SNP'in bağlantı dengesizliği altında olduğu kabul edilir. Prensipte, SNP'lerin allel bağımlılığı büyük iken yüksek LD durumunda oldukları kabul edilir (Lee ve Shatkay, 2009).

Genelde, fiziksel yakınlık içindeki SNP'lerin yüksek LD durumunda olduğu kabul edilir. Rekombinasyon olasılığı iki SNP arasındaki mesafe ile artar (Crawford ve Nickerson, 2005). Bu nedenle, fiziksel yakınlık içindeki SNP'ler atadan oğullarına birlikte geçirilme eğilimindedir. Sonuç olarak bu SNP'lerin allelleri, her biri diğeri ile yüksek oranda ilişkilidir ve bu SNP'leri içeren farklı haplotiplerin sayısı bağlantı dengesizliği nedeniyle beklenilenden çok daha küçüktür (Lee ve Shatkay, 2009).

Son zamanlarda, büyük ölçekli LD çalışmaları (Daly ve ark., 2001; Gabriel ve ark., 2002; Patil ve ark., 2001) insan genomunun kapsamlı LD yapısını anlamak için yapılmıştır. Elde edilen sonuçlar genomun, bloklar olarak adlandırılan farklı bölgelere bölünebileceği hipotezini güçlü bir şekilde destekler niteliktedir. Ayrıca, bir blok içinde rekombinasyonun çok nadir olduğu (yüksek LD) bloklar arasında ise çok yaygın olduğu (düşük LD) gözlemlenmiştir. Sonuç olarak, yüksek LD bir blok içindeki SNP'ler arasında meydana gelir ve böyle SNP'leri içeren farklı haplotiplerin sayısı bir popülasyon içinde şaşırtıcı şekilde küçüktür. Bu gözlem, insan genomunun blok yapısı olarak adlandırılır. Ancak insan genomunun bloklarının nasıl tanımlanacağı konusunda halen bir anlaşma yoktur (Ding ve ark., 2005b; Schulze ve ark., 2004).

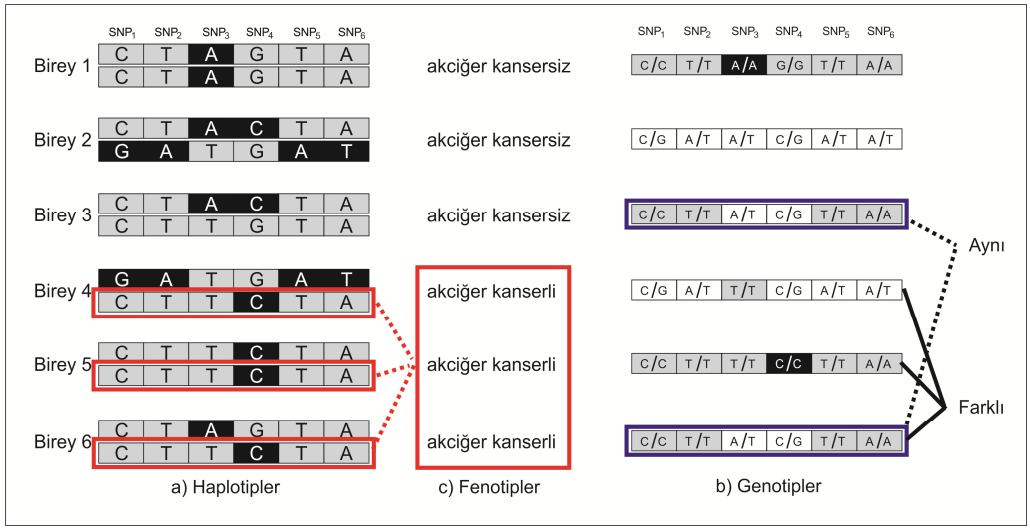
Fiziksel yakınlık içinde olan SNP'ler arasındaki yüksek LD ve insan genomunun blok yapısı nedeniyle haplotiplerin sınırlı sayısı, hastalık-gen ilişkisi için kullanılan sayısal haplotip analizinin temelini teşkil eder.

3.4. Sayısal Haplotip Analizi

Sayısal haplotip analizinin amacı, belli bir hastalıkla yüksek oranda ilişkili olan DNA değişimlerinin kümesini tanımlamaktır. Haplotip, genotip ve hatta tek SNP bilgisi, hedef hastalıkla genetik değişimin ilişkisini tanımlamak için kullanılabilir. Haplotip bilgisinin hastalık-gen ilişkisini araştırmak için kullanılmasına *haplotip analizi* denir. *Tek SNP analizi* ve *genotip analizi* ise bu çalışmalarda sırasıyla tek SNP bilgisi ve genotip bilgisinin kullanılmasına denir (Lee ve Shatkay, 2009).

Haplotip analizi, tek SNP analizi ve genotip analizi ile karşılaştırıldığında birkaç avantaja sahiptir. Tek SNP analizi, bir bireyin fenotipini etkileyen bir kromozom

üstündeki birkaç SNP'in birleşiminin (haplotip) gerekli olduğu ilişkiyi tanımlayamaz (Akey ve ark., 2001; Daly ve ark., 2001; Zhang ve ark., 2002a). Şekil 3.4, bu durumu göstermektedir. Şekil 3.4.a'da çerçeve içerisinde gösterildiği gibi akciğer kanserli üç birey, *CTTCTA* haplotipini paylaşır. Böylece akciğer kanseri fenotipinin *CTTCTA* haplotipi ile ilişkili olduğu sonucu çıkarılabilir. Ancak, eğer altı SNP'in her biri bireysel olarak sınanırsa bu SNP'lerden herhangi biri ile akciğer kanseri fenotipi arasında doğrudan bir ilişki bulunamaz. Örneğin hem akciğer kanserli hem de akciğer kansersiz bireyler, ilk SNP üzerinde *C* alleline veya *G* alleline sahiptir (Lee ve Shatkay, 2009).



Şekil 3.4. Haplotip analizi ve genotip analizi arasındaki fark

Genotipler, *faz* olarak bilinen kaynak kromozom bilgisini içermez. Bu nedenle, genotipler bir haplotip ile hedef hastalık arasındaki mevcut olan açık ilişkiyi gizleyebilir. Örneğin Şekil 3.4.a'da, akciğer kanserli her birey (durum) iki haplotipe sahiptir. Bu haplotiplerden biri, akciğer kanseri fenotipi ile ilişkili olan *CTTCTA* haplotipidir. Diğeri ise her durum için benzersizdir. Bütün durumlar, *CTTCTA* haplotipini paylaşmasına rağmen, Şekil 3.4.c'de onların genotiplerinin tamamı benzersiz haplotipleri nedeniyle farklı gözükürler. Daha da kötüsü akciğer kanserli olan birey 6'nın genotipi akciğer kansersiz olan birey 3'ün genotipi ile benzerdir. Bu nedenle, akciğer kanseri ile yüksek oranda ilişkili olan belli bir genotip tanımlanmaz ve bunun sonucu olarak *CTTCTA* haplotipi ile akciğer kanseri arasındaki gerçek ilişki kaçırılabilir (Lee ve Shatkay, 2009).

Avantajlarına rağmen haplotip analizinin kullanımı haplotip bilgisini elde temek için kullanılan biyomoleküler metotların uzun işlem süresi ve yüksek maliyeti ile

sınırlıdır. Ancak *haplotip fazlama* ve *etiket SNP seçimi* gibi sayısal prosedürler bu problemi çözebilir ve büyük oranda hastalık-gen ilişkisi için haplotip analizinin kullanımına yardımcı olabilir. Haplotip fazlama, genotip verisinden haplotip bilgisini çıkarır. Etiket SNP seçimi hastalık-gen ilişkisini çalışmak için yeterli derecede bilgi verici olan bir haplotip üzerindeki SNP'lerin alt kümesini seçer. Bu sayısal prosedürler, haplotip analizi için kullanılırken bütün prosedür sayısal haplotip analizi olarak adlandırılır (Lee, 2009).

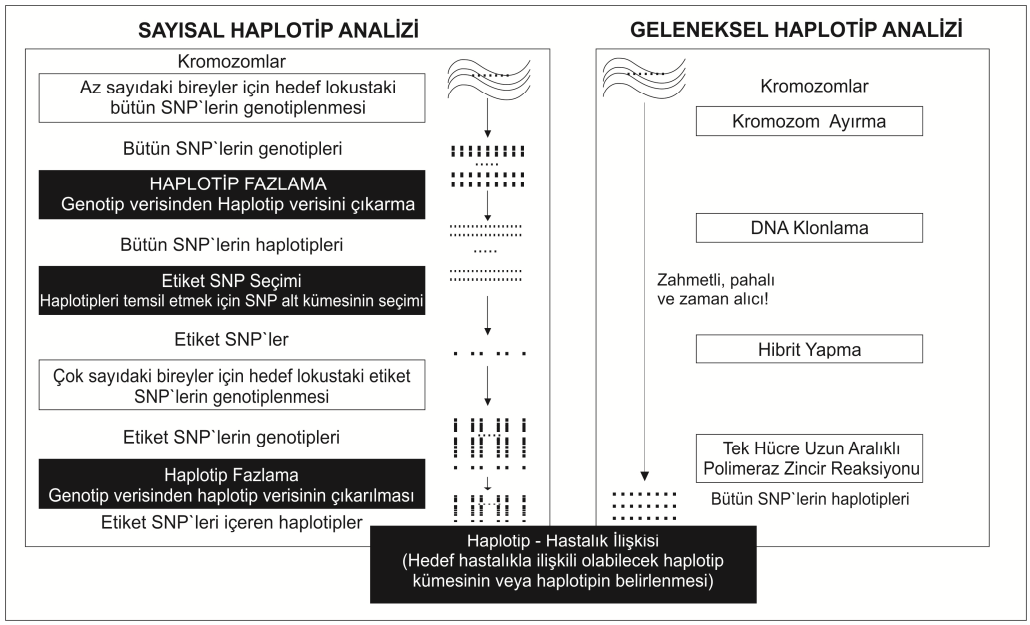
Şekil 3.5, sayısal haplotip analizinin ve geleneksel haplotip analizinin prosedürlerini gösterir. Biyomoleküler deneyler beyaz kutularda gösterilir ve sayısal ve istatistiksel prosedürler siyah kutularda görüntülenir. Sayısal haplotip analizi, iki genotipleme deneyi ile birlikte haplotip fazlama, etiket SNP seçimi ve haplotip-hastalık ilişkisini içerir. Başlangıçta, bireylerin nispeten küçük bir sayısı hedef popülasyondan genotiplenir ve haplotipleri, haplotip fazlama algoritmaları kullanılarak çıkarılır. Daha sonra, etiket SNP seçim algoritmaları, haplotipler üzerindeki SNP'lerin küçük bir alt kümesini seçer. Bu alt küme, küçük bir bilgi kaybı ile belirlenen haplotipleri temsil eder. SNP'lerin seçilen küçük bir sayısını kullanarak ikinci genotipleme, bireylerin büyük bir sayısı için yapılır. Haplotip fazlama algoritmaları bu genotip verisinden haplotipleri çıkarmak için tekrar kullanılır. Sonuç olarak, hedef hastalıkla haplotiplerin bir kümesi veya haplotipin ilişkisini tanımlayan haplotip-hastalık ilişkisi haplotiplere uygulanabilir (Lee ve Shatkay, 2009).

Sayısal haplotip analizinin aksine geleneksel haplotip analizi, haplotip bilgisini doğrudan elde etmek için biyomoleküler deneylere güvenir. Böylece sayısal prosedürlerden çok daha doğru haplotip bilgisini sağlayabilir ve yakın gelecekte biyomüleküler metotlar haplotip analizi için standart bir teknik olabilir (Nothnagel, 2004). Ancak o zamana kadar haplotip fazlama ve etiket SNP seçim prosedürlerinin büyük ölçekli ilişki çalışmaları için oldukça fazla kullanılacağı beklenir (Lee, 2009).

3.5. Etiket SNP Seçimi

Çok büyük ölçekli hastalık çalışmalarında çok sayıdaki bireyler için aday bölge içindeki bütün SNP'leri genotiplemek hâlâ oldukça zaman alıcı ve maliyetlidir. Bu nedenle, hastalık-gen ilişkisini belirlemek için yeterince bilgi verici SNP alt kümesini seçmek çözülmesi gereken kritik bir problemdir. Bu işlem *etiket SNP seçimi* olarak

bilinir. Genellikle, bir haplotip üzerinde seçilen SNP'lere *etiket SNP*'ler seçilmeyen SNP'lere de *etiketlenen SNP*'ler denir (Lee, 2009).



Şekil 3.5. Sayısal haplotip analizi ve geleneksel haplotip analizi

$S=\{s_1, s_2,.., s_m\}$ aday bölge içindeki SNP kümesi ve $D=\{h_1, h_2,.., h_n\}$ kümesi de her biri m SNP'i içeren haplotip kümesi olsun. $h_i \in D$ her bir elemanı 0 ve 1'lerden oluşan m büyüklüklü bir vektördür. Etiket SNP'lerin sayısını k ve $f(T', D)$ ise T' alt kümesinin D orijinal verisini nasıl temsil ettiğini değerlendiren bir fonksiyon olsun. Buna göre etiket SNP seçim problemi aşağıdaki şekilde ifade edilebilir (Lee, 2009).

Problem : Etiket SNP seçimi

Giriş : S SNP kümesi, D haplotip kümesi, k etiket SNP'lerin maksimum sayısı

Çıkış : $T = \text{argmax}_{T' \subset S \,\&\, |T'| \leq k} f(T', D)$

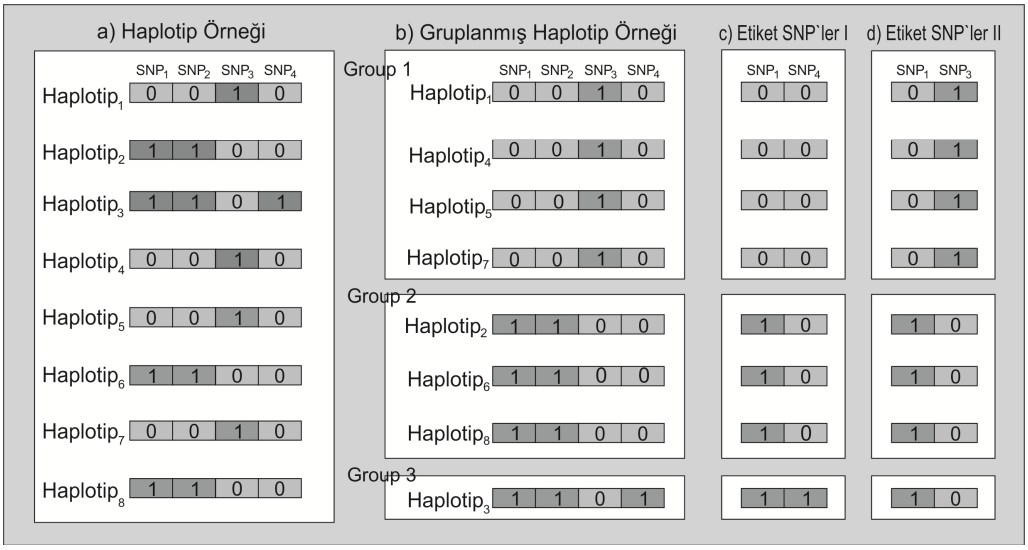
Özet olarak etiket SNP seçim problemini çözmek için orijinal SNP'lerin bütün olası alt kümeleri arasında f değerlendirme fonksiyonuna dayanarak en uygun T SNP alt kümesinin bulunması gerekir.

Haplotip bilgisini ölçmek için kullandıkları yaklaşıma göre etiket SNP seçimi için kullanılan algoritmalar dört gruba ayrılır ve aşağıda sırasıyla açıklanmıştır.

3.5.1. Haplotip farklılığı tabanlı metotlar

Haplotip farklılığına dayanan metotlar, insan genomunun farklı bloklara bölünebildiği ve her bir bloğun yaygın haplotiplerin çok küçük bir sayısını içerdiği varsayımına dayanır (Daly ve ark., 2001; Gabriel ve ark., 2002; Johnson ve ark., 2001; Patil ve ark., 2001). Bu metotlar orijinal veri içindeki sınırlı haplotip farklılığının çoğunu yakalayabilen SNP alt kümesini bulmayı hedefler.

Şekil 3.6, haplotip farklılığına dayanarak etiket SNP kümesinin nasıl seçileceğini gösterir. Şekil 3.6.a'da dört SNP'li sekiz haplotipten oluşan örnek bir veri kümesi görülmektedir. Bir SNP'in majör alleli açık gri içerisinde 0 ile minör alleli ise koyu gri içerisinde 1 ile gösterilmiştir. Her bir allel ya majör ya da minör olması gerektiği için dört SNP'ten oluşan olası farklı haplotiplerin sayısı 2^4'dür. Ancak örnekte farklı haplotiplerin gözlenen sayısı Şekil 3.6.b'de görüldüğü gibi sadece 3'dür. Bu nedenle, iki SNP'li bilgi farklı haplotiplerin sınırlı sayısını benzersiz olarak tanımlamak için yeterli olabilir. Prensip olarak iki SNP'in orijinal veri içerisindeki farklı haplotipleri ayırt edebilme kabiliyeti bütün olası kombinasyonlar denenerek bulunabilir. Daha sonra, en çok ayırt edebilme yeteneğine sahip olan çift etiket SNP'ler olarak seçilir (Lee, 2009).



Şekil 3.6. Sınırlı haplotip farklılığına dayanan etiket SNP seçimi

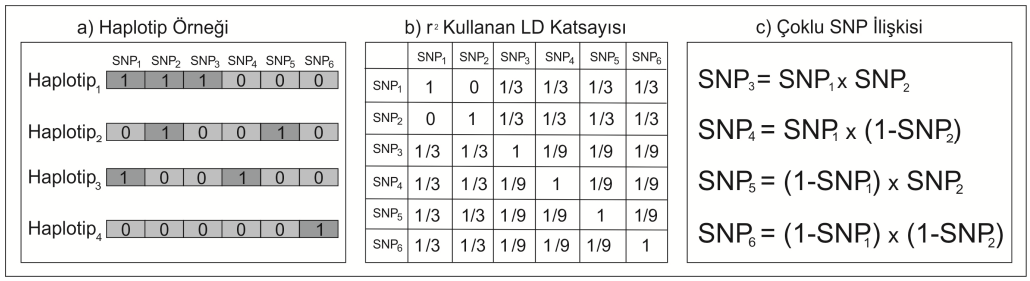
Şekil 3.6.c ve Şekil 3.6.d'de görüldüğü gibi SNP_1 ve SNP_4 başarılı bir şekilde 8 haplotipi 3 farklı gruba bölerken SNP_1 ve SNP_3 haplotiplerden sadece dördünü doğru olarak bir farklı küme içerisine yerleştirir. Diğer dört haplotip farklı olmalarına rağmen

aynı grupta gösterilir. Bu yüzden {SNP_1, SNP_4} alt kümesi tarafından elde edilen sınırlı haplotip farklılığı 8 iken {SNP_1, SNP_3} için bu ölçü sadece 4'tür (Lee, 2009).

3.5.2. SNP'ler arası ilişki katsayısına dayanan metotlar

Bu metotlar, etiket SNP kümesinin bir haplotip üzerindeki bir hastalık lokusunu tahmin edebilecek en küçük SNP alt kümesi olması gerektiği fikrini savunur (Ao ve ark., 2005; Byng ve ark., 2003; Carlson ve ark., 2004; Goldstein ve ark., 2003; Ke ve ark., 2005; Wu ve ark., 2003). Ancak hastalık lokusu genellikle aranılan yerdir ve daha önceden bilinemez. Bunun yerine SNP'ler arasındaki ilişki tahmin için kullanılır. Prensip olarak bir haplotip üzerindeki bütün SNP'ler etiket SNP'lerden en az biri ile yüksek oranda ilişkilidir ve etiket SNP kümesi seçiminde bu kurala dikkat edilir. Bu nedenle, bir hastalık ile ilişkili SNP etiket SNP olarak seçilmese dahi bu SNP ile hedef hastalık arasındaki ilişki, yüksek oranda ilişkili olduğu etiket SNP'den dolaylı olarak çıkarılabilir. Birçok çalışma içinde SNP'lerin rastgele olmayan bu ilişkisi, ilişki katsayısını tahmin etmek için kullanılır (Lee, 2009).

Bütün ilişki katsayısı tabanlı yaklaşımlar $O(cnm^2)$ karmaşıklığına sahiptir. Burada c kümelerin sayısı, n haplotiplerin sayısı ve m SNP'lerin sayısıdır. Bu nedenle bu metotlar genelde blok bölme prosedürü gerektirmediği için haplotip farklılığına dayanan metotlardan daha hızlı çalışır. İlişki katsayısı tabanlı metotların ana eksikliği, çoklu SNP bağımlılıklarını yakalayamamasında ve diğer metotlardan daha çok etiket SNP seçme eğiliminde olmasında yatar (Burkett ve ark., 2005; Goldstein ve ark., 2003; Ke ve ark., 2005; Schulze ve ark., 2004). Şekil 3.7, ilişki katsayısı tabanlı metotların bu zayıflığını gösterir. Şekil 3.7.a'da, altı SNP'li dört haplotipten oluşan örnek bir veri kümesi görülmektedir. Eğer SNP'ler arasındaki LD katsayısını ölçmek için en yaygın LD ölçüsü olan r^2 korelasyon katsayısı (Goldstein ve ark., 2003) kullanılırsa Şekil 3.7.b'de görüldüğü gibi hiçbir SNP çifti 0.5'den daha büyük LD katsayısına sahip değildir. Bu nedenle ilişki katsayısı tabanlı metotlar altı SNP'in tamamını etiket SNP olarak seçer. Ancak Şekil 3.7.c'de gösterildiği gibi SNP 3, 4, 5 ve 6'nın alleli, SNP 1 ve 2'nin allelleri tarafından mükemmel bir şekilde temsil edilir. Böylece çoklu SNP bağımlılıkları dikkate alınarak sadece SNP 1 ve 2 ile altı SNP'in tamamı temsil edilebilir (Lee, 2009).



Şekil 3.7. SNP'ler arasındaki LD katsayısı ve çoklu SNP bağımlılıkları

3.5.3. Fenotip ilişkisine dayanan metotlar

Bu yöntemler fenotip bilgisinin mevcut olduğunu kabul eder ve hastalıksız bireylerden hastalıklı bireyleri ayırt edebilecek SNP kümesini bulmayı denerler. Bulunan bu SNP'ler daha sonra etiket SNP kümesi olarak kullanılırlar. Bu açıdan bakıldığında etiket SNP seçimi küçük bir hata ile iki sınıf (hastalıklı/hastalıksız) arasındaki ayırt edilebilir özellikleri seçmeyi hedefleyen bir çeşit özellik seçim problemidir. Sonuç olarak sınıf etiketleri ve özellikler arasındaki ilişkiyi ölçen sınıflama teknikleri ve istatistiksel testler bu yöntemlerde kullanılır (Lee, 2009).

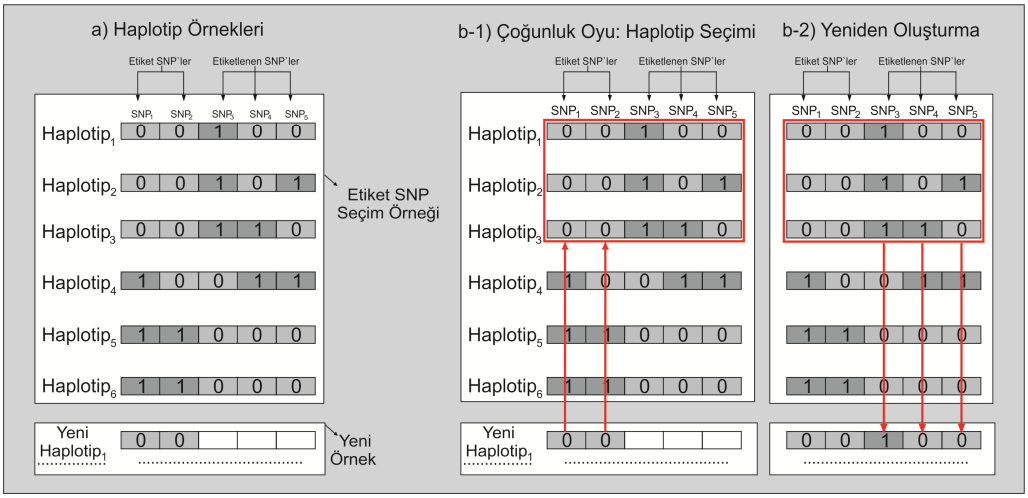
Fenotip ilişki tabanlı metotlar sayısal haplotip analizinin amacıyla doğrudan ilişkilidir. Fenotip sınıflama tabanlı yaklaşımların ana sınırlaması, fenotip bilgilerine olan ihtiyaçlarında yatar. İlaveten, etiket SNP seçimi için kullanılan haplotiplerin sayısı nispeten küçüktür. Bu nedenle küçük örnekleri sınıflamak için seçilen etiket SNP'ler daha büyük örnekler için uygulanamaz. Bu işlem hastalık-gen ilişki çalışmasının performansını doğrudan etkileyebilir.

3.5.4. Etiketlenen SNP tahminine dayanan metotlar

Etiket SNP seçiminde kullanılan diğer bir yaklaşım ise etiketlenen (geriye kalan) SNP tahminine dayanan yöntemlerdir. Bu yöntemlerde etiket SNP seçimi sadece SNP'lerin belli bir kümesini kullanarak orijinal haplotip verisinin yeniden oluşturulması problemi olarak dikkate alınmıştır. Böylece bu yöntemlerde, küçük bir hata ile seçilmeyen (etiketlenen) SNP'leri tahmin edebilecek SNP'lerin kümesinin seçilmesi amaçlanmıştır. Genellikle seçilen etiket SNP'ler genotiplendikten sonra geriye kalan SNP'lerin allelleri etiket SNP'lerin allelleri kullanılarak tahmin edilir ve hastalık-gen ilişkisi bütünüyle oluşturulan haplotip verisine dayanılarak yönetilir. Bu nedenle bu

yöntemler seçilen etiket SNP kümesi ile birlikte etiketlenen SNP kümesi için bir tahmin kuralı sunarlar (Lee, 2009).

Şekil 3.8.a'da, beş SNP'li altı haplotipten oluşan örnek bir veri kümesi görülmektedir. Bu veri kümesinde SNP 1 ve 2 etiket SNP olarak seçilir. Yeni bir örnek içinde her bir haplotipin genotiplenmeyen allellerini (etiketlenen SNP'ler) oluşturmak için bu örnekteki etiket SNP'lere dikkat edilir. Bu örnekteki etiket SNP'ler ile aynı değerlere sahip olan haplotiplerin etiketlenen SNP'lerine bakılır (Şekil 3.8.b-1). Bu SNP'lerin çok sık olarak karşılaşılan allelleri, yeni haplotipteki her bir etiketlenen SNP'in alleli olarak belirlenir. Şekil 3.8.b-2'de görüldüğü gibi, çoğunluk oylama kuralı yeni haplotipe, nadir olan allellerden daha ziyade yaygın olan allleleri atama eğilimindedir (Lee, 2009).



Şekil 3.8. Etiketlenen SNP tahmin tabanlı metotlarda çoğunluk oylaması

İlişki katsayısı tabanlı metotlardan farklı olarak etiketlenen SNP tahmin tabanlı metotlar etiket SNP kümesini seçmek için çoklu SNP bağımlılıklarını kullanır. Bunun sonucu olarak seçilen etiket SNP'lerin sayısı çoğunlukla ilişki katsayısı tabanlı metotlarınkinden daha küçüktür (De Bakker ve ark., 2006). İlaveten, bütün dinamik programlama metotları (Bafna ve ark., 2003; Halldorsson ve ark., 2004b; Halperin ve ark., 2005) verilen ölçüye göre global optimumu bulmayı garanti eder. Ancak bu metotların tahmin yeteneği küçük sınırlandırılmış konum, etiket SNP'lerin sabit sayısı gibi kısıtlamalar ile sınırlıdır.

İnsan genom projesinin tamamlanması ile birlikte genom üzerinde yapılan yoğun araştırmalardan biri de belli bir hastalıkla yüksek oranda ilişkili olabilecek DNA değişimini veya DNA değişim kümesini tanımlayacak hastalık-gen ilişkisidir. Haplotip

fazlama ve etiket SNP seçimi olmak üzere iki prosedürden oluşan sayısal haplotip analizi, büyük çaplı ilişki çalışmalarını oluşturmak için oldukça pratik bir yapı sağlar. Bu çalışmalar masrafsız, hızlı ve oldukça doğru bir performans gösterir. Düşük verili biyomoleküler deneyler, daha az masraflı oluncaya kadar sayısal haplotip analizi kullanılmaya devam edecektir. Bu bölümde sayısal haplotip analizi ile ilgili genel kavramlar ve etiket SNP seçimiyle ilgili olarak kullanılan yaklaşımlar tanıtılmıştır (Lee ve Shatkay, 2009).

Haplotip fazlama ile karşılaştırıldığında etiket SNP seçimi üzerine yapılan araştırmalar daha sonra başlamıştır. Bu nedenle bu tezde etiket SNP seçim problemi üzerine çalışmalar yapılmıştır. Yapılan araştırmalarda etiketlenen SNP tahmin tabanlı yaklaşımların diğerlerine göre birkaç avantaja sahip olduğu düşünülmektedir. Bu avantajlar ön blok bölme işleminin ve fenotip bilgisinin gerekmemesi, çoklu SNP ilişkilerinin kullanılabilmesi olarak sayılabilir. Ancak etiketlenen SNP tahmin tabanlı yöntemlerde, genellikle sayısal karmaşıklığı yüksek olan dinamik programlama prosedürleri kullanılmasına rağmen tahmin performanslarının geliştirilmesi gerekmektedir. Bu nedenle bu tez çalışmasında etiket SNP seçim problemi için diğer yapay zekâ teknikleri kullanılmıştır.

4. MATERYAL VE YÖNTEM

4.1. Kullanılan Veri Kümeleri

Etiket SNP seçimi için geliştirilen ve yukarıda detaylı bir şekilde anlatılan GA-SVM, Parametre Optimizasyonlu GA-SVM ve Parametre Optimizasyonlu CLONTagger yöntemlerinin performanslarını değerlendirmek için farklı büyüklükte ve farklı SNP sayılarına sahip veri kümeleri kullanılmıştır. Kullanılan bu veri kümelerinin özellikleri aşağıda verilmiştir.

4.1.1. ACE veri kümesi

ACE (Angiotensin Converting Enzyme) veri kümesi (Reider ve ark., 1999), 11 bireye ait olan 22 haplotipten oluşur. *17q23* kromozomu üzerindeki *24 Kb*'lık genomik bölgede bulunan 52 adet biallelik SNP'i içerir.

4.1.2. ABCB1 veri kümesi

ABCB1 (ATP-Binding Cassette, sub-family B) veri kümesi (Kroetz ve ark., 2003), P-glikoproteininden sorumlu bir gendir. Genom dizisinin *74 Kb*'lık bölgesi üzerinde uzanır. 247 bireye ait olan 494 haplotipten oluşur ve 27 biallelik SNP içerir.

4.1.3. LPL veri kümesi

LPL (Human Lipoprotein Lipase) veri kümesi (Clark ve ark., 1998), *19q13.22* kromozomu üzerindeki *5.5 Kb*'lık genom bölgesinde uzanır. 71 bireyden alınan 142 haplotipi içerir. Her bir haplotip 88 SNP'den oluşur. Bu veri kümesindeki 142 haplotipten 88 tanesi benzersizdir. Bu tez çalışmasında yapılan uygulamalarda bu 88 haplotip kullanılmıştır.

4.1.4. 5q31 veri kümesi

5q31 veri kümesi (Daly ve ark., 2001), 129 aile üçlüsünün *5q31* kromozomu üzerideki *616 Kb*'lık bölgeden elde edilmiştir. Bu veri kümesi 103 biallellik SNP içerir.

Bu tez çalışmasında yapılan uygulamalarda, bu veri kümesi içerisindeki çocuk popülasyonundan elde edilen örnekler kullanılmıştır.

4.1.5. STEAP ve TRPM8 veri kümeleri

Bu veri kümeleri, Hapmap'ten (International HapMap Consortium, 2003) elde edilen 30 CEPH aile üçlüsüne ait olan örnekleri içerir. STEAP ve TRPM8 gen bölgeleri içerisindeki biallelik SNP'lerin sayısı sırasıyla 22 ve 101'dir. Bu tez çalışmasında yapılan uygulamalarda, bu veri kümeleri içerisindeki ebeveyn popülasyonundan elde edilen örnekler kullanılmıştır.

4.1.6. Popülasyon D'nin D9 veri kümesi

D9 veri kümesi (Gabriel ve ark., 2002), Yoruba popülasyonundan alınan 30 aile üçlüsüne ait olan toplam 180 haplotipten oluşur. Her bir haplotip 49 biallelik SNP içerir.

4.1.7. ENm013, ENr112 ve ENr113 veri kümeleri

Bu veri kümeleri, Hapmap ENCODE (International HapMap Consortium, 2003) projesinden elde edilen 30 CEPH aile üçlüsüne ait olan örneklerden oluşur. ENm013, ENr112 ve ENr113 veri kümeleri sırasıyla *7q21.13, 2p16.3* ve *4q26* kromozomları üzerindeki *500 Kb*'lık bölgelerden elde edilmiştir. Her bir bölge içindeki biallelik SNP'lerin sayısı sırasıyla 361, 412 ve 515'dir. Bu tez çalışmasında yapılan uygulamalarda bu veri kümeleri içerisindeki ebeveyn popülasyonundan elde edilen örnekler kullanılmıştır.

4.2. Kullanılan Yöntemler

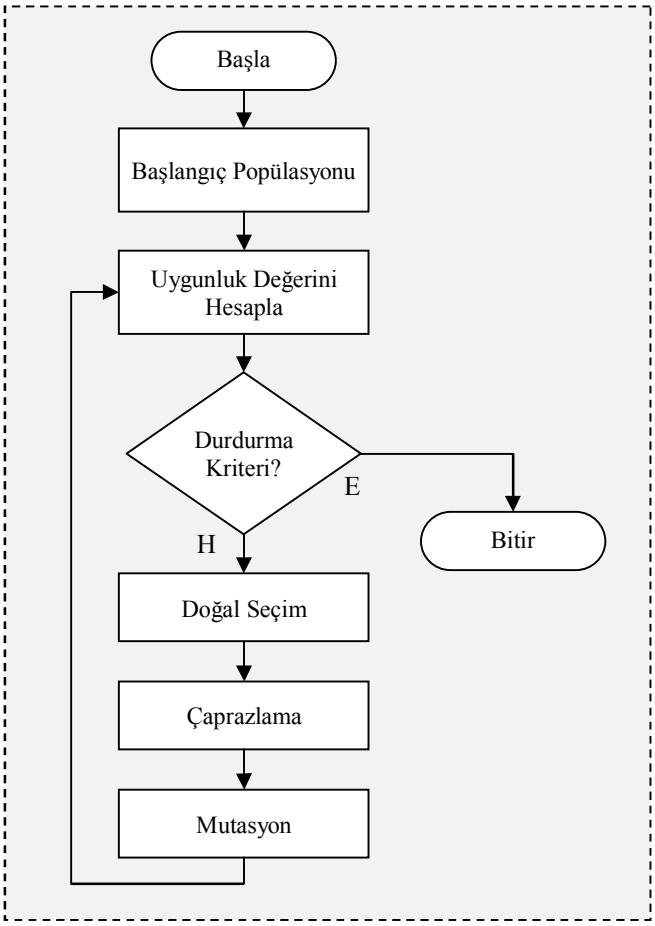
4.2.1. Genetik algoritma

Genetik algoritmalar arama ve optimizasyon için kullanılan sezgisel yöntemlerdir. Geniş ve kapsamlı arama algoritmalarının aksine genetik algoritmalar, en iyiyi seçmek için bütün farklı durumları üretmez. Bu nedenle mükemmel çözüme ulaşamayabilir. Ancak zaman kısıtlamalarını dikkate alan en yakın çözümlerden biridir.

GA, daha önceden hiçbir bilgisi olmamasına karşın, olayları ve bilgiyi öğrenme ve toplama yeteneğine sahiptir.

GA'lar verilen bir veri kümesi için en iyi çözümü (parametreler kümesini) bulan doğrusal olmayan optimizasyon araçlarıdır. GA parametreleri biyolojideki genleri temsil ederken parametrelerin toplu kümesi de kromozomu oluşturmaktadır. GA, mümkün çözümler kümesinin rasgele üretilmesiyle başlar. Parametreleriyle her bir çözüm, arama uzayında (kromozom veya nesil uzayı) uygunluk fonksiyonunun özel bir noktasını üretir. Her bir iterasyonda oluşan bu farklı nesillere popülasyon adı verilir. Her yeni nesil, rasgele bilgi değişimi ile oluşturulan diziler içinde hayatta kalanların birleştirilmesi ile elde edilir. Bu yeni neslin eskilerden daha iyi olması beklenir (Angeline, 1995).

GA'nın basit akış şeması Şekil 4.1'de verilmiştir. Şekilden de görülebildiği gibi başlangıç popülasyonunun oluşturulması, uygunluk değerlendirmesi, doğal seçim, çaprazlama ve mutasyon işlemleri GA'nın temel prosedürlerini oluşturmaktadır.



Şekil 4.1. GA'nın akış şeması

4.2.1.1. Başlangıç popülasyonu

Optimizasyon döngüsü başlamadan önce, optimize edilmesi gereken parametrelerin istenilen şekle dönüştürülmesi gerekir. Bu işleme kodlama (encoding) denir. Kodlama, GA için önemli bir konudur. Çünkü sistemden elde edilen bilgiye bakış açısı büyük ölçüde sınırlandırılabilir. Gen stringi probleme özel bilgiyi depolar. Gen olarak adlandırılan her bir öğe, genellikle değişkenler stringi olarak ifade edilir. Değişkenler ikili veya reel sayı şeklinde gösterilebilir ve aralığı probleme özel olarak tanımlanır. Probleme özel gen stringleri belirlendikten sonra kromozomların oluşturulması gerekir. Kromozomlar, algoritmaya giriş olarak verilen popülasyon sayısına göre rastgele olarak üretilir. Bu kromozomlar başlangıç popülasyon matrisinin satırlarını oluştrur. Parametreler ise bu matriste sütunlar olarak temsil edilir (Özsağlam ve Çunkaş, 2008).

4.2.1.2. Uygunluk değerinin hesaplanması

Rastgele olarak üretilen başlangıç popülasyonundaki her bir kromozomun uygunluk değerinin hesaplanması gerekir. Bu işlem için amaç fonksiyon kullanılır. Amaç fonksiyonu, giriş parametrelerine göre çıkış üreten bir fonksiyondur. Her bir kromozom için hesaplanan uygunluk değerleri dikkate alınarak durdurma kriteri sınanır. Eğer durdurma kriteri sağlanırsa algoritma sona erer. Aksi durumda popülasyona doğal seçim işlemi uygulanır (Özsağlam ve Çunkaş, 2008).

4.2.1.3. Doğal seçim

Popülasyondaki en yüksek uygunluk değerine sahip olan bireylerin hayatta kalması diğerlerinin ise kaldırılması gerekir (Angeline, 1995). Doğal seçim, algoritmanın her iterasyonunda meydana gelir. Rulet tekerleği, rastgele, ağırlıklı, ağırlıklı rastgele, eşik değer, şeçkinlik, turnuva ve hiyerarşik gibi metodlar GA içerisinde kullanılan doğal seçim yöntemleridir. Bu yöntemlerden rulet tekerleği seçim metodu, GA'lar içerisinde yaygın olarak kullanılmaktadır (Goldberg ve Deb, 1991).

4.2.1.4. Çaprazlama

Doğal seçim sonucunda üretilen popülasyonu geliştirmek için yeni popülasyondaki bireylere çaprazlama oranı dikkate alınarak çaprazlama işlemi uygulanır. Çaprazlama işlemi uygulanacak olan bireyler genellikle rastgele olarak seçilir. Çaprazlama işlemi sonucunda ebeveyn kromozomlardan yavru kromozomlar elde edilir (Özsağlam ve Çunkaş, 2008). Genellikle, GA içerisinde çaprazlama operatörü olarak tek nokta, çok noktalı ve düzenli gibi operatörler kullanılmaktadır.

4.2.1.5. Mutasyon

Çaprazlama işlemi sonucunda elde edilen popülasyonu geliştirmek için mutasyon oranınına göre yeni popülasyondaki bireylere mutasyon işlemi uygulanır. Mutasyon işlemi sonucunda popülasyonun sadece bazı bitleri değişir. Mutasyona uğratılacak bitleri tespit etmek için bütün kromozomlardaki her bir bit pozisyonu için 0 ile 1 arasında rasgele sayılar üretilir. Eğer bir kromozomun herhangi bir biti için üretilen bu sayı, mutasyon oranından daha küçük ise bu kromozomdaki ilişkili bit 0 ise 1'e veya 1 ise 0'a değiştirilerek mutasyona uğratılır (Özsağlam ve Çunkaş, 2008).

Mutasyon işleminden sonra yeni popülasyondaki bütün bireylerin uygunluk değerleri tekrar hesaplanır ve en iyi uygunluk değerine sahip birey belirlenir. Bu işlemler algoritmaya giriş olarak verilen jenerasyon sayısı kadar tekrarlanır. Her bir jenerasyon sonucunda elde edilen bireylerden en iyi uygunluk değerine sahip olan birey algoritmanın sonucu olarak geri döndürülür.

4.2.2. Destek vektör makinesi

SVM; öğrenme, sınıflandırma, kümeleme, yoğunluk tahmini ve regresyon kuralları üretmek için kullanılan bir eğitme algoritmasıdır. SVM'nin ilk olarak 1960'lı yıllarda V. Vapnik tarafından teorik temelleri atılmış ve daha sonra sınıflandırma konusunda önerilmiştir (Cores ve Vapnik, 1995).

SVM'de amaç, veri noktalarını mümkün olduğu kadar iyi sınıflandıran ve iki sınıf noktaya ayıran optimum ayırıcı düzlemin bulunmasıdır. Yani iki sınıf arasındaki uzaklığın maksimum olduğu durumun bulunması amaçlanmaktadır. Bu sınıflandırma mantığının temel taşlarını ise her iki sınıfın uç noktalarında bulunan ve eğitim örnekleri

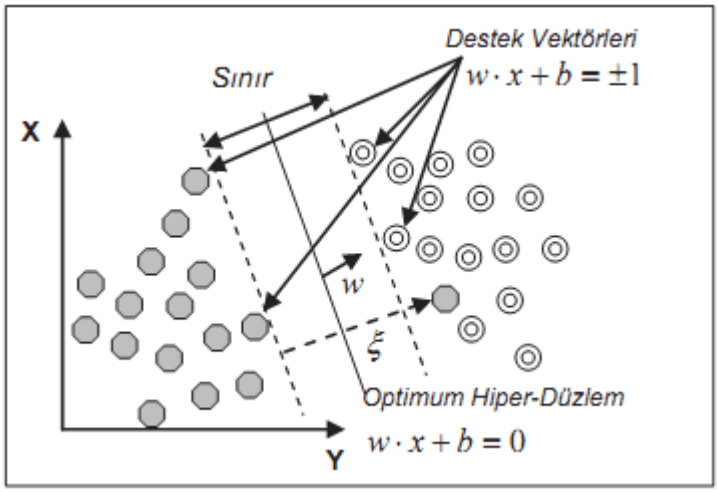
arasından seçilen destek vektörler oluşturmaktadır. Ayırıcı düzlemin optimum yapılması genelleme yeteneğini maksimum düzeye çıkaracaktır. Ancak bu verileri iki sınıfa ayıracak çok fazla sayıda ayırıcı düzlem bulunmaktadır. Öncelikle ayırıcı düzlem sayısı, sınıflar arası mesafe ölçüsüyle sınıf genişliği ölçüsünün çarpımı 1 alınarak sınırlandırılır. Daha sonra bu ayırıcı düzlemler arasında optimum olan bulunur. Optimum ayırıcı düzlem, her iki sınıfın en uç verileri arasındaki mesafenin (iki sınıfın destek vektörleri arasındaki mesafe) maksimum olduğu durumu sağlayan ayırıcı düzlemdir. Bu ayırıcı düzlem bahsedilen aralığın tam ortasından geçmelidir (Çomak, 2008).

Sınıflandırma problemlerinin birkaç türü vardır. Bunlardan en temel olanı eğitim hatasının bulunmadığı doğrusal ayrılabilirlik durumudur. Bu durumda SVM, giriş uzayı üzerinde doğrusal bir ayırıcı fonksiyon oluşturur. Bir başka durum, eğitim hatasının bulunmadığı doğrusal olmayan ayrılabilirlik durumudur. Bu durumda SVM sınıflayıcısı, giriş uzayı üzerinde doğrusal sınıflandırma işlemini gerçekleştiremez. Bu nedenle, ilk önce çekirdek fonksiyonlarının yardımıyla giriş uzayından nitelik uzayına dönüşüm yapılır. Daha sonra SVM sınıflayıcısı, nitelik uzayı üzerinde doğrusal ayırıcı fonksiyonunu oluşturur. Eğitim hatasının bulunduğu durumlarda da benzer işlemler yapılır. Fakat sisteme pozitif değerli bir esneklik parametresi eklenir. Hem sınıflandırma hem de regresyon işlemlerinde, öğrenme problemi ikinci dereceden amaç fonksiyonuna sahip bir optimizasyon problemi olarak temsil edilir. SVM regresyon metodundaki temel fikir, eldeki eğitim verilerinin karakterini mümkün olduğunca gerçeğe yakın bir şekilde yansıtan ve istatistiksel öğrenme teorisine uyan doğrusal ayırıcı fonksiyonun bulunmasıdır. Sınıflandırmaya benzer bir şekilde regresyonda da doğrusal olmayan durumların işlenebilmesi için çekirdek fonksiyonları kullanılır (Çomak, 2008).

SVM eğitim işlemi, ikinci dereceden bir optimizasyon probleminin çözümü gibi ele alınabilir. Çekirdek fonksiyonlarının kullanımı ile birlikte SVM, yüksek boyutlu nitelik uzayında doğrusal bir modele dönüştürülür. Giriş uzayından daha yüksek boyutlu nitelik uzayına haritalama yapılması, tüm gerekli hesaplamalar giriş uzayında yapıldığı için SVM'nin eğitim işlemine ek bir hesap yükü getirmez. Bu yüzden SVM, en karmaşık doğrusal olmayan öğrenme işlemlerinde bile basit bir öğrenme algoritmasıyla görevini yerine getirebilir.

Şekil 4.2'den görülebildiği gibi SVM, sınıflandırma durumunda nitelik uzayında mevcut eğitme verilerine göre optimum ayırma düzlemini öğrenmeye çalışır. Regresyon durumunda ise, mevcut eğitme verilerine göre giriş ve çıkış uzayları arasındaki

haritalama fonksiyonunu öğrenmeye çalışır. SVM başlıca üç bileşenden oluşmaktadır; istatistiksel öğrenme teorisi, optimizasyon algoritması ve çekirdek fonksiyonları. Böylece, SVM çekirdek fonksiyonlarıyla dönüştürülmüş olan nitelik uzayında, iyi bir genelleme teorisinin ışığında ve optimizasyon teorisinin yardımıyla doğrusal olarak eğitilen bir öğrenme makinesi olarak tanımlanabilir (Çomak, 2008).



Şekil 4.2. SVM sınıflandırıcısı

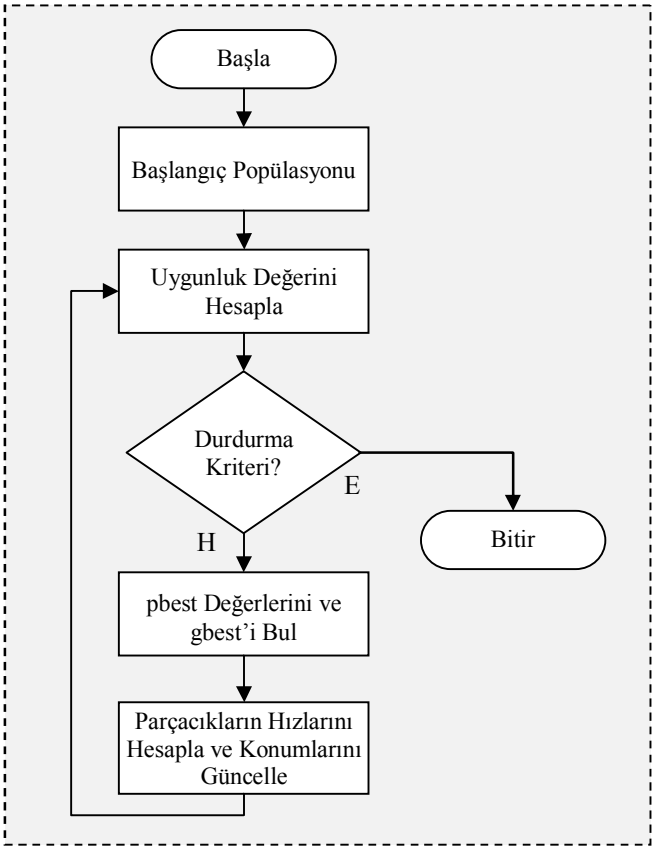
SVM ile gerçekleştirilecek bir sınıflandırma işlemi için kullanılacak çekirdek fonksiyonu ve bu fonksiyona ait en uygun parametrelerin belirlenmesi esastır. Ayrıca tüm SVM'ler için düzenleme parametresi olan C'nin de kullanıcı tarafından belirlenmesi gerekir.

4.2.3. Parçacık sürü optimizasyon algoritması

PSO, Kennedy ve Eberhart (1995) tarafından geliştirilmiş popülasyon temelli sezgisel bir optimizasyon tekniğidir. Kuş veya balık sürülerinin sosyal davranışlarından esinlenilerek geliştirilmiştir. PSO, GA gibi evrimsel hesaplama teknikleri ile birçok benzerlikler gösterir. Sistem random çözümlerden oluşan bir popülasyonla başlatılır ve en iyi çözüm için jenerasyonları güncelleyerek arama yapılır. GA'nın aksine PSO'da çaprazlama ve mutasyon gibi evrimsel operatörler yoktur. PSO'da parçacık denilen potansiyel çözümler, mevcut en iyi çözümleri takip ederek problem uzayında gezinirler. GA ile karşılaştırıldığında PSO çok az parametreye sahiptir ve gerçekleştirilmesi kolaydır. PSO, temel olarak sürüde bulunan bireylerin pozisyonunun, sürünün en iyi pozisyona sahip olan bireyine yaklaştırılmasına dayanır. Bu yaklaşma hızı rasgele gelişen durumdur ve çoğu

zaman sürü içinde bulunan bireyler yeni hareketlerinde bir önceki konumdan daha iyi konuma gelirler ve bu süreç hedefe ulaşıncaya kadar devam eder (Özsağlam ve Çunkaş, 2008).

PSO'nun basit akış şeması Şekil 4.3'de verilmiştir. Şekilden de görülebildiği gibi başlangıç popülasyonunun oluşturulması, uygunluk değerlendirmesi, pbest değerlerinin ve gbest'in bulunması, parçacık hızlarının hesaplanması ve konumlarının güncellenmesi PSO'nun temel prosedürlerini oluşturmaktadır (Özsağlam ve Çunkaş, 2008).



Şekil 4.3. PSO'nun akış şeması

4.2.3.1. Başlangıç popülasyonu

PSO'da her bir çözüm arama uzayındaki bir kuştur ve bunlar bir "parçacık" olarak isimlendirilir. Parçacıklar, algoritmaya giriş olarak verilen popülasyon büyüklüğüne göre rastgele olarak üretilir. Bu parçacıklar başlangıç popülasyon matrisinin satırlarını oluştrur. Parametreler ise bu matriste sütunlar olarak temsil edilir.

4.2.3.2. Uygunluk değerinin hesaplanması

Rastgele olarak üretilen başlangıç popülasyonundaki her bir parçacığın uygunluk değerinin hesaplanması gerekir. Bu işlem için amaç fonksiyon kullanılır. Amaç fonksiyonu, giriş parametrelerine göre çıkış üreten bir fonksiyondur. Her bir parçacık için hesaplanan uygunluk değerleri dikkate alınarak durdurma kriteri sınanır. Eğer durdurma kriteri sağlanırsa algoritma sona erer.

4.2.3.3. pbest değerlerinin ve gbest'in bulunması

Algoritmanın her iterasyonunda, her bir parçacık *pbest* ve *gbest* değerlerine göre güncellenir. pbest değeri bir parçacığın o ana kadar bulduğu en iyi uygunluk değeridir. Ayrıca bu değer daha sonra kullanılmak üzere hafıza tutulur. *gbest* değeri ise popülasyondaki herhangi bir parçacık tarafından o ana kadar elde edilmiş en iyi uygunluk değeridir ve bu değer popülasyon için global en iyi değerdir.

4.2.3.4. Parçacık hızlarının hesaplanması ve konumlarının güncellenmesi

pbest değerleri ve gbest bulunduktan sonra her bir parçacığın hızları hesaplanır ve konumları güncellenir. Bu işlem için sırasıyla Denklem 4.1 ve Denklem 4.2 kullanılır (Özsağlam ve Çunkaş, 2008).

$$v_{ij}^{k+1} = w\ x\ v_{ij}^{k} + \ c_1\ x\ r_1\ x\ \left(pbest_{ij}^{k} - x_{ij}^{k}\right) + c_2\ x\ r_2\ x\ (gbest_{j}^{k} - x_{ij}^{k}) \quad (4.1)$$

$$x_{ij}^{k+1} = x_{ij}^{k} + \ v_{ij}^{k+1} \quad (4.2)$$

Burada *c1* ve *c2* öğrenme faktörleridir ve sırasıyla parçacığın kendi tecrübelerine göre ve sürüdeki diğer parçacıkların tecrübelerine göre hareketini yönlendirir. *r1* ve *r2*, *[0,1]* aralığındaki rastgele değerlerdir. *w* atalet ağırlığıdır ve küçük atalet ağırlığı local aramaya olanak tanırken büyük atalet ağırlığı global aramaya imkan verir. $pbest_{ij}^{k}$, *i.* parçacığın şimdiye kadar ulaşılan en iyi çözümü ve $gbest_{j}^{k}$ ise popülasyon içindeki herhangibir parçacık tarafından şimdiye kadar ulaşılan global en iyi çözümdür. v_{ij}^{k} ve x_{ij}^{k}, sırasıyla, *i.* parçacığın hızı ve çözümüdür.

Güncelleme işleminden sonra yeni popülasyondaki bütün parçacıkların uygunluk değerleri tekrar hesaplanır. Bu işlemler algoritmaya giriş olarak verilen iterasyon sayısı kadar tekrarlanır.

4.2.4. Klonal seçim algoritması

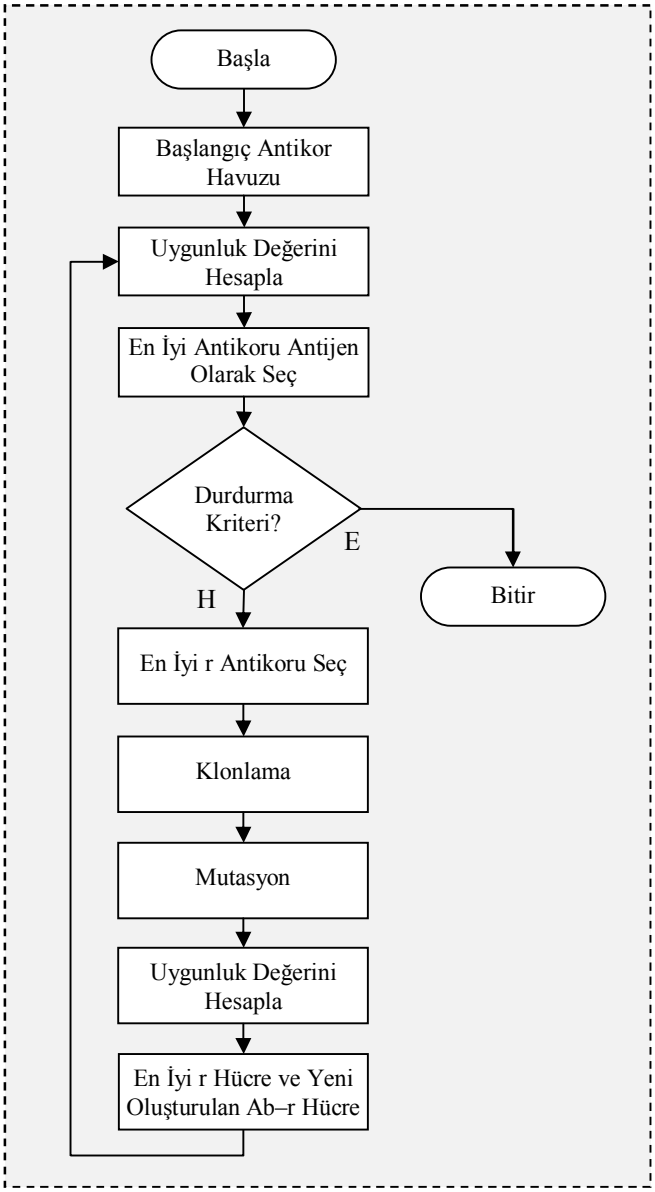
CLONALG algoritması, bağışıklık sisteminin bir antijenik uyarana karşı temel özelliklerini tanımlamak içim kullanılır. Bu algoritma iki temel esas üzerine şekillendirilmiştir: Birincisi sadece antijeni tanıyan hücreler çoğalma için seçilirler. İkincisi seçilen ve çoğalan hücreler duyarlılık olgunlaşması işlemine tabi tutularak, antijene olan duyarlılıkları artırılır (Kodaz, 2007).

CLONALG standart bir GA ile karşılaştırıldığında lokal optimumlardan oluşan çok çeşitli çözüm kümesi üretebilmektedir. GA'lar ise tüm popülasyonu en iyi bireye benzetmeye çalıştırmaktadır. Gerçekte iki algoritmanın da kodlama ve hesaplama yöntemleri çok farklı değildir. Fakat araştırma işlemlerini gerçekleştirirlerken esinlendikleri kaynak, kullandıkları notasyon ve gerçekleştirdikleri işlemlerin sırası bakımından farklılık arz etmektedir. Her iki yöntemde de çözüme ulaşmada popülasyon büyüklüğü önemli bir parametredir. GA'larda popülasyon içerisindeki her bir birey kromozom olarak adlandırılırken CLONALG algoritmasında antikor olarak adlandırılmaktadır. GA'larda optimal ya da optimale yakın çözüme ulaşmada kullanılan en önemli operatör çaprazlama operatörüdür. CLONALG algoritmasında çaprazlama operatörü kullanılmamaktadır. Mutasyon her iki algoritmada da kullanılan bir operatördür. Mutasyon operatörü, algoritmaların yeni çözümlere ulaşma sürecinde ince ayar yapmasını sağlamaktadır (Kodaz, 2007).

CLONALG algoritması ilk olarak De Castro ve Von Zuben (1999) tarafından optimizasyon problemleri için ortaya konulmuştur. Bu algoritmanın akış diyagramı Şekil 4.4.'de verilmiştir.

4.2.4.1. Başlangıç antikor havuzu

Başlangıç antikor havuzu, ikili bir matris ile temsil edilir. Bu matriste satırlar popülasyondaki antikorları, sütunlar ise parametreleri temsil eder. Antikorlar, algoritmaya giriş olarak verilen popülasyon büyüklüğüne göre rastgele olarak üretilir.



Şekil 4.4. CLONALG'nin akış şeması

4.2.4.2. Uygunluk değerlendirmesi

Rastgele olarak üretilen başlangıç antikor havuzundaki her bir antikorun uygunluk değerinin hesaplanması gerekir. Bu işlem için amaç fonksiyon kullanılır. Amaç fonksiyonu, giriş parametrelerine göre çıkış üreten bir fonksiyondur. Her bir antikor için hesaplanan uygunluk değerleri dikkate alınarak durdurma kriteri sınanır. Eğer durdurma kriteri sağlanırsa algoritma sona erer.

4.2.4.3. Klonlama işlemi

Popülasyondaki antikorların en iyi uygunluk değerine sahip olan ilk *r* tanesine klonlama işlemi uygulanır. Her bir antikor için üretilecek olan klon sayısı Denklem 4.3'de gösterilen formül kullanılarak hesaplanır (Kodaz, 2007).

$$C = round(\frac{\beta . Ab}{i}) \tag{4.3}$$

Burada β bir klonlama faktörüdür. *Ab,* antikor havuzundaki antikorların sayısı ve *i,* o anki antikorun uygunluk değeri açısından sırasını gösterir.

4.2.4.4. Mutasyon işlemi

Klonlanan antikorlara mutasyon işlemi uygulanır. Öncelikle klonlanan antikorlarla antijen arasındaki duyarlılıklar hesaplanır. Daha sonra bütün antikorlar için hesaplanan duyarlılık değerleri ile orantılı olarak mutasyon işlemi uygulanır. Duyarlılıkların hesaplanmasında Denklem 4.4'de verilen Hamming mesafe formülü kullanılır. Burada p_{ij}, antikor havuzundaki *i.* antikorun *j.* bitini, Ag_j ise antijenin *j.* bitini gösterir (Kodaz, 2007).

$$D_i = \sum_{j=1}^{n} \delta, \quad \delta = \begin{Bmatrix} 1 & p_{ij} \neq Ag_j \\ 0 & Diğer \end{Bmatrix} \tag{4.4}$$

Mutasyon işlemine tabi tutulan antikorların uygunluk değerleri hesaplanır. En iyi uygunluk değerine sahip ilk r antikor ve yeni oluşturulan Ab-r antikor ile yeni popülasyon oluşturulur ve bir sonraki iterasyondan devam edilir.

5. GELİŞTİRİLEN YÖNTEMLER

Daha önceki bölümde de bahsedildiği gibi etiket SNP seçimi konusunda bugüne kadar farklı çalışmalar yapılmıştır. Bu çalışmalar içerisinde etiketlenen SNP tahmini tabanlı yaklaşımların avantajları dikkate alınarak, etiket SNP seçim probleminde uygulanmak üzere üç farklı yöntem geliştirilmiştir. Bu yöntemler GA-SVM, Parametre Optimizasyonlu GA-SVM ve Parametre Optimizasyonlu CLONTagger olarak adlandırılmıştır. Aşağıda bu yöntemlerin detayları sırasıyla açıklanmıştır.

5.1. GA-SVM Metodu

Etiket SNP'lerin tahmin doğruluğunun artırılması için GA-SVM olarak adlandırılan hibrit bir metot önerilmiştir. Bu yöntemde, etiket SNP'ler GA tarafından seçilmekte ve geriye kalan SNP'lerin tahmini ise SVM tarafından yapılmaktadır.

Şekil 5.1, GA-SVM metodunun algoritmasını göstermektedir. Şekilden de görülebildiği gibi bu metot başlangıç popülasyonunun oluşturulması, uygunluk değerlendirmesi, doğal seçim, çaprazlama, mutasyon ve ayarlama prosedürlerini içeren modüllerden oluşmaktadır.

Girişler: H; Haplotip matrisi
N_G; Jenerasyon sayısı
q; Popülasyondaki birey sayısı
C_R; Çaprazlama oranı
M_R; Mutasyon oranı
N; Etiket SNP sayısı

Çıkışlar: TS = $\{t_1, t_2, \ldots, t_N\}$; Etiket SNP'lerin kümesi
PA; Etiket SNP kümesinin tahmin doğruluğu

Algoritma:

Başla
1. H matrisine ve q değerine göre başlangıç popülasyonunu oluştur
2. Popülasyonun uygunluğunu değerlendir
3. N_G defa aşağıdaki işlemleri tekrarla
3.1. Doğal seçim mekanizmasını kullanarak daha iyi bireyleri hayatta bırak
3.2. C_R çaprazlama oranını kullanarak yeni bireyleri oluştur
3.3. M_R oranını kullanarak bazı bireyleri mutasyona uğrat
3.4. Seçilen etiket SNP'lerin sayısını ayarla
3.5. Popülasyonun uygunluğunu değerlendir
3.6. Mevcut popülasyondaki en iyi bireyi sakla
4. En iyi uygunluk değerli bireyi seç
5. Etiket SNP kümesini (TS) ve seçilen bireyin tahmin doğruluğunu (PA) geri döndür
Bitir

Şekil 5.1. GA-SVM Algoritması

5.1.1. GA tarafından etiket SNP'lerin seçilmesi

5.1.1.1. Başlangıç popülasyonu

Bir veri kümesi *m* satırlı ve *n* sütunlu ikili *H* matrisi ile temsil edilir. Bu matris içinde her bir satır belli bir haplotipi, her bir sütun ise belli bir SNP'i temsil eder (He ve Zelikovsky, 2007). Böyle bir matristeki her bir haplotip, n bitli ikili bir vektör ile temsil edilir. Bu matrise dayanarak, GA *q* satırlı ve *n* sütunlu bir *P* popülasyon matrisi oluşturur. Buradaki *q* değeri popülasyondaki bireylerin sayısıdır ve genellikle 10 ile 200 arasında rastgele değer olarak alınır (Guo ve ark., 2007; Sağ ve Çunkaş, 2009). *P* matrisinin her bir $p_{ij} \in \{0, 1\}$ elemanı *i.* bireyin *j.* SNP'inin değerini temsil eder. Bu matristeki bir birey içerisindeki 1 değeri ilişkili SNP'in etiket SNP olduğunu, 0 değeri ise ilişkili SNP'in etiketlenen SNP olduğunu yani değerinin tahmin edilmesi gerektiğini gösterir. *P* popülasyon matrisindeki bireylerin tamamı *N* ile gösterilen aynı sayıdaki etiket SNP'e sahiptir. Ancak farklı bireyler bu etiket SNP'lerin farklı kombinasyonlarına sahip olabilir. Şekil 5.2'de örnek bir *P* popülasyon matrisi verilmiştir. Bu matris *q=5* birey ve *n=15* SNP'den oluşur ve her bir birey *N=5* etiket SNP'e sahiptir.

	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	SNP_8	SNP_9	SNP_{10}	SNP_{11}	SNP_{12}	SNP_{13}	SNP_{14}	SNP_{15}
Birey 1	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1
Birey 2	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0
Birey 3	1	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Birey 4	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0
Birey 5	0	0	0	1	0	0	1	1	0	0	0	0	1	0	1

Şekil 5.2. 5 birey ve 15 SNP'den oluşan popülasyon matrisi. Her bir birey 5 etiket SNP ve 10 etiketlenen SNP'den oluşur.

Bir GA, aday çözümlerden oluşan sabit bir popülasyon büyüklüğü ile sürdürülen iteratif bir prosedürdür (Zou ve ark., 2008). Bu algoritmanın her bir iterasyonunda, yeni popülasyonu üretmek için doğal seçim, çaprazlama ve mutasyon olmak üzere üç genetik operatör uygulanır. Yeni popülasyonun kromozomları sonraki bölümde Denklem 5.1 ile verilen uygunluk fonksiyonu kullanılarak değerlendirilir. Bu değerlendirmelere göre daha önceki kromozomlardan daha iyi olanlar aday çözümler olarak belirlenir (Zou ve ark., 2008).

5.1.1.2. Uygunluk değerlendirmesi

GA içerisinde, popülasyonların uygunluk değerlendirmesi için birini dışarıda bırak çapraz doğrulama (LOOCV), 10 kat çapraz doğrulama ve 5 kat çapraz doğrulama gibi yöntemler kullanılabilir. LOOCV yönteminin tahmin doğruluğu diğerlerine göre daha iyi olduğu ve karşılaştırma yapılan diğer çalışmaların tamamında bu yöntem kullanıldığı için GA-SVM algoritmasının uygunluk değerlendirmesi işleminde de bu yöntem kullanılmıştır (Arlot ve Celisse, 2010; He ve Zelikovsky, 2007; Lee ve Shatkay, 2006; Lin ve Leu, 2010; Yang ve ark., 2008). LOOCV metoduna göre *j.* iterasyonda öncelikle *j.* haplotip *H* matrisinden kaldırılır. Daha sonra geriye kalan haplotiplerden GA kullanılarak etiket SNP'ler seçilir ve bu seçilen etiket SNP'ler kaldırılan haplotipteki etiketlenen SNP'leri (SNP'lerin geriye kalanını) tahmin etmek için kullanılır. Bu işlem, *H* matrisindeki bütün haplotipler doğrulama verisi olarak kullanılıncaya kadar bütün $j=1,2,\ldots,m$ değerleri için tekrarlanır. Doğru olarak tahmin edilen SNP sayısının toplam tahmin edilen SNP sayısına oranı tahmin doğruluğunu (uygunluk değerini) verir ve Denklem 5.1'deki eşitlik ile hesaplanır. Burada N_c doğru olarak tahmin edilen SNP sayısını, N_a ise toplam tahmin edilen SNP sayısını gösterir.

$$PA = N_c / N_a \tag{5.1}$$

5.1.1.3. Doğal seçim

Popülasyondaki en yüksek uygunluk değerine sahip olan bireylerin hayatta kalması diğerlerinin ise kaldırılması gerekir (Angeline, 1995). Doğal seçim, algoritmanın her iterasyonunda meydana gelir. GA içerisinde rulet tekerleği, rastgele, ağırlıklı, turnuva, hiyerarşik v.b. seçim metodları kullanılabilir. GA-SVM algoritması içerisinde, yaygın olarak kullanılan rulet tekerleği seçim metodu kullanılmıştır (Goldberg ve Deb, 1991). Bu metot, her bir kromozomun uygunluğu ile orantılı bir alana sahip olan bir tekerleğin dönmesine benzer. Bu metoda göre q sayıda bireyin sıralı kümülatif ihtimallerini içeren bir A kümesi ve 0 ile 1 aralığında rastgele olarak üretilen q sayıdan oluşan bir B kümesi oluşturulur. Daha sonra her bir $b_j \in B$ sayısı için $c_i = min\{a_i \in A: a_i \geq b_j\}$ sayısı seçilir. $C = \{c_i\}_{i=1}^{q}$ yeni popülasyon kümesi oluşturulur.

5.1.1.4. Çaprazlama işlemi

Doğal seçim sonucunda üretilen popülasyonu geliştirmek için yeni popülasyondaki bireylere C_R oranlı bir çaprazlama işlemi uygulanır. Genellikle, çaprazlama işlemi uygulanacak olan bireyler rastgele olarak seçilir. GA-SVM algoritmasında ebeveyn kromozomlardan yavru kromozomları elde etmek için düzenli çaprazlama operatörü kullanılmıştır (Sywerda, 1989). Bu operatörü uygulamak için öncelikle *0.5* karıştırma oranlı bir çaprazlama maskesi oluşturulur (Prügel-Bennett, 2001). Bu maske çaprazlama işleminin uygulanacağı ebeveyn kromozomların bitlerini tespit etmek için kullanılır. Çaprazlama maskesi üzerindeki 1 değeri bu bite karşılık gelen SNP'lerin iki ebeveyn arasında çaprazlanacağı, 0 değeri ise bu bite karşılık gelen SNP'lerin değiştirilmeden kalacağı anlamına gelir. Bu algoritmada C_R=*0.9* çaprazlama oranı kullanılmıştır (Lin ve ark., 2003). Birey 1 ve 3 (Şekil 5.2) üzerinde uygulanan düzenli çaprazlama işlemi, Şekil 5.3'de görülmektedir. Bu örnekte 3. satır *0.5* karıştırma oranı ile üretilen çaprazlama maskesidir.

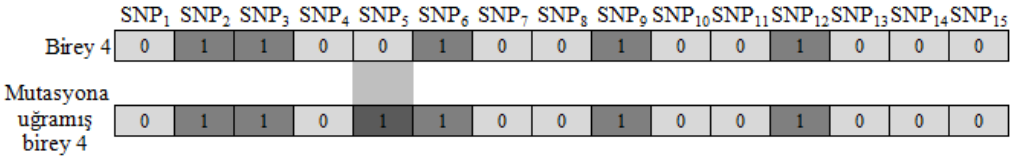
	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	SNP_8	SNP_9	SNP_{10}	SNP_{11}	SNP_{12}	SNP_{13}	SNP_{14}	SNP_{15}
Birey 1	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1
Birey 3	1	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Çaprazlama Maskesi	1	0	0	1	0	1	1	0	0	0	1	0	1	0	1
Yavru 1	1	0	1	0	0	0	1	1	0	0	1	1	0	0	0
Yavru 2	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1

Şekil 5.3. Birey 1 ve 3'e uygulanan düzenli çaprazlama işlemi

5.1.1.5. Mutasyon işlemi

Çaprazlama işlemi sonucunda elde edilen popülasyonu geliştirmek için yeni popülasyondaki bireylere M_R oranı ile mutasyon işlemi uygulanır. Mutasyon işlemi popülasyonun sadece bazı bitlerini değiştirir. Mutasyona uğratılacak bitleri tespit etmek için bütün kromozomlardaki her bir bit pozisyonu için 0 ile 1 arasında rasgele sayılar üretilir. Eğer bir kromozomun herhangi bir biti için üretilen bu sayı, M_R değerinden daha küçük ise bu kromozomdaki ilişkili bit 0 ise 1'e veya 1 ise 0'a değiştirilerek mutasyona uğratılır. GA-SVM algoritması için M_R=*0.01* mutasyon oranı kullanılmıştır

(Lin ve ark., 2003). Şekil 5.4'de birey 4'ün SNP_5'i (Şekil 5.2) üzerinde uygulanan örnek bir mutasyon işlemi gösterilmiştir.



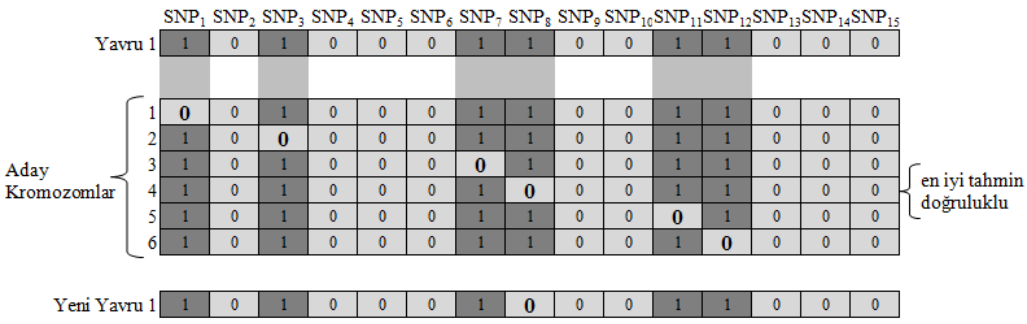
Şekil 5.4. Birey 4'ün SNP_5'i üzerinde uygulanan mutasyon işlemi

5.1.1.6. Etiket SNP'lerin sayısının ayarlanması

Çaprazlama ve mutasyon işlemlerinden sonra kromozomlardaki etiket SNP'leri gösteren 1'lerin sayısı değişebilir. Bu yüzden her bir kromozom için etiket SNP'lerin sayısının (M) algoritmaya giriş olarak verilen etiket SNP'lerin sayısına (N) eşitlenmesi gerekir. Bu problemin çözümü için şu ana kadar iki yaklaşım önerilmiştir (Mahdevar ve ark., 2010; Yang ve ark., 2008). Mahdevar ve ark. (2010)'nın önerdiği rastgele arama algoritmasına göre eğer $M<N$ ise etiket SNP'lerin istenilen sayısına ulaşmak için N-M adet SNP'in rastgele olarak etiket SNP'ler grubuna eklenmesi gerekir. Eğer $M>N$ ise M-N adet etiket SNP'in yine rastgele bir şekilde etiket SNP'ler grubundan çıkarılması gerekir. Ancak etiket SNP'lerin sayısının bu şekildeki ayarlanması işlemi, farklı iterasyonlarda çok farklı tahmin doğruluklarının elde edilmesine neden olur (Yang ve ark., 2008). Bu nedenle Yang ve ark., (2008), farklı bir yaklaşım olan lokal arama algoritmasını önermişlerdir. Lokal arama algoritmasına göre, etiket SNP'lerin sayısı ayarlanırken aday kromozomlar içerisinden en iyi tahmin doğruluklu olan kromozom yeni kromozom olarak belirlenir. Her bir aday kromozom için tahmin doğruluğunun hesaplanması işleminde LOOCV metodu kullanılır.

Bir rastgele arama algoritmasında, iterasyonlar arasındaki tahmin doğruluklarında meydana gelen önemli farkı minimize etmek için iterasyonların sayısının artırılması gerekir. Bu işlem ise etiket SNP'lerin seçim işleminin çok zaman almasına sebep olur (Goldberg, 1989). Benzer şekilde lokal arama algoritmasında ise en iyi tahmin doğruluklu yeni kromozomu bulmak için kullanılan LOOCV metodu çok zaman alır (Yang ve ark., 2008). Bu nedenle çoğu uygulamalar için pratik değildir (Arlot ve Celisse, 2010).

GA-SVM algoritmasında, Yang ve ark. (2008)'nın önerdiği gibi lokal arama algoritması kullanılmıştır. Ancak bu algoritma içinde aday kromozomlar içinden en iyi tahmin doğruluklu yeni kromozomun bulunması işlemi için, LOOCV metodu ile karşılaştırıldığında, hemen hemen aynı sonuçları üretmesine karşılık daha hızlı çalışan 10 kat çapraz doğrulama yöntemi kullanılmıştır (İlhan ve ark., 2011a). 10 kat çapraz doğrulama metodunda, haplotip stringlerini içeren *H* veri kümesi 10 eşit parçaya bölünür ve bu parçalar birer birer çıkarılarak sırasıyla test kümesi olarak kullanılır. Şekil 5.3'deki yavru 1 kromozomunun içerdiği etiket SNP'leri gösteren 1'lerin sayısı *M=6*'dır. Bu sayının algoritmaya giriş olarak verilen *N=5*'e düşürülmesi gerekir. Bunun için bu kromozomdaki 1'ler sırasıyla 0'a dönüştürülerek 6 farklı aday kromozom elde edilir. Bu aday kromozomlardan en iyi tahmin doğruluklu olan kromozom yeni yavru 1 kromozomu olarak belirlenir. Şekil 5.5 bu işlemi göstermektedir.



Şekil 5.5. Yavru 1 kromozomundaki 1'lerin birer birer 0'a dönüştürülerek aday kromozomların oluşturulması

Lokal arama metoduna göre etiket SNP'lerin sayısının ayarlanması işleminde tahmin doğrulukları hesaplanan aday kromozomların toplam sayısı eğer *M<N* ise Denklem 5.2 ile eğer *M>N* ise Denklem 5.3 ile hesaplanır. Bu denklemlerde $k = |N - M|$ şeklindedir.

$$T = \sum_{j=1}^{k} (n - M - j + 1) \tag{5.2}$$

$$T = \sum_{j=1}^{k} (M - j + 1) \tag{5.3}$$

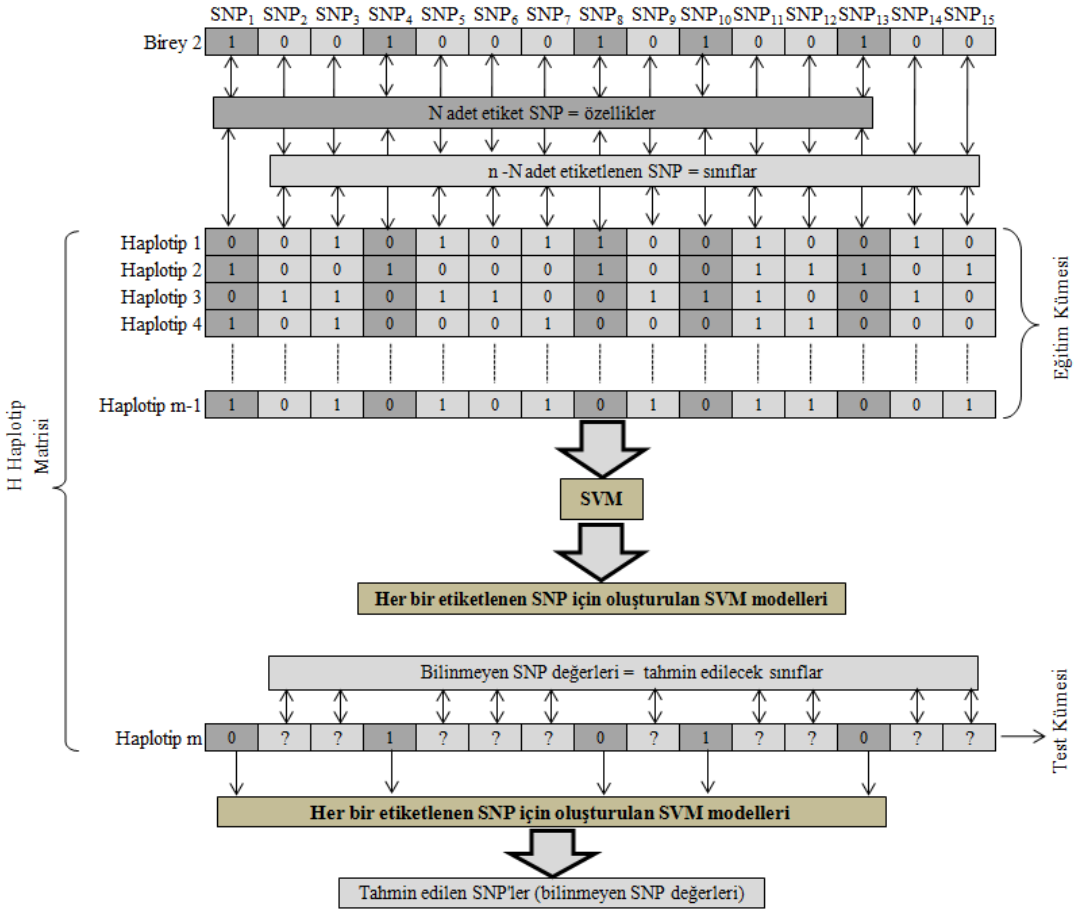
Etiket SNP sayısının ayarlanmasından sonra yeni popülasyondaki bütün bireylerin LOOCV metodu yardımıyla uygunluk değerleri (tahmin doğrulukları) hesaplanır ve en iyi uygunluk değerine sahip birey belirlenir. Bu işlem algoritmaya giriş olarak verilen N_G jenerasyon (iterasyon) sayısı kadar tekrarlanır. Her bir jenerasyon sonucunda elde edilen bireylerden en iyi uygunluk değerine sahip olan birey algoritmanın sonucu olarak geri döndürülür.

5.1.2. SVM tarafından SNP'lerin geri kalanının tahmin edilmesi

SNP'lerin geri kalanlarının değerlerini tahmin etmek için kullanılan yöntemlerden daha önceki bölümlerde bahsedilmişti. Bu yöntemler arasında SVM, yapay sinir ağı gibi diğer veri madenciliği yaklaşımları ile rekabet edebilecek oranda doğru sonuçlar ürettiği için biyoinformatik alanlarında özellikle tercih edilir (Chanda ve ark., 2009; He ve Zelikovsky, 2007). Şekil 5.6'da, SNP'lerin geri kalanının tahmini işlemi gösterilmektedir. Şekilden de görülebildiği gibi SVM ilk olarak, eğitim kümesi olarak verilen haplotiplerdeki SNP değerlerini kullanarak bir model inşa eder. Daha sonra test kümesindeki haplotipe ait olan SNP'lerin geri kalanlarının değerleri (bilinmeyen SNP değerleri) bu model ve GA ile elde edilen etiket SNP'ler kullanılarak tahmin edilir.

SVM tabanlı tahmin yönteminde, H matrisinde bulunan her bir haplotip sırasıyla bir test kümesi olarak düşünülür. Burada her bir etiket SNP belli bir özelliği, SNP'lerin geri kalanının her biri de belli bir sınıfı temsil eder. Doğru olarak tahmin edilen SNP sayısının toplam tahmin edilen SNP sayısına oranı tahmin doğruluğu olarak adlandırılır ve Denklem 5.1'deki eşitlik ile hesaplanır.

SNP'lerin geri kalanının değerlerinin tahmin edilmesi için SVM yazılımı olarak *RBF* (radial basis function) fonksiyonlu *Libsvm* yazılımı kullanıldı (Chang ve Lin, 2011). Libsvm, γ ve C parametreleri ile birlikte kullanılır. γ, RBF çekirdek fonksiyonunun genliğini düzenleyerek SVM'nin genelleştirme yeteneğini kontrol eden bir parametre ve C ise eğitim hatasını minimize etmek ve marjini maksimize etmek arasındaki ayrımı kontrol eden bir parametredir (Cores ve Vapnik, 1995; Vapnik, 1998). *Libsvm* tarafından SNP'lerin geri kalanının değerlerinin tahmin edilmesi için He ve Zelikovsky (2007) tarafından da tavsiye edildiği gibi $\gamma=0.1$ ve $C=0.05$ değerleri kullanılmıştır.

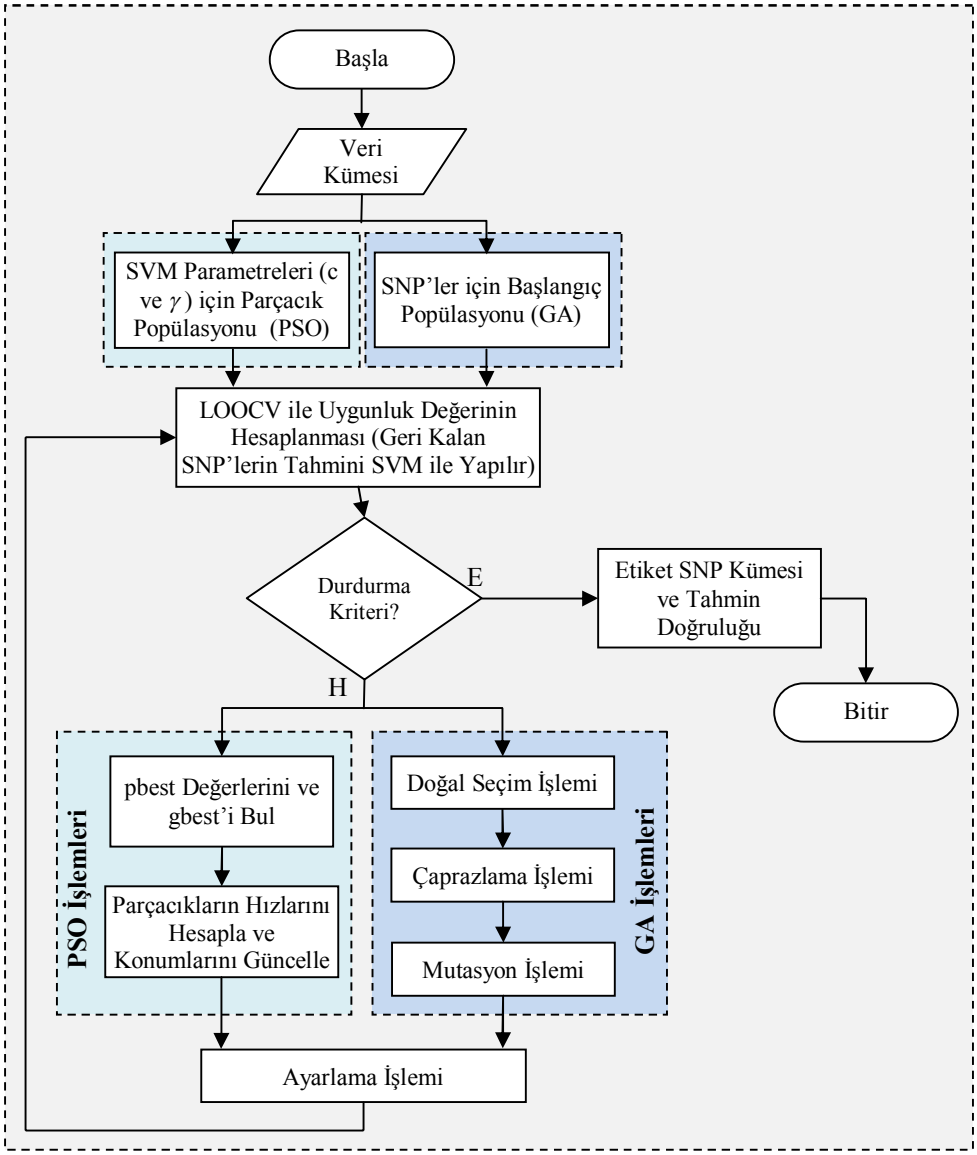


Şekil 5.6. Haplotip m'ye ait SNP'lerin geri kalanının tahmin işlemi

5.2. Parametre Optimizasyonlu GA-SVM Metodu

Geliştirilen parametre optimizasyonlu GA-SVM metodunda da etiket SNP seçim yöntemi olarak GA, etiket SNP dışındaki SNP'lerin tahmin yöntemi olarak da SVM kullanılmıştır. Ayrıca algoritmanın tahmin doğruluğunun değerlendirilmesi için de LOOCV yöntemi kullanılmıştır. Ancak GA-SVM metodundan farklı olarak geliştirilen bu yöntemde, SVM'nin γ ve C parametreleri sabit değerler olarak değil PSO algoritmasıyla (Kennedy ve Eberhart, 1995) optimize edilerek kullanılmıştır. Yapılan deneyler parametre optimizasyonlu GA-SVM metodu ile daha yüksek tahmin doğruluklarının elde edilebileceğini göstermiştir.

Şekil 5.7, parametre optimizasyonlu GA-SVM metodunun akış diyagramını göstermektedir. Şekilden de görülebildiği gibi bu metot başlangıç popülasyonlarının oluşturulması, uygunluk değerinin hesaplanması, pbest değerlerinin ve gbest'in bulunması, parçacık hızlarının hesaplanması ve konumlarının güncellenmesi, doğal seçim, çaprazlama, mutasyon ve ayarlama prosedürlerinden oluşmaktadır.



Şekil 5.7. Parametre optimizasyonlu GA-SVM algoritmasının akış şeması

5.2.1. GA tarafından etiket SNP'lerin seçilmesi

5.2.1.1. SNP'ler için başlangıç popülasyonu

SNP'lerin başlangıç popülasyonu ikili matris ile temsil edilir. Bu matriste popülasyondaki bireyler satırlar, SNP'ler ise sütunlar olarak gösterilir. Algoritmanın girişi, *m* satırlı ve *n* sütunlu ikili bir *H* matrisidir. Bu matris içinde her bir satır belli bir haplotipi, her bir sütun ise belli bir SNP'i temsil eder (He ve Zelikovsky, 2007). Burada, her bir haplotip, n bitli ikili bir vektör ile temsil edilir. Bu matrise dayanarak, GA N_p

satırlı ve *n* sütunlu bir P_{GA} popülasyon matrisi oluşturur. Buradaki N_p değeri genellikle 10 ile 200 arasında rastgele olarak seçilen popülasyondaki bireylerin sayısıdır (Guo ve ark., 2007; Sağ ve Çunkaş, 2009). P_{GA} matrisinin her bir $p_{ij} \in \{0, 1\}$ elemanı, *i.* bireyin *j.* SNP'inin değerini temsil eder. Bu matristeki 1 değeri ilişkili SNP'in etiket SNP olduğunu, 0 değeri ise ilişkili SNP'in etiketlenen SNP olduğunu yani değerinin tahmin edilmesi gerektiğini gösterir. P_{GA} popülasyon matrisindeki bireylerin tamamı *N* ile gösterilen aynı sayıdaki etiket SNP'e sahiptir. Ancak farklı bireyler bu etiket SNP'lerin farklı kombinasyonlarına sahip olabilir. Şekil 5.8'de $N_p=5$ birey ve $n=15$ SNP'den oluşan örnek bir P_{GA} popülasyon matrisi verilmiştir. Bu matriste her bir birey $N=5$ etiket SNP'e sahiptir.

	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	SNP_8	SNP_9	SNP_{10}	SNP_{11}	SNP_{12}	SNP_{13}	SNP_{14}	SNP_{15}
Birey 1	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1
Birey 2	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0
Birey 3	1	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Birey 4	0	1	1	0	0	1	0	0	1	0	0	1	0	0	0
Birey 5	0	0	0	1	0	0	1	1	0	0	0	0	1	0	1

Şekil 5.8. 5 birey ve 15 SNP'den oluşan popülasyon matrisi. Her bir birey 5 etiket SNP ve 10 etiketlenen SNP içerir.

5.2.1.2. Uygunluk değerlendirmesi

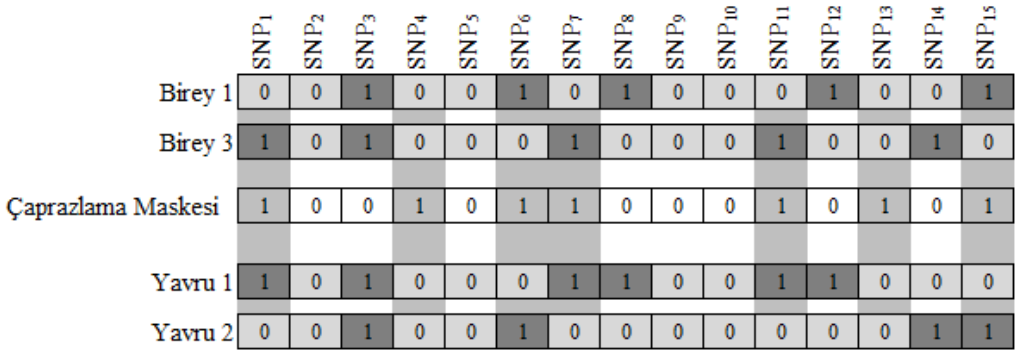
Geliştirilen parametre optimizasyonlu GA-SVM metodunda uygunluk değerlendirme yöntemi olarak LOOCV yöntemi kullanılmıştır (Arlot ve Celisse, 2010; He ve Zelikovsky, 2007; Lee ve Shatkay, 2006; Lin ve Leu, 2010; Yang ve ark., 2008). LOOCV metoduna göre, *j.* iterasyonda öncelikle *j.* haplotip *H* matrisinden kaldırılır. Daha sonra geriye kalan haplotiplerden GA kullanılarak etiket SNP'ler seçilir ve bu seçilen etiket SNP'ler kaldırılan haplotipteki etiketlenen SNP'leri (SNP'lerin geriye kalanını) tahmin etmek için kullanılır. Bu işlem *H* matrisindeki bütün haplotipler doğrulama verisi olarak kullanılıncaya kadar bütün $j=1,2,\ldots,m$ değerleri için tekrarlanır. Doğru olarak tahmin edilen SNP sayısının toplam tahmin edilen SNP sayısına oranı tahmin doğruluğunu (uygunluk değerini) verir ve Denklem 5.1'deki eşitlik ile hesaplanır.

5.2.1.3. Doğal seçim

Popülasyondaki en yüksek uygunluk değerine sahip olan bireylerin hayatta kalması diğerlerinin ise kaldırılması gerekir (Angeline, 1995). Doğal seçim, algoritmanın her iterasyonunda meydana gelir. Parametre optimizasyonlu GA-SVM algoritmasında, yaygın olarak kullanılan rulet tekerleği seçim metodu kullanılmıştır (Goldberg ve Deb, 1991). Bu metoda göre, N_p sayıda bireyin sıralı kümülatif ihtimallerini içeren bir A kümesi ve 0 ile 1 aralığında rastgele olarak üretilen N_p sayıdan oluşan bir B kümesi oluşturulur. Daha sonra her bir $b_j \in B$ sayısı için $c_i = min\{a_i \in A: a_i \geq b_j\}$ sayısı seçilir. Sonuç olarak $C = \{c_i\}_{i=1}^{q}$ yeni popülasyon kümesi oluşturulur.

5.2.1.4. Çaprazlama işlemi

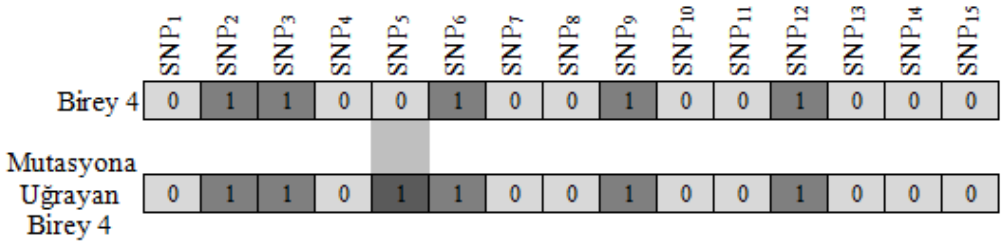
Doğal seçim sonucunda üretilen popülasyonu geliştirmek için, yeni popülasyondaki bireylere C_R oranlı bir çaprazlama işlemi uygulanır. Genellikle çaprazlama işlemi uygulanacak olan bireyler rastgele seçilir. Parametre optimizasyonlu GA-SVM algoritmasında ebeveyn kromozomlardan yavru kromozomları elde etmek için düzenli çaprazlama operatörü kullanılmıştır (Sywerda, 1989). Bu operatörü uygulamak için öncelikle, *0.5* karıştırma oranlı bir çaprazlama maskesi oluşturulur (Prügel-Bennett, 2001). Bu maske, çaprazlama işleminin uygulanacağı ebeveyn kromozomların bitlerini tespit etmek için kullanılır. Çaprazlama maskesi üzerindeki 1 değeri bu bite karşılık gelen SNP'lerin iki ebeveyn arasında çaprazlanacağı, 0 değeri ise bu bite karşılık gelen SNP'lerin değiştirilmeden kalacağı anlamına gelir. Bu algoritmada $C_R = 0.9$ çaprazlama oranı kullanılmıştır (Lin ve ark., 2003). Birey 1 ve 3 (Şekil 5.8) üzerinde uygulanan düzenli çaprazlama işlemi Şekil 5.9'da görülmektedir. Bu örnekte 3. satır *0.5* karıştırma oranı ile üretilen çaprazlama maskesidir.



Şekil 5.9. Birey 1 ve 3 üzerinde uygulanan düzenli çaprazlama işlemi ve elde edilen yavru bireyler

5.2.1.5. Mutasyon işlemi

Çaprazlama işlemi sonucunda elde edilen popülasyonu geliştirmek için popülasyondaki bireylere M_R oranı ile mutasyon işlemi uygulanır. Mutasyon işlemi, popülasyonun sadece bazı bitlerini değiştirir. Mutasyona uğratılacak bitleri tespit etmek için bütün kromozomlardaki her bir bit pozisyonu için 0 ile 1 arasında rasgele sayılar üretilir. Eğer bir kromozomun herhangi bir biti için üretilen bu sayı M_R değerinden daha küçük ise bu kromozomdaki ilişkili bit 0 ise 1'e veya 1 ise 0'a değiştirilerek mutasyona uğratılır. Bu algoritma için, M_R=*0.01* mutasyon oranı kullanılmıştır (Lin ve ark., 2003). Şekil 5.10'da birey 4'ün SNP_5'i (Şekil 5.8) üzerinde uygulanan örnek bir mutasyon işlemi gösterilmiştir.



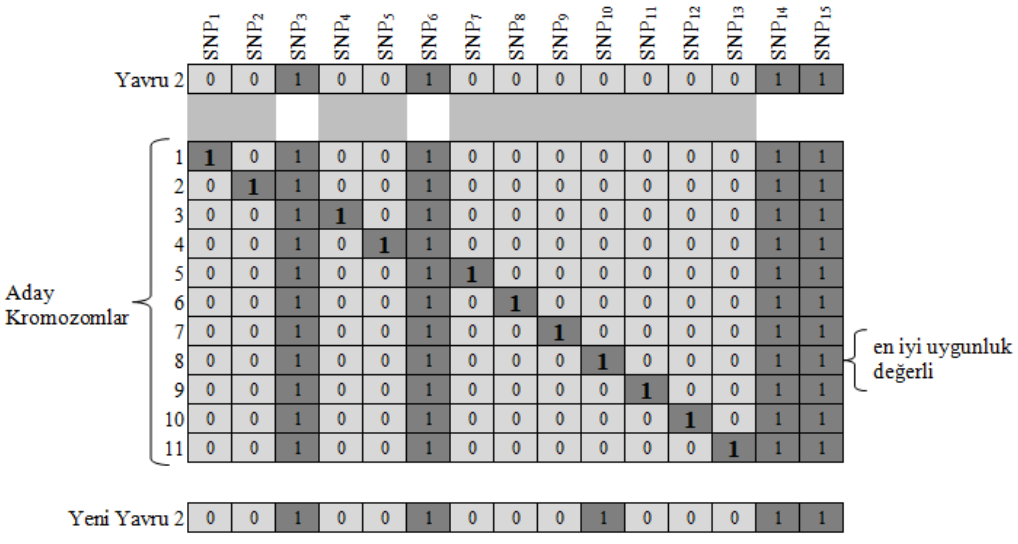
Şekil 5.10. Birey 4'ün SNP_5'i üzerinde uygulanan mutasyon işlemi

5.2.1.6. Etiket SNP'lerin sayısının ayarlanması

Çaprazlama ve mutasyon işlemlerinden sonra kromozomlardaki etiket SNP'leri gösteren 1'lerin sayısı değişebilir. Bu yüzden her bir kromozom için etiket SNP'lerin sayısının (M) parametre optimizasyonlu GA-SVM algoritmasına giriş olarak verilen etiket SNP'lerin sayısına (N) eşitlenmesi gerekir. Bu problemin çözümü için şu ana

kadar rastgele arama ve lokal arama olmak üzere iki yaklaşım önerilmiştir (Mahdevar ve ark., 2010; Yang ve ark., 2008). Bölüm 4.1.1.6'da da bahsedildiği gibi rastgele arama algoritmasında, iterasyonlar arasındaki tahmin doğruluklarında meydana gelen önemli farkı minimize etmek için iterasyonların sayısının artırılması gerekir. Bu işlem ise etiket SNP'lerin seçim işleminin çok zaman almasına sebep olur (Goldberg, 1989). Lokal arama algoritmasında ise en iyi tahmin doğruluklu yeni kromozomu bulmak için kullanılan LOOCV metodu çok zaman alır (Yang ve ark., 2008). Bu nedenle çoğu uygulamalar için pratik değildir (Arlot ve Celisse, 2010).

GA-SVM algoritmasında olduğu gibi parametre optimizasyonlu GA-SVM algoritmasında da Yang ve ark. (2008)'nın önerdiği gibi lokal arama algoritması kullanılmıştır. Ancak bu algoritma içinde aday kromozomlar içerisinde en iyi tahmin doğruluklu yeni kromozomun bulunması işlemi için 10 kat çapraz doğrulama yöntemi kullanılmıştır (İlhan ve ark., 2011b). Şekil 5.9'daki yavru 2 kromozomunun içerdiği etiket SNP'leri gösteren 1'lerin sayısı $M=4$'dür. Bu sayının algoritmaya giriş olarak verilen $N=5$'e eşitlenmesi gerekir. Bunun için bu kromozomdaki 0'ler sırasıyla 1'e dönüştürülerek 11 farklı aday kromozom elde edilir. Bu aday kromozomlardan en iyi tahmin doruluklu olan kromozom yeni yavru 2 kromozomu olarak belirlenir. Şekil 5.11 bu işlemi göstermektedir.



Şekil 5.11. Yavru 2 kromozomundaki 0'ların sırayla 1'e dönüştürülerek aday kromozomların oluşturulması

Lokal arama metoduna göre, etiket SNP'lerin sayısının ayarlanması işleminde tahmin doğrulukları hesaplanan aday kromozomların toplam sayısı Denklem 5.2 ve Denklem 5.3 ile hesaplanır.

Etiket SNP sayısının ayarlanmasından sonra yeni popülasyondaki bütün bireylerin LOOCV metodu yardımıyla uygunluk değerleri (tahmin doğrulukları) hesaplanır ve en iyi uygunluk değerine sahip birey belirlenir. Bu işlem algoritmaya giriş olarak verilen N_G jenerasyon (iterasyon) sayısı kadar tekrarlanır. Her bir jenerasyon sonucunda elde edilen bireylerden en iyi uygunluk değerine sahip olan birey algoritmanın sonucu olarak geri döndürülür.

5.2.2. PSO tarafından SVM parametrelerinin (C ve γ) optimize edilmesi

Parametre optimizasyonlu GA-SVM algoritmasında da etiket SNP'lerin değerlerini kullanarak etiketlenen SNP'lerin değerlerinin tahmin edilmesi için SVM ve SVM sınıflayıcısı içerisinde de en yaygın çekirdek fonksiyonu olan *RBF* (radial basis function) fonksiyonu kullanılmıştır. RBF, γ ve C parametreleri ile birlikte kullanılır. γ RBF çekirdek fonksiyonunun genliğini düzenleyerek SVM'nin genelleştirme yeteneğini kontrol eden bir parametre ve C ise eğitim hatasını minimize etmek ve marjini maksimize etmek arasındaki ayrımı kontrol eden bir parametredir (Cores ve Vapnik, 1995; Vapnik, 1998). Ancak verilen bir problem için hangi C ve γ değerlerinin en iyi olduğu daha önceden bilinmez. Bu nedenle bir takım parametre ayarlamalarının yapılması gerekir (Hsu ve ark., 2010). Buradaki amaç, sınıflayıcının bilinmeyen veriyi doğru olarak tahmin edebilmesi için gerekli en iyi C ve γ parametrelerini belirlemektir (Hsu ve ark., 2010). Bu amaca ulaşmak için parametre optimizasyonlu GA-SVM algoritmasında PSO algoritması kullanılmıştır.

5.2.2.1. SVM parametreleri için başlangıç popülasyonu

SVM parametreleri için başlangıç popülasyonu bir matris ile temsil edilir. Bu matristeki satırlar popülasyondaki parçacıkları, sütunlar ise SVM parametrelerini temsil eder. PSO algoritması iki sütunlu ve N_p satırlı P_{PSO} popülasyon matrisini oluşturur. Buradaki sütunlar C ve γ parametrelerini, satırlar ise N_p tane parçacığı temsil eder. P_{PSO} parçacık matrisinin satır sayısı GA tarafından oluşturulan P_{GA} popülasyon matrisinin

satır sayısına eşittir. P_{PSO} matrisinin her bir $p_{ij} \in [0, 1]$ elemanı *i.* parçacık için C ve γ parametrelerinin değerlerini temsil eder. Şekil 5.12'de, örnek bir popülasyon matrisi verilmiştir. Şekilden de görülebildiği gibi bu popülasyon matrisi 5 parçacık ve iki parametreden (C ve γ) oluşur.

	C	γ
Parçacık 1	0,08	0,5
Parçacık 2	0,1	0,35
Parçacık 3	0,05	0,06
Parçacık 4	0,7	0,83
Parçacık 5	0,95	0,2

Şekil 5.12. Parçacık popülasyon matrisi 5 parçacık içerir. Her bir parçacık C ve γ parametrelerini içerir.

5.2.2.2. Uygunluk değerlendirmesi

Parametre optimizasyonlu GA-SVM yönteminde, PSO algoritmasının uygunluk fonksiyonu olarak GA için hesaplanan uygunluk fonksiyonu (Denklem 5.1) kullanılır. P_{GA} popülasyon matrisindeki (GA için popülasyon matrisi) her bir birey için hesaplanan uygunluk değeri, aynı zamanda P_{PSO} popülasyon matrisindeki (PSO için popülasyon matrisi) karşılık gelen parçacığın uygunluk değeri olarak kullanılır. Örneğin, Şekil 5.8'deki birey 2'nin hesaplanan uygunluk değeri 0.93 ise bu değer aynı zamanda Şekil 5.12'deki parçacık 2'nin uygunluk değeridir.

5.2.2.3. pbest değerlerinin ve gbest'in bulunması

PSO algoritmasının her iterasyonunda, her bir parçacık iki "en iyi" değere göre güncellenir. Bunlardan ilki, bir parçacığın o ana kadar bulduğu en iyi uygunluk değeridir. Ayrıca bu değer daha sonra kullanılmak üzere hafızada tutulur ve "*pbest*" yani parçacığın en iyi değeri olarak isimlendirilir. Diğeri ise popülasyondaki herhangi bir parçacık tarafından o ana kadar elde edilmiş en iyi uygunluk değeridir. Bu değer popülasyon için global en iyi değerdir ve "*gbest*" olarak isimlendirilir.

5.2.2.4. Parçacık hızlarının hesaplanması ve konumlarının güncellenmesi

Bir önceki adımda bulunan parçacıkların en iyi değerleri (pbest) ve global en iyi değere (gbest) göre, popülasyondaki her bir parçacığın hızları hesaplanır ve konumları güncellenir. Parçacıkların hızlarının hesaplanması ve konumlarının güncellenmesi için sırasıyla Denklem 5.4 ve Denklem 5.5 kullanılır.

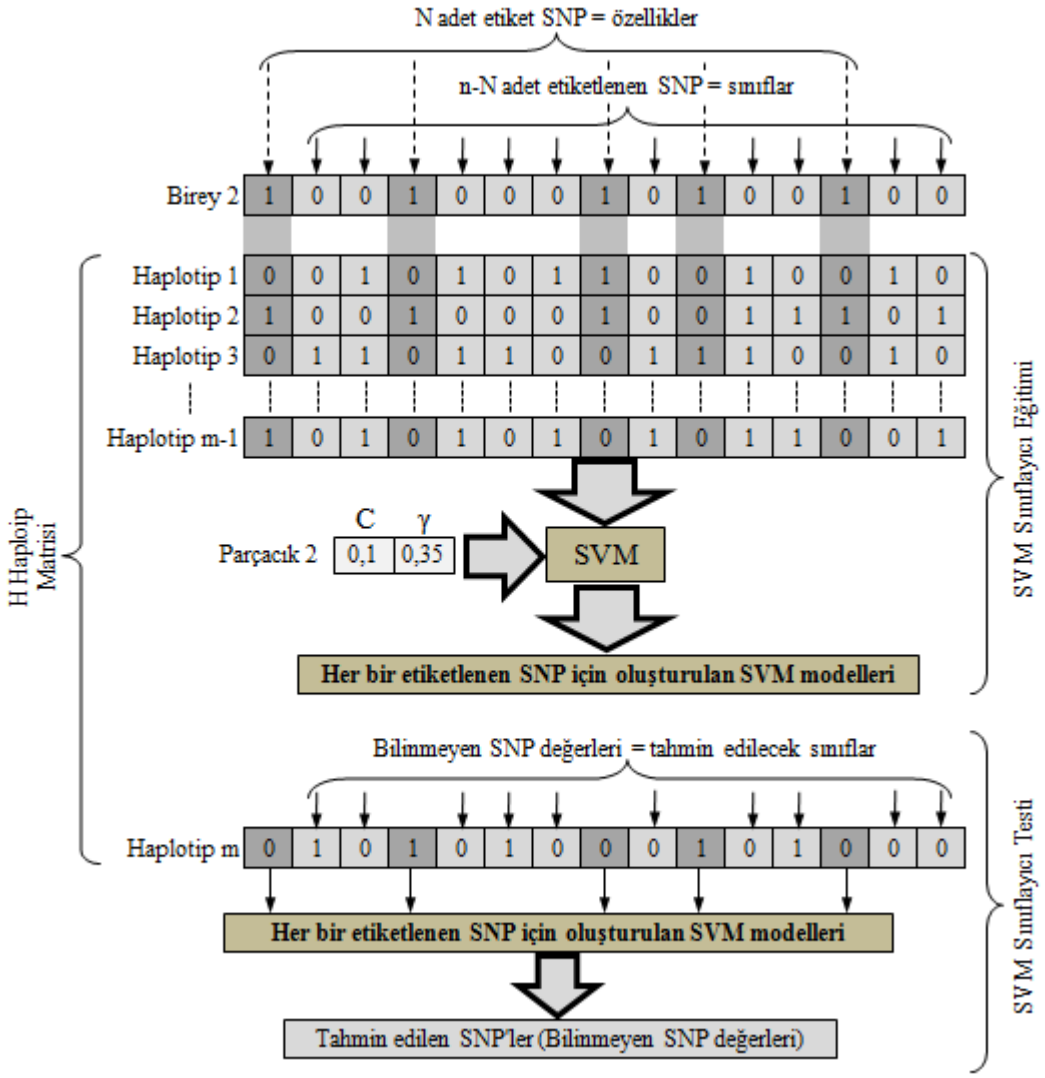
$$v_{ij}^{k+1} = w\ x\ v_{ij}^{k} + \ c_1\ x\ r_1\ x\ \left(pbest_{ij}^{k} - x_{ij}^{k}\right) + c_2\ x\ r_2\ x\ (gbest_{j}^{k} - x_{ij}^{k}) \quad (5.4)$$

$$x_{ij}^{k+1} = x_{ij}^{k} + \ v_{ij}^{k+1} \quad (5.5)$$

Burada $i=1,2,...,N_p$, $k=1,2,...,N_G$ ve $j=1,2$'dir ve N_p popülasyonun büyüklüğünü, N_G iterasyon sayısını ve 2 problem uzayının boyutunu gösterir (C ve γ). v_{ij}^{k} ve x_{ij}^{k}, sırasıyla, *i.* parçacığın hızı ve çözümüdür (pozisyonu). $pbest_{ij}^{k}$, *i.* parçacığın şimdiye kadar ulaşılan en iyi çözümü ve $gbest_{j}^{k}$ ise popülasyon içindeki herhangibir parçacık tarafından şimdiye kadar ulaşılan global en iyi çözümdür. *c1* ve *c2* öğrenme faktörleridir ve sırasıyla parçacığın kendi tecrübelerine göre ve sürüdeki diğer parçacıkların tecrübelerine göre hareketini yönlendirir. Bu çalışmada *c1* ve *c2* öğrenme faktörlerinin her ikiside *2* olarak alınmıştır (Reider ve ark., 1999). *r1* ve *r2*, *[0,1]* aralığındaki rastgele değerlerdir. *w* atalet ağırlığıdır ve küçük atalet ağırlığı local aramaya olanak tanırken büyük atalet ağırlığı global aramaya imkan verir. w atalet ağırlığının 0.8, 1 ve 1.2 değerleri ile yaptığımız deneyler, bu parametrenin 1 değeri için elde ettiğimiz sonuçların (tahmin doğruluğu) diğerlerine göre daha iyi olduğunu göstermiştir. Bu nedenle, bu çalışmada mümkün olduğu kadar daha iyi bir çözüme ulaşmak için *w=1* olarak alınmıştır (Song ve Eberhart, 1998).

5.2.3. SVM tarafından SNP'lerin geri kalanının tahmin edilmesi

Şekil 5.13'de, SNP'lerin geri kalanının tahmini işlemi gösterilmektedir. Şekilden de görülebildiği gibi SVM ilk olarak eğitim kümesi olarak verilen haplotiplerdeki SNP değerlerini kullanarak bir model inşa eder. Daha sonra test kümesindeki haplotipe ait olan SNP'lerin geri kalanlarının değerleri (bilinmeyen SNP değerleri), bu model ve GA ile elde edilen etiket SNP'ler kullanılarak tahmin edilir.



Şekil 5.13. Haplotip m'ye ait SNP'lerin geri kalanının tahmin işlemi

SVM tabanlı tahmin yönteminde, *H* matrisinde bulunan her bir haplotip sırasıyla bir test kümesi olarak düşünülür. Burada her bir etiket SNP belli bir özelliği, SNP'lerin geri kalanının her biri de belli bir sınıfı temsil eder. Doğru olarak tahmin edilen SNP sayısının toplam tahmin edilen SNP sayısına oranı tahmin doğruluğu olarak adlandırılır ve Denklem 5.1'deki eşitlik ile hesaplanır.

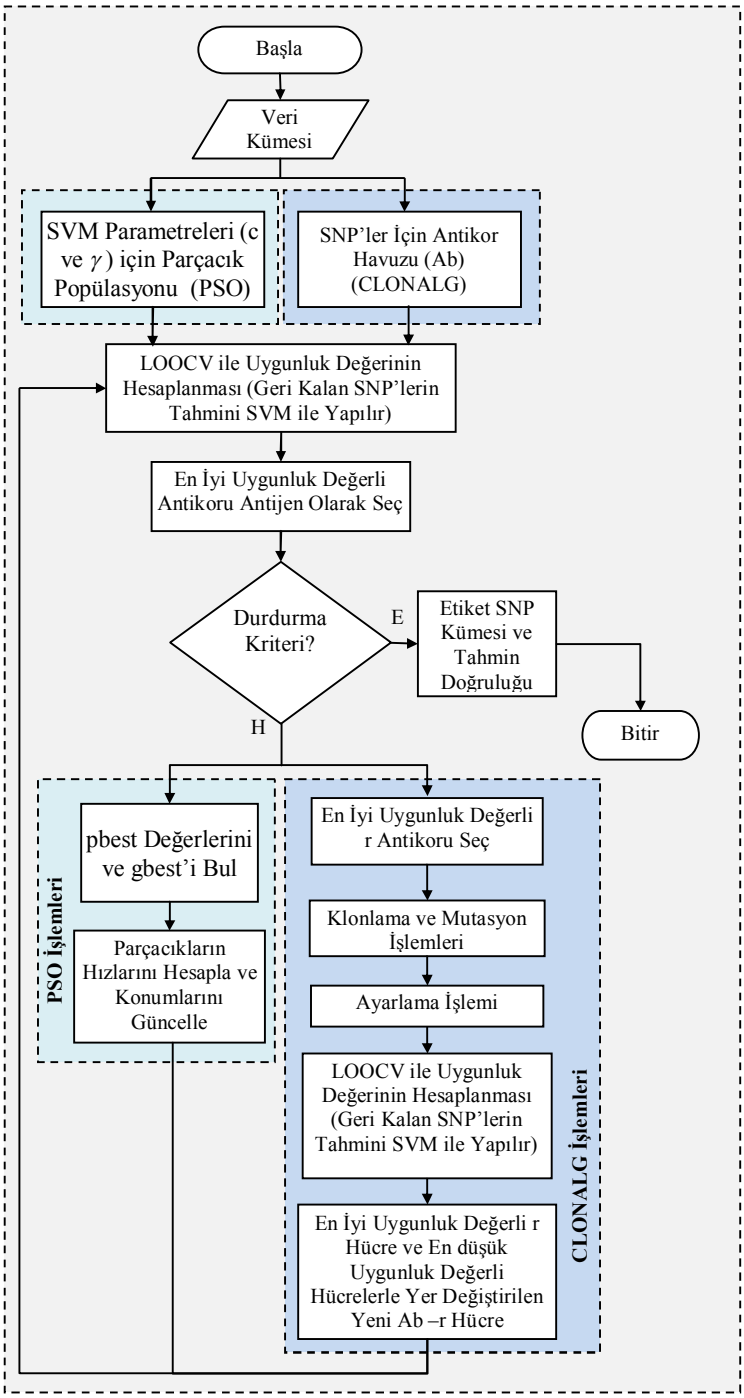
SNP'lerin geri kalanının değerlerinin tahmin edilmesi için SVM yazılımı olarak *RBF* (radial basis function) fonksiyonlu *Libsvm* yazılımı kullanılmıştır (Chang ve Lin, 2011). Libsvm, *C* ve γ parametreleri ile birlikte kullanılır. Parametre optimizasyonlu GA-SVM metodunda PSO algoritması bu iki parametrenin optimizasyonu için kullanılmıştır.

GA tarafından oluşturulan popülasyon matrisindeki bir bireyin uygunluk değeri hesaplanırken PSO tarafından oluşturulan parçacık popülasyon matrisinde aynı satıra karşılık gelen parçacıktaki C ve γ parametreleri kullanılır. Örneğin Şekil 5.13'deki birey 2'nin uygunluk değerini hesaplamak için parçacık popülasyon matrisindeki parçacık 2 kullanılır. Hesaplanan bu değer hem GA için birey 2'nin uygunluk değeri hem de PSO için parçacık 2'nin uygunluk değeri olarak kabul edilir.

5.3. Parametre Optimizasyonlu CLONTagger Metodu

Etiket SNP'lerin tahmin doğruluğunu artırmak için yeni hibrit bir metot önerilmiş ve bu metot parametre optimizasyonlu CLONTagger metodu olarak adlandırılmıştır. Bu yöntemde, etiket SNP'leri seçmek için klonal seçim algoritması (De Castro ve Von Zuben, 1999) ve geri kalan SNP'lerin değerlerini tahmin etmek için de SVM (Vapnik, 1998) kullanılmıştır. SVM'nin C ve γ parametreleri PSO (Kennedy and Eberhart, 1995) tarafından optimize edilmiştir.

Şekil 5.14'deki akış diyagramından da görülebildiği gibi bu metot başlangıç popülasyonlarının oluşturulması, uygunluk değerinin hesaplanması, pbest değerlerinin ve gbest'in bulunması, parçacık hızlarının hesaplanması ve konumlarının güncellenmesi, klonlama, mutasyon ve ayarlama prosedürlerinden oluşmaktadır.



Şekil 5.14. Parametre optimizasyonlu CLONTagger metodunun akış şeması

5.3.1. CLONALG tarafından etiket SNP'lerin seçilmesi

5.3.1.1. SNP'ler için başlangıç antikor havuzu

SNP'ler için başlangıç antikor havuzu, ikili bir matris ile temsil edilir. Bu matriste satırlar popülasyondaki antikorları, sütunlar ise SNP'leri temsil eder.

Algoritmanın girişi *m* satırlı ve *n* sütunlu ikili bir *H* matrisidir. Bu matris içinde her bir satır belli bir haplotipi, her bir sütun ise belli bir SNP'i temsil eder (He ve Zelikovsky, 2007). Bu matrise dayanarak, CLONALG algoritması, *Ab* satırlı ve *n* sütunlu bir P_{CSA} popülasyon matrisi oluşturur. Buradaki *Ab* değeri popülasyondaki antikorların sayısıdır. P_{CSA} matrisinin her bir $p_{ij} \in \{0, 1\}$ elemanı, *i.* antikorun *j.* SNP'inin değerini temsil eder. Bu matristeki 1 değeri ilişkili SNP'in etiket SNP olduğunu, 0 değeri ise ilişkili SNP'in etiketlenen SNP olduğunu yani değerinin tahmin edilmesi gerektiğini gösterir (İlhan ve ark., 2011a). P_{CSA} popülasyon matrisindeki antikorların tamamı *N* ile gösterilen aynı sayıdaki etiket SNP'e sahiptir. Ancak farklı antikorlar bu etiket SNP'lerin farklı kombinasyonlarına sahip olabilir. Şekil 5.15'de *Ab=5* antikor ve *n=15* SNP'den oluşan örnek bir P_{CSA} popülasyon matrisi verilmiştir. Bu matriste her bir antikor *N=5* etiket SNP'e sahiptir.

	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	SNP_8	SNP_9	SNP_{10}	SNP_{11}	SNP_{12}	SNP_{13}	SNP_{14}	SNP_{15}
Antikor 1	0	0	0	1	0	0	1	1	0	0	1	0	0	1	0
Antikor 2	1	0	0	1	0	0	0	1	0	1	0	0	1	0	0
Antikor 3	1	0	1	0	0	1	0	0	0	0	1	0	0	0	1
Antikor 4	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0
Antikor 5	0	0	0	0	1	0	1	1	0	0	0	1	0	0	1

Şekil 5.15. Örnek bir popülasyon matrisi

5.3.1.2. Uygunluk değerlendirmesi

Parametre optimizasyonlu CLONTagger metodunda, LOOCV yöntemi uygunluk değerlendirme yöntemi olarak kullanılmıştır. LOOCV metoduna göre, *j.* iterasyonda öncelikle *j.* haplotip *H* matrisinden kaldırılır. Daha sonra geriye kalan haplotiplerden CLONALG kullanılarak etiket SNP'ler seçilir ve bu seçilen etiket SNP'ler kaldırılan haplotipteki etiketlenen SNP'leri (SNP'lerin geriye kalanını) tahmin etmek için kullanılır. Bu işlem, *H* matrisindeki bütün haplotipler doğrulama verisi olarak kullanılıncaya kadar bütün $j=1,2,\ldots,m$ değerleri için tekrarlanır. Tahmin doğruluğu (uygunluk değeri) Denklem 5.1'deki eşitlik ile hesaplanır.

5.3.1.3. Klonlama İşlemi

Popülasyondaki antikorların en iyi uygunluk değerine sahip olan ilk *r* tanesine klonlama işlemi uygulanır. Parametre optimizasyonlu CLONTagger algoritmasında her bir antikor için üretilecek olan klon sayısı Denklem 5.6'da gösterilen formül kullanılarak hesaplanır.

$$C = round(\frac{\beta . Ab}{i}) \tag{5.6}$$

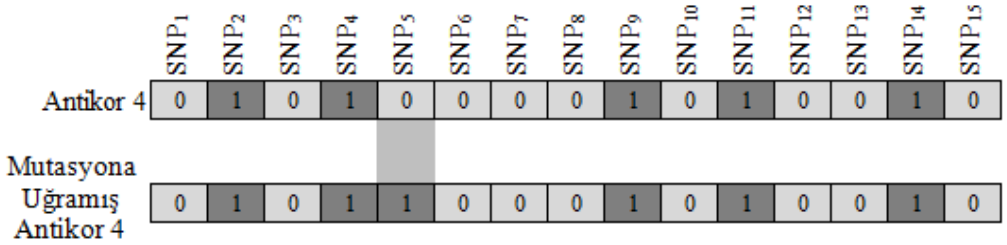
Burada β bir klonlama faktörüdür. *Ab,* antikor havuzundaki antikorların sayısı ve *i,* o anki antikorun uygunluk değeri açısından sırasını gösterir. Parametre optimizasyonlu CLONTagger algoritmasının deneysel sonuçları için β klonlama faktörü *1* ve *Ab* antikor sayısı da *20* olarak alınmıştır.

5.3.1.4. Mutasyon işlemi

Klonlanan antikorlara mutasyon işlemi uygulanır. Öncelikle klonlanan antikorlarla antijen arasındaki duyarlılıklar hesaplanır. Duyarlılıkların hesaplanmasında Denklem 5.7'de verilen Hamming mesafe formülü kullanılır. Burada p_{ij}, P_{CSA} popülasyonundaki *i.* antikorun *j.* SNP'ini, Ag_j ise antijenin *j.* SNP'ini gösterir.

$$D_i = \sum_{j=1}^{n} \delta, \quad \delta = \begin{Bmatrix} 1 & p_{ij} \neq Ag_j \\ 0 & Diğer \end{Bmatrix} \tag{5.7}$$

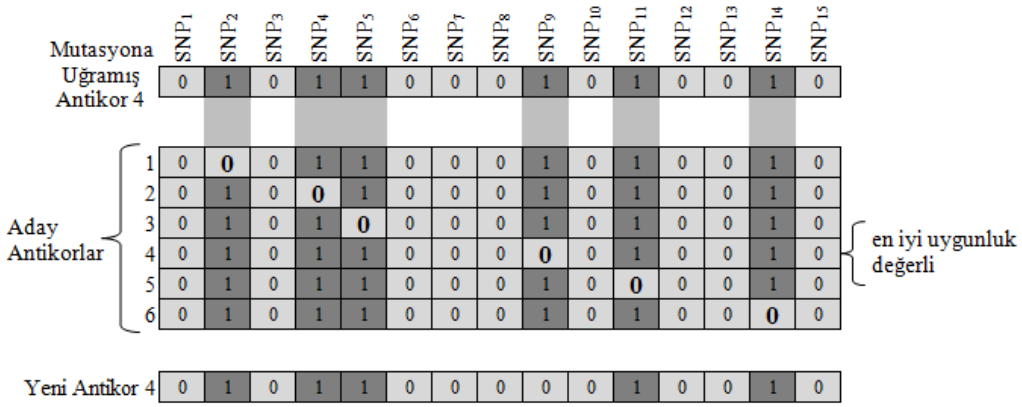
Klonlanan bütün antikorlar için hesaplanan duyarlılık değerleri ile orantılı olarak mutasyon işlemi uygulanır. Önerilen parametre optimizasyonlu CLONTagger algoritmasında, duyarlılık değeri yüksek olan antikorların daha fazla sayıda biti, duyarlılık değeri düşük olan antikorların ise daha az sayıda biti mutasyon işlemine tabi tutulur. Örneğin 2 duyarlılık değerine sahip bir antikorun sadece tek bir biti için mutasyon işlemi uygulanır. Hangi bitin mutasyona uğrayacağı ise algoritmanın girişine uygulanan SNP sayısına göre rastgele olarak belirlenir. Şekil 5.15'deki antikor 4'ün SNP_5'i üzerindeki mutasyon işlemi Şekil 5.16'da verilmektedir.



Şekil 5.16. Örnek bir mutasyon işlemi

5.3.1.5. Etiket SNP'lerin sayısının ayarlanması

Mutasyon işleminden sonra kromozomlardaki etiket SNP'leri gösteren 1'lerin sayısı değişebilir. Bu yüzden her bir antikor için etiket SNP'lerin sayısının (*M*) parametre optimizasyonlu CLONTagger algoritmasına giriş olarak verilen etiket SNP'lerin sayısına (*N*) eşitlenmesi gerekir. Bu problemin çözümünde de GA-SVM ve parametre optimizasyonlu GA-SVM algoritmasındaki gibi lokal arama algoritması kullanılmıştır. Yine bu algoritma içinde de aday antikorlar içerisinde en iyi tahmin doğruluklu yeni antikorun bulunması işlemi için 10 kat çapraz doğrulama yöntemi kullanılmıştır (İlhan ve ark., 2011a). Şekil 5.16'daki örnekte mutasyona uğramış antikor 4 içindeki etiket SNP'leri gösteren 1'lerin sayısı *M=6*'dır. Bu sayının algoritmaya giriş olarak verilen *N=5*'e eşitlenmesi gerekir. Bunun için, bu antikordaki 1'ler sırasıyla 0'a dönüştürülerek 6 farklı aday antikor elde edilir. Bu aday antikorlardan en iyi tahmin doğruluklu olan antikor yeni antikor 4 olarak belirlenir. Şekil 5.17 bu işlemi göstermektedir.



Şekil 5.17. Mutasyona uğramış antikor 4 üzerindeki etiket SNP'lerin sayısının ayarlanması işlemi

Lokal arama metoduna göre etiket SNP'lerin sayısının ayarlanması işleminde, tahmin doğrulukları hesaplanan aday antikorların toplam sayısı Denklem 5.2 ve Denklem 5.3 ile hesaplanır.

Etiket SNP sayısının ayarlanmasından sonra yeni popülasyondaki bütün antikorların LOOCV metodu yardımıyla uygunluk değerleri (tahmin doğrulukları) hesaplanır ve daha sonra en yüksek uygunluk değerine sahip *r* sayıdaki antikor ile en düşük uygunluk değerli antikorlar ile yer değiştirilmek için üretilen *Ab-r* sayıdaki antikor ile yeni popülasyon oluşturulur. Bu işlem algoritmaya giriş olarak verilen N_G jenerasyon (iterasyon) sayısı kadar tekrarlanır. Her bir jenerasyon sonucunda elde edilen antikorlardan en iyi uygunluk değerine sahip olan antikor algoritmanın sonucu olarak geri döndürülür.

5.3.2. PSO tarafından SVM parametrelerinin (C ve γ) optimize edilmesi

Bu çalışmada, etiket SNP'lerin değerlerini kullanarak etiketlenen SNP'lerin değerlerinin tahmin edilmesi için SVM ve SVM sınıflayıcısı içerisinde de çekirdek fonksiyonu olarak *RBF* (radial basis function) fonksiyonu kullanılmıştır. RBF, *C* ve γ parametreleri ile birlikte kullanılır. Daha öncede bahsedildiği gibi verilen bir problem için hangi *C* ve γ değerlerinin en iyi olduğu daha önceden bilinmez. Bu nedenle bir takım parametre ayarlamalarının yapılması gerekir (Hsu ve ark., 2010). Buradaki amaç, sınıflayıcının bilinmeyen veriyi doğru olarak tahmin edebilmesi için gerekli en iyi *C* ve γ parametrelerini belirlemektir (Hsu ve ark., 2010). Bu amaça ulaşmak için parametre optimizasyonlu CLONTagger yönteminde PSO algoritması kullanılmıştır.

5.3.2.1. SVM parametreleri için başlangıç popülasyonu

SVM parametreleri için başlangıç popülasyonu bir matris ile temsil edilir. Bu matristeki satırlar popülasyondaki parçacıkları, sütunlar ise SVM parametrelerini temsil eder. PSO algoritması, iki sütunlu ve N_p satırlı P_{PSO} popülasyon matrisini oluşturur. Buradaki sütunlar *C* ve γ parametrelerini, satırlar ise N_p tane parçacığı temsil eder. P_{PSO} parçacık matrisinin satır sayısı CLONALG tarafından oluşturulan P_{CSA} popülasyon matrisinin satır sayısına eşittir. P_{PSO} matrisinin her bir $p_{ij} \in [0, 1]$ elemanı, *i.* parçacık için *C* ve γ parametrelerinin değerlerini temsil eder. Şekil 5.18'de örnek bir popülasyon

matrisi verilmiştir. Şekilden de görülebildiği gibi bu popülasyon matrisi 5 parçacık ve iki parametreden (C ve γ) oluşur.

	C	γ
Parçacık 1	0,08	0,5
Parçacık 2	0,1	0,35
Parçacık 3	0,05	0,06
Parçacık 4	0,7	0,83
Parçacık 5	0,95	0,2

Şekil 5.18. Örnek bir parçacık popülasyon matrisi

5.3.2.2. Uygunluk değerlendirmesi

Parametre optimizasyonlu CLONTagger yönteminde, PSO algoritmasının uygunluk fonksiyonu olarak CLONALG için hesaplanan uygunluk fonksiyonu kullanılır. P_{CSA} popülasyon matrisindeki (CLOANLG için popülasyon matrisi) her bir antikor için hesaplanan uygunluk değeri aynı zamanda P_{PSO} popülasyon matrisindeki (PSO için popülasyon matrisi) karşılık gelen parçacığın uygunluk değeri olarak kullanılır. Örneğin, Şekil 5.15'deki antikor 2'nin hesaplanan uygunluk değeri 0.87 ise bu değer aynı zamanda Şekil 5.18'deki parçacık 2'nin uygunluk değeridir.

5.3.2.3. pbest değerlerinin ve gbest değerinin bulunması

PSO algoritmasının her iterasyonunda, her bir parçacık, iki "en iyi" değere göre güncellenir. Bunlardan ilki bir parçacığın o ana kadar bulduğu en iyi uygunluk değeridir. Ayrıca bu değer daha sonra kullanılmak üzere hafızada tutulur ve "*pbest*" yani parçacığın en iyi değeri olarak isimlendirilir. Diğeri ise popülasyondaki herhangi bir parçacık tarafından o ana kadar elde edilmiş en iyi uygunluk değeridir. Bu değer popülasyon için global en iyi değerdir ve "*gbest*" olarak isimlendirilir.

5.3.2.4. Parçacık hızlarının hesaplanması ve konumlarının güncellenmesi

Bir önceki adımda bulunan parçacıkların en iyi değerleri (pbest) ve global en iyi değere (gbest) göre, popülasyondaki her bir parçacığın hızları hesaplanır ve konumları

güncellenir. Parçacıkların hızlarının hesaplanması ve konumlarının güncellenmesi için sırasıyla Denklem 5.4 ve Denklem 5.5 kullanılır.

Bu denklemlerde $i=1,2,...,Ab$, $k=1,2,...,N_G$ ve $j=1,2$'dir ve Ab popülasyonun büyüklüğünü, N_G iterasyon sayısını ve 2 problem uzayının boyutunu gösterir (C ve γ). v_{ij}^k ve x_{ij}^k, sırasıyla, *i.* parçacığın hızı ve çözümüdür (pozisyonu). $pbest_{ij}^k$, *i.* parçacığın şimdiye kadar ulaşılan en iyi çözümü ve $gbest_j^k$ ise popülasyon içindeki herhangibir parçacık tarafından şimdiye kadar ulaşılan global en iyi çözümdür. *c1* ve *c2* öğrenme faktörleridir ve sırasıyla parçacığın kendi tecrübelerine göre ve sürüdeki diğer parçacıkların tecrübelerine göre hareketini yönlendirir. Bu çalışmada *c1* ve *c2* öğrenme faktörlerinin her ikiside *2* olarak alınmıştır (Reider ve ark., 1999). *r1* ve *r2*, *[0,1]* aralığındaki rastgele değerlerdir. *w* atalet ağırlığının 0.8, 1 ve 1.2 değerleri ile yaptığımız deneyler, bu parametrenin 1 değeri için elde ettiğimiz sonuçların (tahmin doğruluğu) diğerlerine göre daha iyi olduğunu göstermiştir. Bu nedenle, bu çalışmada mümkün olduğu kadar daha iyi bir çözüme ulaşmak için *w=1* olarak alınmıştır (Song ve Eberhart, 1998).

5.3.3. SVM tarafından SNP'lerin geri kalanının tahmin edilmesi

SNP'lerin geri kalanının tahmini işlemi Şekil 5.13'de gösterilmiştir. Şekilden de görülebildiği gibi SVM ilk olarak eğitim kümesi olarak verilen haplotiplerdeki SNP değerlerini kullanarak bir model inşa eder. Daha sonra test kümesindeki haplotipe ait olan SNP'lerin geri kalanlarının değerleri (bilinmeyen SNP değerleri) bu model ve CLONALG ile elde edilen etiket SNP'ler kullanılarak tahmin edilir.

SVM tabanlı tahmin yönteminde, *H* matrisinde bulunan her bir haplotip sırasıyla bir test kümesi olarak düşünülür. Burada her bir etiket SNP belli bir özelliği, SNP'lerin geri kalanının her biri de belli bir sınıfı temsil eder. Doğru olarak tahmin edilen SNP sayısının toplam tahmin edilen SNP sayısına oranı tahmin doğruluğu olarak adlandırılır ve Denklem 5.1'deki eşitlik ile hesaplanır.

Parametre optimizasyonlu CLONTagger yönteminde de SVM yazılımı olarak *RBF* (radial basis function) fonksiyonlu *Libsvm* yazılımı kullanılmıştır (Chang ve Lin, 2011). Libsvm, *C* ve γ parametreleri ile birlikte kullanılır. Bu metodda PSO algoritması bu iki parametrenin optimizasyonu için kullanılmıştır.

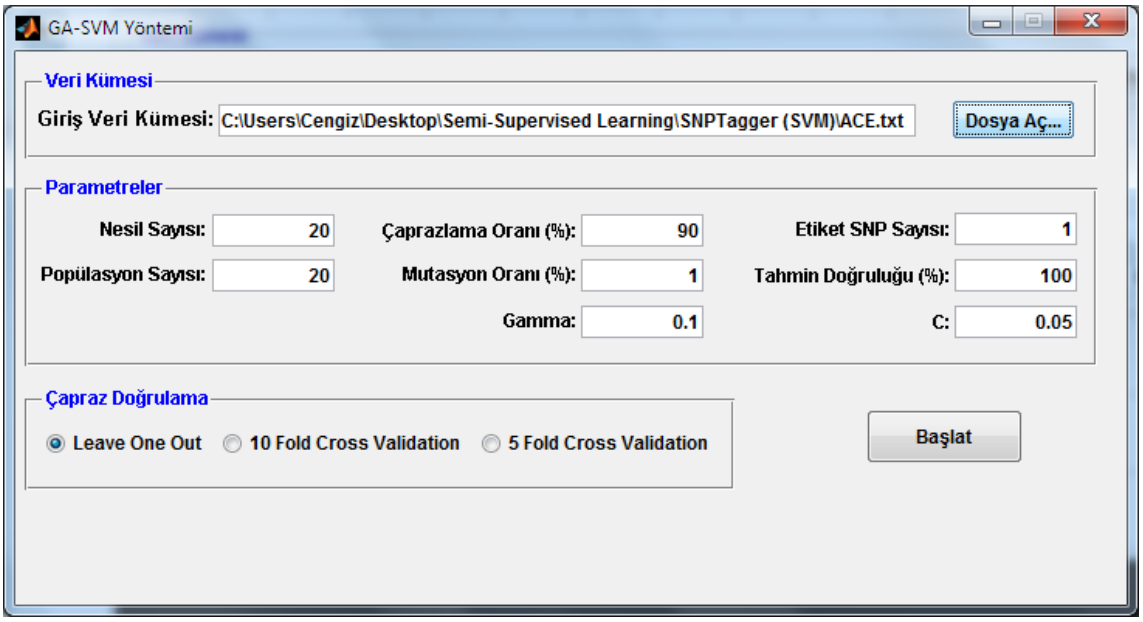
CLONALG tarafından oluşturulan popülasyon matrisindeki bir antikorun uygunluk değeri hesaplanırken PSO tarafından oluşturulan parçacık popülasyon matrisinde aynı satıra karşılık gelen parçacıktaki C ve γ parametreleri kullanılır. Örneğin, antikor 2'nin uygunluk değerinin hesaplanması işlemi parçacık popülasyon matrisindeki parçacık 2 kullanılarak yapılır. Hesaplanan bu değer hem CLONALG için antikor 2'nin uygunluk değeri hem de PSO için parçacık 2'nin uygunluk değeri olarak kabul edilir.

6. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

Bu tez çalışmasında geliştirilen GA-SVM, Parametre Optimizasyonlu GA-SVM ve Parametre Optimizasyonlu CLONTagger yöntemleri *Matlab 7.4* programı kullanılarak kodlanmıştır. Önerilen yöntemlerin performanslarını değerlendirmek için farklı özelliklerdeki veri kümeleri üzerinde çeşitli deneyler yapılmıştır. Deneylerde *Microsoft Windows 7* işletim sitemini kullanan *Intel Core I5-2430M@2.40 GHz* işlemciye ve *4 GB* belleğe sahip bilgisayar kullanılmıştır. Elde edilen sonuçlar diğer yöntemler ile karşılaştırılmıştır. Aşağıda, geliştirilen bu yöntemlerin farklı veri kümeleri üzerindeki uygulama sonuçlarına yer verilmiştir.

6.1. GA-SVM Yönteminin Uygulama Sonuçları

GA-SVM yönteminin Matlab programında geliştirilen grafiksel arayüzü Şekil 6.1'de verilmiştir. Şekilden de görülebildiği gibi veri kümesi seçimi ve parametre girişleri bu arayüz aracılığıyla yapılmaktadır.



Şekil 6.1. GA-SVM yönteminin Matlab'da oluşturulan grafiksel arayüzü

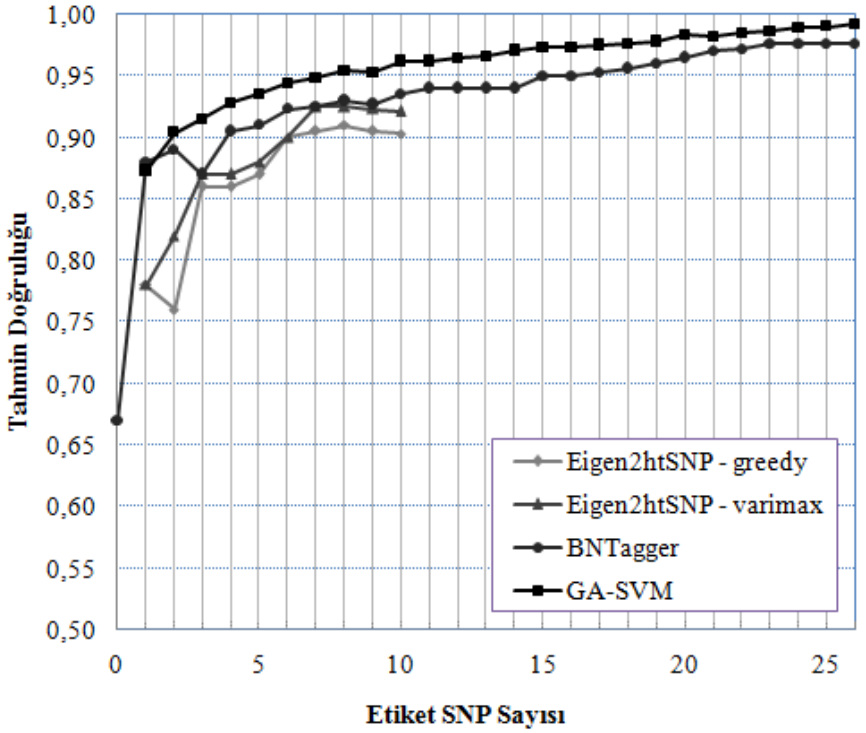
GA-SVM yönteminde, genetik algoritma için kullanılacak en uygun parametre değerlerini belirlemek için çeşitli deneyler yapılmıştır. Jenerasyon sayısının ve popülasyon büyüklüğünün 20'ye kadar olan değerleri için kullanılan veri kümeleri

üzerinde farklı etiket SNP sayılarında düzenli bir artış gözlenmiştir. Ancak jenerasyon sayısının ve popülasyon büyüklüğünün 20'den daha büyük değerleri için ise farklı etiket SNP sayılarında tahmin doğruluklarında önemli bir gelişme gözlenmemiştir. Bu nedenle genetik algoritma için jenerasyon sayısı ve popülasyon büyüklüğü *20* olarak kabul edilmiştir. Ayrıca en iyi çaprazlama ve mutasyon oranlarını tespit etmek için de çeşitli deneyler yapılmıştır. Yapılan bu deneyler, çaprazlama ve mutasyon oranlarındaki değişikliklerin GA-SVM algoritmasının tahmin doğruluğunu önemli oranda etkilemediğini göstermiştir. Bu nedenle Lin ve ark. (2003) tarafından da önerildiği gibi çaprazlama ve mutasyon oranları olarak sırasıyla *0.9* ve *0.01* değerleri kabul edilmiştir.

SVM içinde kullanılacak en uygun γ ve C parametre ikilisini belirlemek için de çeşitli deneyler yapılmıştır. Farklı veri kümeleri üzerinde yapılan bu deneyler içerisinde, $\gamma=0.1$ ve $C=0.05$ değerleri için GA-SVM algoritmasının tahmin doğruluğunun etiket SNP'lerin sayısının artmasıyla üssel olarak arttığı gözlenmiştir. Bu nedenle He ve Zelikovsky (2007) tarafından da önerildiği gibi γ ve C parametreleri için sırasıyla *0.1* ve *0.05* değerleri kullanılmıştır.

6.1.1. ACE veri kümesi

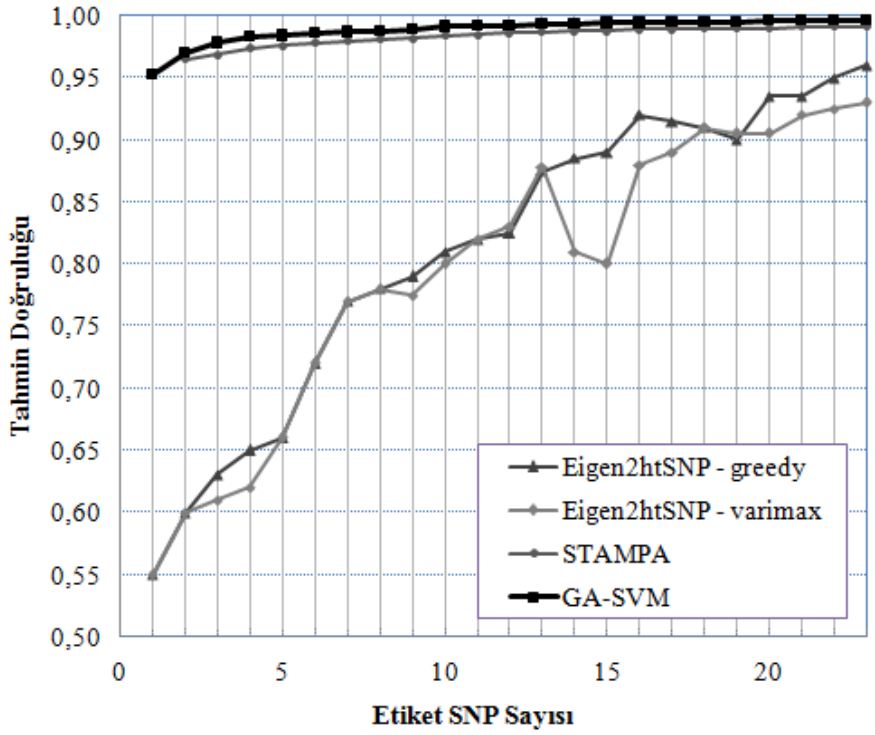
Önerilen GA-SVM metodu, 52 SNP'li ACE veri kümesi için BNTagger (Lee ve Shatkay, 2006) metodu ve Eigen2htSNP (Lin ve Altman, 2004) metodları ile karşılaştırılmıştır. Bu veri kümesi için elde edilen tahmin doğrulukları Şekil 6.2'de görülmektedir. Eigen2htSNP metotlarının önerildiği çalışmada, bu veri kümesi için sadece 1 ile 10 aralığındaki etiket SNP'ler için elde edilen sonuçlar verilmiştir. Bu nedenle Şekil 6.2'de de Eigen2htSNP metotları için bu aralıktaki sonuçlar gösterilmiştir. Şekilden de görüldüğü gibi GA-SVM metodu ile tek etiket SNP için %87.3 tahmin doğruluğu elde edilirken BNTagger metodu ile %88 tahmin doğruluğu elde edilir. Diğer bütün etiket SNP sayıları için ise GA-SVM metodu, BNTagger ve Eigen2htSNP metodlarından önemli oranda daha iyi bir performans sergilemiştir. Şekil 6.2'den de görülebildiği gibi GA-SVM metodunda etiket SNP sayısı artarken tahmin doğruluğu da düzenli olarak artmaktadır. Buna karşılık diğer metotların tahmin doğruluklarında dalgalanmalar görülmektedir. Önerilen GA-SVM metodu, 1 ile 10 etiket SNP aralığında BNTagger ve Eigen2htSNP metotlarından ortalama olarak %2.2 ve %5.1 daha yüksek bir tahmin doğruluğu sergilemiştir.



Şekil 6.2. ACE veri kümesi için GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.1.2. ABCB1 veri kümesi

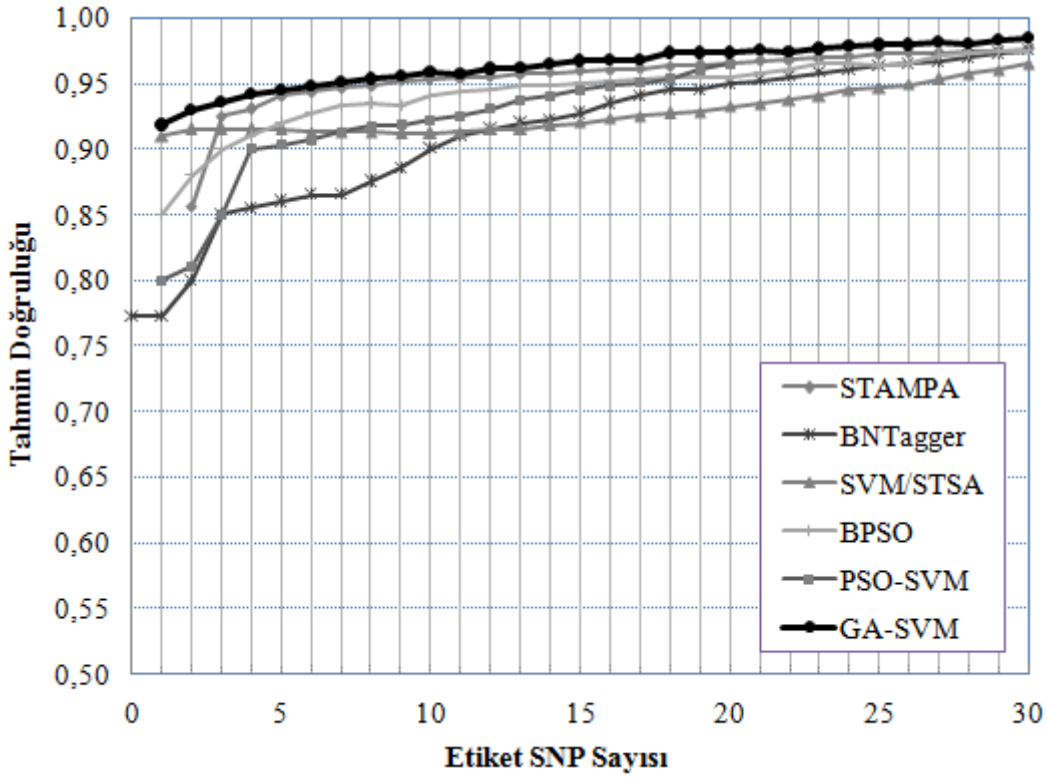
27 SNP'li ABCB1 veri kümesi için GA-SVM metodu Eigen2htSNP ve STAMPA (Halperin ve ark., 2005) metodları ile karşılaştırılmıştır. Bir etiket SNP için önerilen GA-SVM metodu %95.3 tahmin doğruluğuna ulaşırken Eigen2htSNP metotları %55 tahmin doğruluğuna ulaşmıştır. Benzer şekilde iki etiket SNP için GA-SVM metodu ile %97 tahmin doğruluğu elde edilirken STAMPA ile %96.5 tahmin doğruluğu elde edilmiştir. Oysa aynı etiket SNP sayısı için Eigen2htSNP metotları ile sadece %60 tahmin doğruluğuna ulaşılmıştır. Şekil 6.3'den de görülebileceği gibi 2 ile 23 etiket SNP aralığında GA-SVM metodu, Eigen2htSNP ve STAMPA metotlarından sırasıyla %16.6 ve %0.63 daha fazla tahmin doğruluğu sağlamıştır.



Şekil 6.3. ABCB1 veri kümesi için GA-SVM, STAMPA ve Eigen2htSNP metotları tarafından elde edilen tahmin doğrulukları

6.1.3. LPL veri kümesi

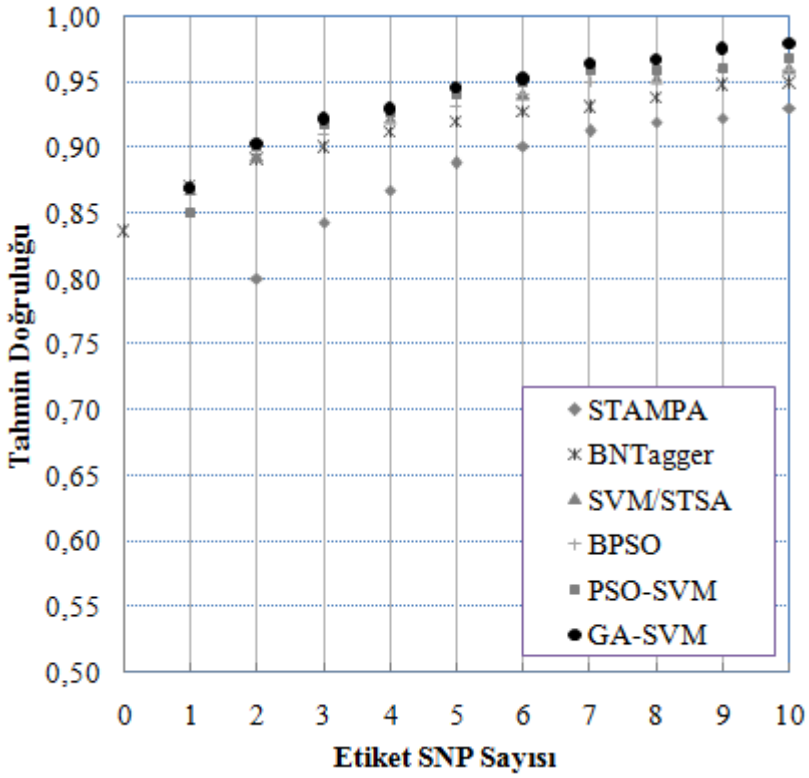
Önerilen GA-SVM yöntemi, 88 benzersiz haplotip içeren 88 SNP'li LPL veri kümesi üzerinde STAMPA, BNTagger, SVM/STSA (He ve Zelikovsky, 2007), BPSO (Yang ve ark., 2008) ve PSO-SVM (Lin ve Leu, 2010) metotları ile karşılaştırılmıştır. Şekil 6.4'den de görülebileceği gibi 2 ile 20 etiket SNP aralığında GA-SVM metodu, PSO-SVM, BPSO, SVM/STSA, BNTagger ve STAMPA metodlarından sırasıyla ortalamada %3.64, %2.12, %3.92, %5.89 ve %1.03 daha fazla oranda tahmin doğruluğu sergilemiştir.



Şekil 6.4. 88 benzersiz haplotip içeren LPL veri kümesi üzerinde GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.1.4. 5q31 veri kümesi

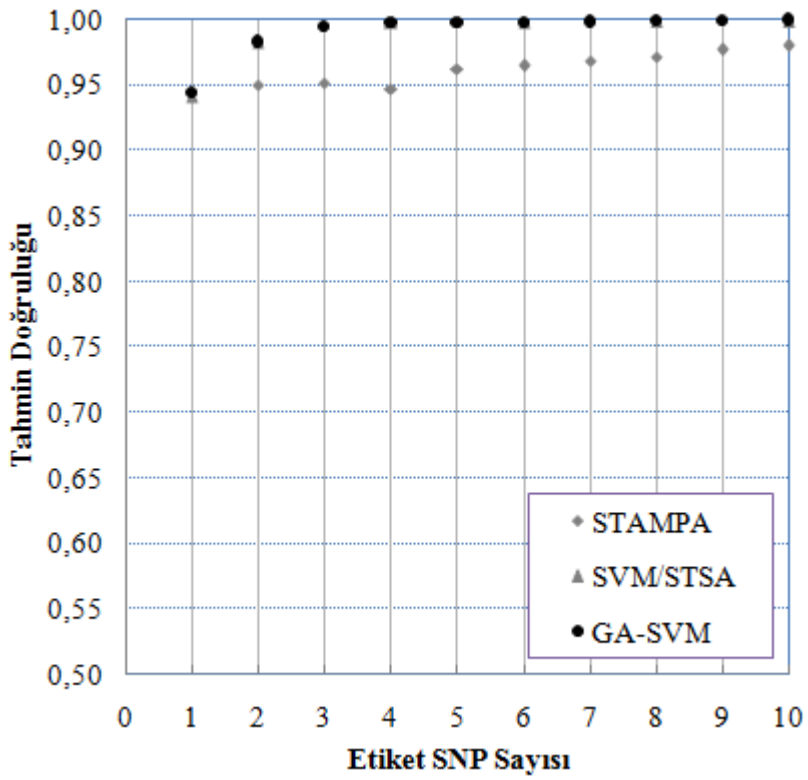
Önerilen GA-SVM yöntemi, 103 biallelik SNP içeren 5q31 veri kümesi üzerinde STAMPA, BNTagger, SVM/STSA, BPSO ve PSO-SVM metotları ile karşılaştırılmıştır. Bu veri kümesi üzerinde 2 ile 10 etiket SNP aralığında önerilen yöntem, diğerleri içerisinde en iyi olan yönteme göre ortalamada %0.6 oranında daha iyi bir tahmin doğruluğu sergilemiştir. Şekil 6.5, GA-SVM ve diğer yöntemler tarafından elde edilen tahmin doğruluklarını göstermektedir.



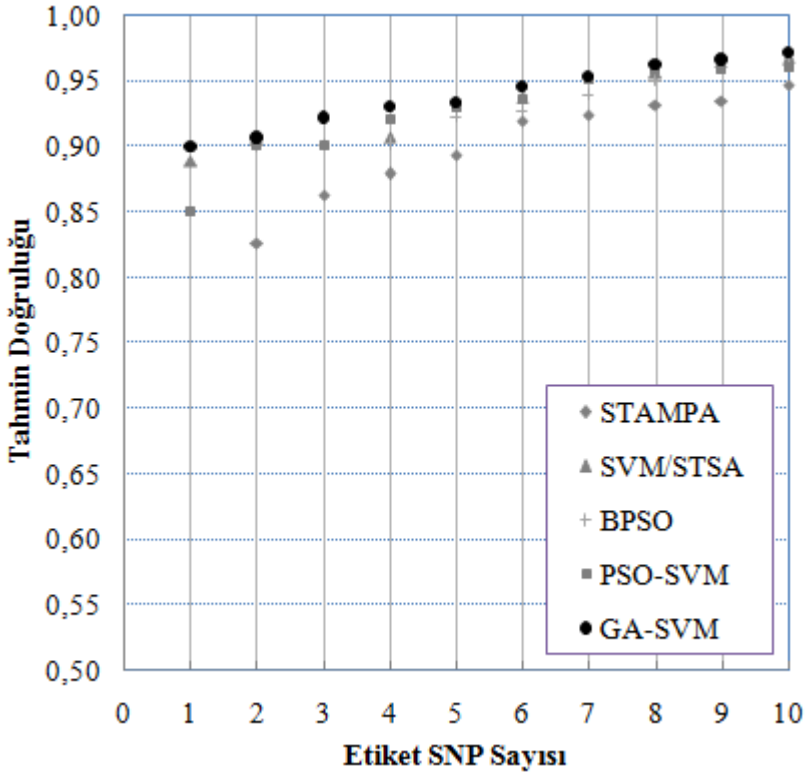
Şekil 6.5. 5q31 veri kümesi üzerinde GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.1.5. STEAP ve TRPM8 veri kümeleri

Önerilen GA-SVM yöntemi, 22 SNP'li STEAP veri kümesi üzerinde STAMPA ve SVM/STSA yöntemleri ile 101 SNP'li TRPM8 veri kümesi üzerinde de STAMPA, SVM/STSA, BPSO ve PSO-SVM yöntemleri ile karşılaştırılmıştır. STEAP veri kümesi üzerinde elde edilen sonuçlar Şekil 6.6, TRPM8 veri kümesi üzerinde elde edilen sonuçlar ise Şekil 6.7'de gösterilmektedir.



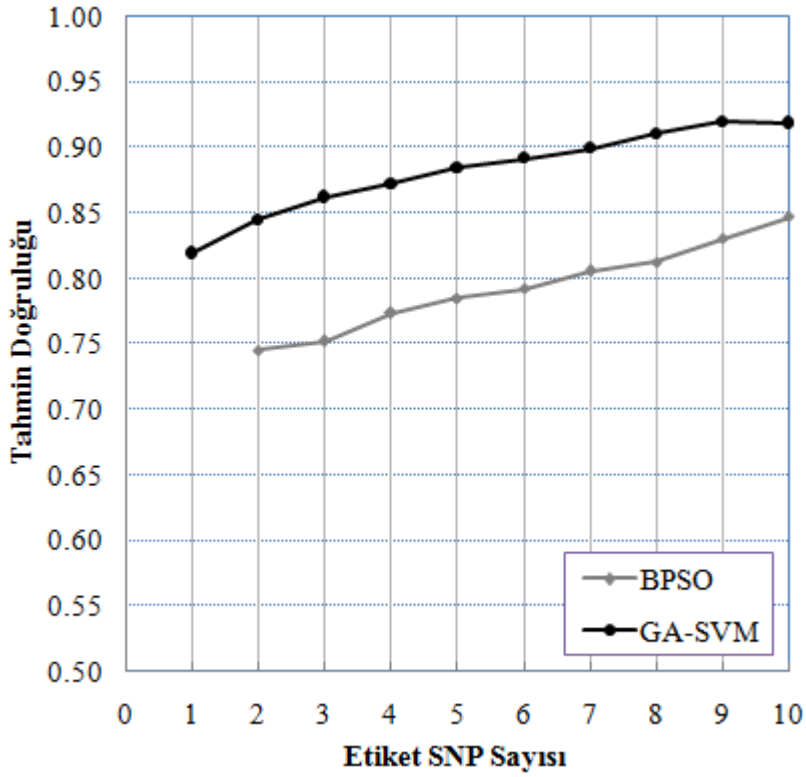
Şekil 6.6. STEAP veri kümesi üzerinde GA-SVM, STAMPA ve SVM/STSA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.7. TRPM8 veri kümesi üzerinde GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.1.6. Popülasyon D'nin D9 veri kümesi

GA-SVM yöntemi, 180 haplotip içeren 49 SNP'li popülasyon D'nin D9 veri kümesi üzerinde BPSO yöntemi ile karşılaştırılmıştır. 2 ile 10 etiket SNP aralığında önerilen yöntem, BPSO yöntemine göre %9.5 oranında daha iyi bir tahmin doğruluğu göstermiştir. Şekil 6.8, D9 veri kümesi üzerinde her iki yöntem tarafından elde edilen tahmin doğruluklarını göstermektedir.

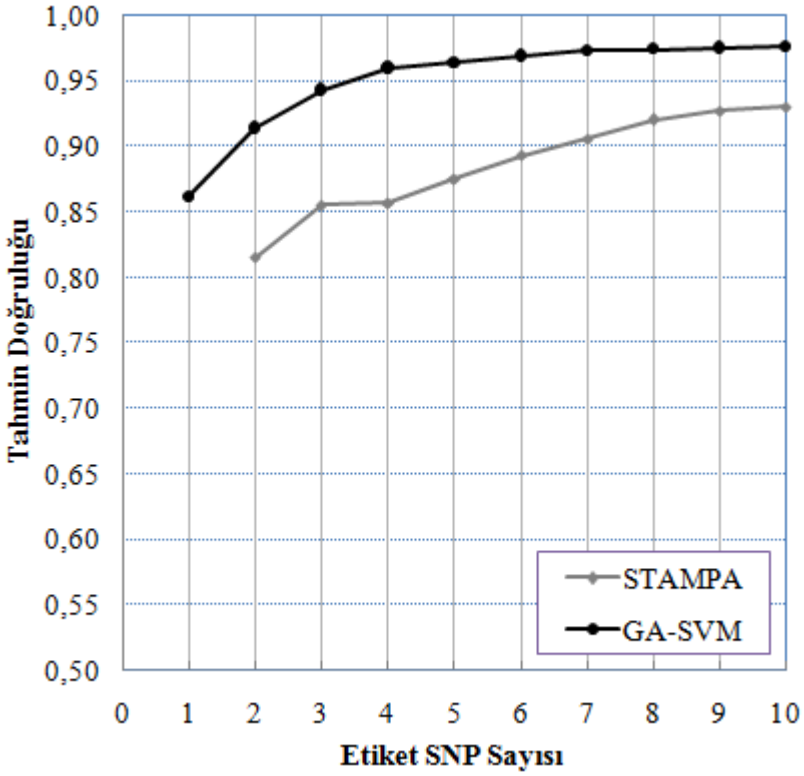


Şekil 6.8. D9 veri kümesi üzerinde GA-SVM ve BPSO metotları tarafından elde edilen tahmin doğrulukları

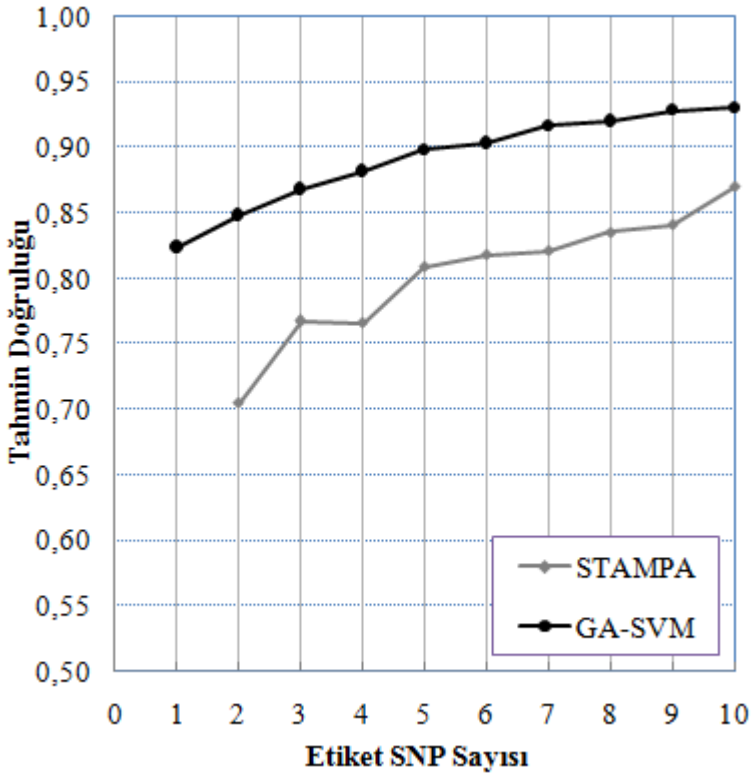
6.1.7. ENm013, ENr112 ve ENr113 veri kümeleri

ENm013, ENr112 ve ENr113 veri kümeleri sırasıyla 361, 412 ve 515 SNP içerir. Daha önce kullanılan veri kümeleri ile kıyaslandığında oldukça büyük veri kümeleridir. Önerilen yöntemin daha büyük veri kümeleri üzerindeki performansını değerlendirmek amacıyla bu veri kümeleri tercih edilmiştir. Bu veri kümeleri üzerinde elde edilen sonuçlar STAMPA metoduyla karşılaştırılmış ve aşağıdaki şekillerde verilmiştir. Şekil 6.9, Şekil 6.10 ve Şekil 6.11'den de görülebildiği gibi önerilen GA-SVM yöntemi,

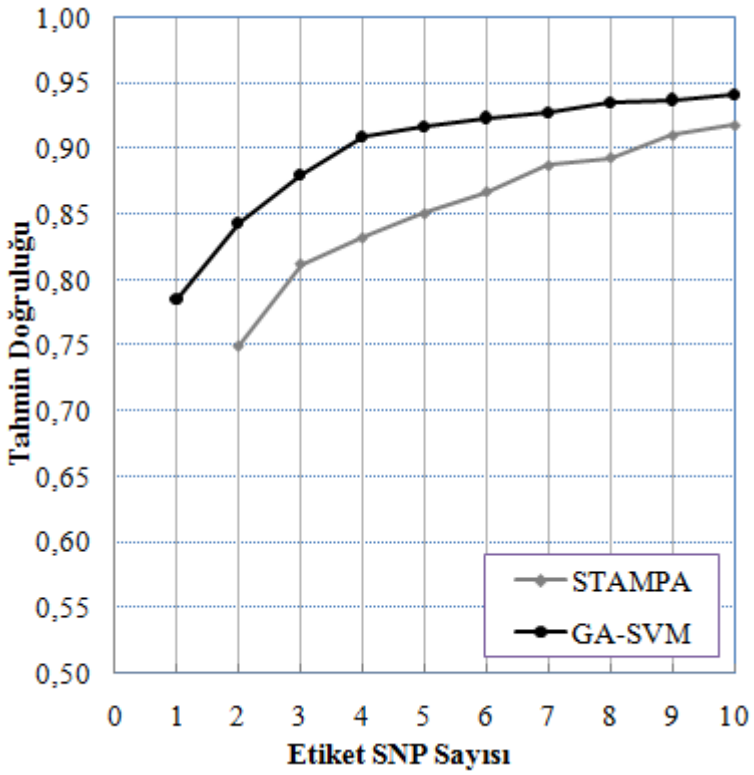
ENm013, ENr112 ve ENr113 veri kümeleri üzerinde yapılan deneylerde 2 ile 10 etiket SNP aralığında STAMPA metoduna göre ortalamada %7.5, %9.5 ve %5.4 daha yüksek tahmin doğruluğu sergilemiştir.



Şekil 6.9. ENm013 veri kümesi üzerinde GA-SVM ve STAMPA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.10. ENr112 veri kümesi üzerinde GA-SVM ve STAMPA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.11. ENr113 veri kümesi üzerinde GA-SVM ve STAMPA metotları tarafından elde edilen tahmin doğrulukları

6.1.8. GA-SVM yönteminin çalışma süresi

Önerilen yöntemin farklı sayıdaki etiket SNP'ler için elde edilen çalışma süreleri Çizelge 6.1'de verilmiştir. SVM/STSA yönteminin 5q31, TRPM8 ve STEAP veri kümeleri üzerindeki çalışma süreleri ise orjinal yayınlarından alınmıştır. STAMPA metodunun 2 ile 10 etiket SNP aralığında tüm veri kümeleri üzerindeki çalışma süreleri deneyler yapılarak belirlenmiş ancak bu aralıkta süre değişimi çok az olduğu için tabloda belirtilmemiştir. Örneğin STAMPA metodu, kullandığımız en küçük veri kümesi olan ACE için 2'den 10'a kadar olan tüm etiket SNP'leri yaklaşık 2 saniyede belirlemektedir. Buna karşılık önerilen metot, 2 etiket SNP'i 2 dakika 45 saniyede belirlerken 10 etiket SNP'i ise 6 dakika 52 saniyede tespit etmektedir. 5q31 veri kümesi için 2'den 10'a kadar olan tüm etiket SNP'leri STAMPA metodu 7 saniyede seçmektedir. Buna karşılık SVM/STSA metodu, 2 etiket SNP'i 5 saatte 10 etiket SNP'i ise 24 saatte seçmektedir. Önerilen metot ise 2 etiket SNP'i 3 saatte 10 etiket SNP'i ise 8 saat 55 dakikada belirlemektedir. En büyük veri kümesi olan ENr113 üzerinde yapılan deneyler, STAMPA metodunun 2'den 10'a kadar olan tüm etiket SNP'leri yaklaşık 6 dakikada seçtiğini göstermiştir. Önerilen metot ise 2 etiket SNP'i 6 saat 40 dakikada 10 etiket SNP'i ise 28 saat 3 dakikada belirlemiştir. Yapılan deneylerden de anlaşıldığı gibi GA-SVM metodu, SVM/STSA metoduna göre daha hızlı çalışırken STAMPA medoduna göre daha yavaş çalışmaktadır.

Çizelge 6.1. Kullanılan veri kümeleri üzerinde farklı etiket SNP sayıları için GA-SVM ve SVM/STSA yöntemlerinin çalışma süreleri

Veri kümeleri (Hap, SNP)	Yöntemler	Etiket SNP Sayısı									
		1	2	3	4	5	6	7	8	9	10
ACE (22, 52)	GA-SVM	1 d 56 sn	2 d 45 sn	3 d 28 sn	3 d 38 sn	5 d 26 sn	5 d 23 sn	5 d 45 sn	5 d 23 sn	6 d 44 sn	6 d 52 sn
ABCB1 (494, 27)	GA-SVM	1 s 19 d	1 s 8 d	1 s 3 d	51 d 58 sn	57 d 14 sn	52 d 4 sn	50 d 59 sn	44 d 32 sn	41 d 49 sn	37 d 35 sn
LPL (88, 88)	GA-SVM	7 d 38 sn	12 d 50 sn	13 d 56 sn	21 d 40 sn	22 d 54 sn	23 d 54 sn	29 d 45 sn	36 d 6 sn	35 d 48 sn	40 d 45 sn
D9 (180, 49)	GA-SVM	17 d 29 sn	21 d 2 sn	25 d 44 sn	29 d 56 sn	35 d 56 sn	43 d 47 sn	54 d 45 sn	57 d 4 sn	57 d 59 sn	1 s 3 d
5q31 (258, 103)	SVM/STSA	3 s	5 s	-	11 s	-	16 s	-	18 s	-	24 s
	GA-SVM	2 s 12 d	3 s	3 s 43 d	4 s 5 d	4 s 49 d	5 s 36 d	6 s 59 d	6 s 30 d	8 s 13 d	8 s 55 d
TRPM8 (120, 101)	SVM/STSA	1 s	2 s	-	5 s	-	9 s	-	16 s	-	23 s
	GA-SVM	20 d 51 sn	24 d 3 sn	24 d 17 sn	26 d 21 sn	33 d 6 sn	41 d 36 sn	42 d 37 sn	47 d 17 sn	49 d 20 sn	53 d 7 sn
STEAP (120, 22)	SVM/STSA	14 d	27 d	-	1 s	-	2 s	-	3 s	-	4 s
	GA-SVM	2 d 41 sn	2 d 37 sn	2 d 18 sn	2 d 40 sn	2 d 45 sn	3 d 1 sn	2 d 40 sn	2 d 24 sn	2 d 27 sn	2 d 20 sn
ENm013 (120, 361)	GA-SVM	2 s 10 d	2 s 55 d	3 s 51 d	4 s 43 d	5 s 41 d	7 s 13 d	8 s 45 d	9 s 10 d	9 s 43 d	11 s 23 d
ENr112 (120, 412)	GA-SVM	3 s 11 d	3 s 45 d	5 s 18 d	6 s 23 d	8 s 36 d	10 s	11 s 53 d	15 s 16 d	21 s 6 d	23 s 36 d
ENr113 (120, 515)	GA-SVM	5 s 8 d	6 s 40 d	8 s 26 d	10 s 33 d	13 s 53 d	16 s 41 d	19 s 26 d	21 s 40 d	24 s 43 d	28 s 3 d

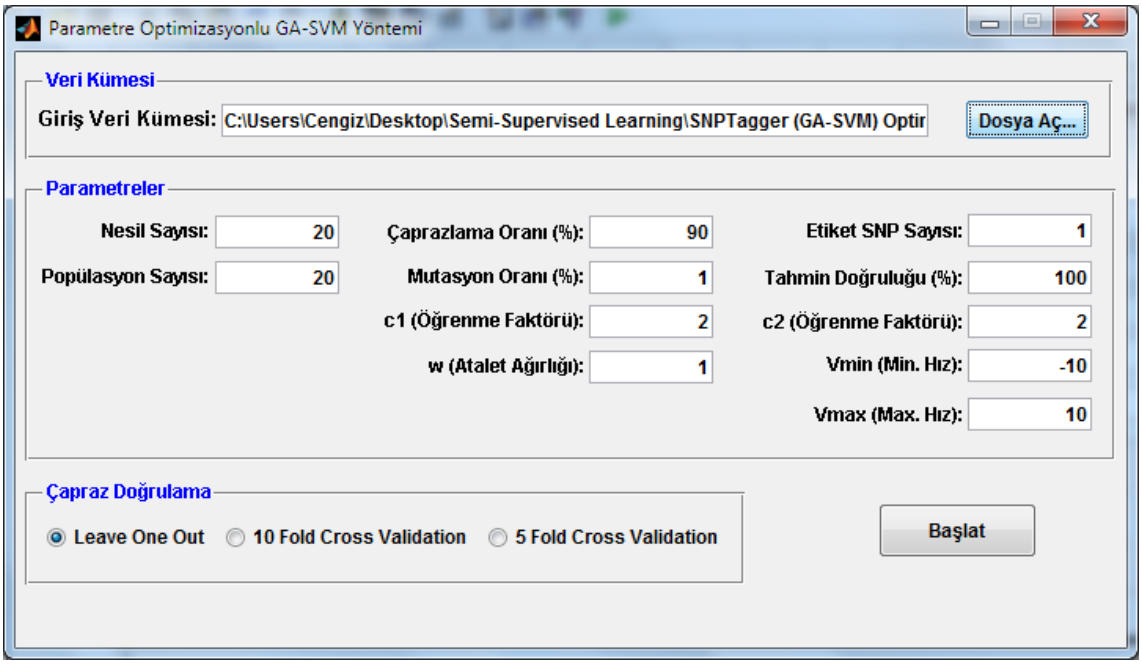
Hap: Haplotip sayısı, SNP: SNP sayısı, s: saat, d: dakika, sn: saniye

6.2. Parametre Optimizasyonlu GA-SVM Yönteminin Uygulama Sonuçları

Parametre optimizasyonlu GA-SVM yönteminin Matlab programında geliştirilen grafiksel arayüzü Şekil 6.12'de verilmiştir. Şekilden de görülebildiği gibi veri kümesi seçimi ve parametre girişleri bu arayüz aracılığıyla yapılmaktadır.

Parametre optimizasyonlu GA-SVM yönteminde, genetik algoritma için jenerasyon sayısı ve popülasyon büyüklüğü *20*, çaprazlama ve mutasyon oranları da sırasıyla *0.9* ve *0.01* olarak kabul edilmiştir. Ayrıca PSO için jenerasyon sayısı ve popülasyon büyüklüğü *20*, her iki öğrenme faktörü *2* ve atelet ağırlığı da *1* olarak kabul edilmiştir. *C* ve γ parametrelerinin her ikisi için de $[10^{-2}, 10^{2}]$ arama aralığı kullanılmış, V_{min} değeri *-10* ve V_{max} değeri de *10* olarak kabul edilmiştir. Bu parametreler çeşitli deneyler yapılarak deneme yanılma yoluyla belirlenmiştir. Böylece parametre

optimizasyonlu GA-SVM yönteminin bütün veri kümeleri üzerindeki uygulama sonuçları bu parametreler kullanılarak elde edilmiştir.



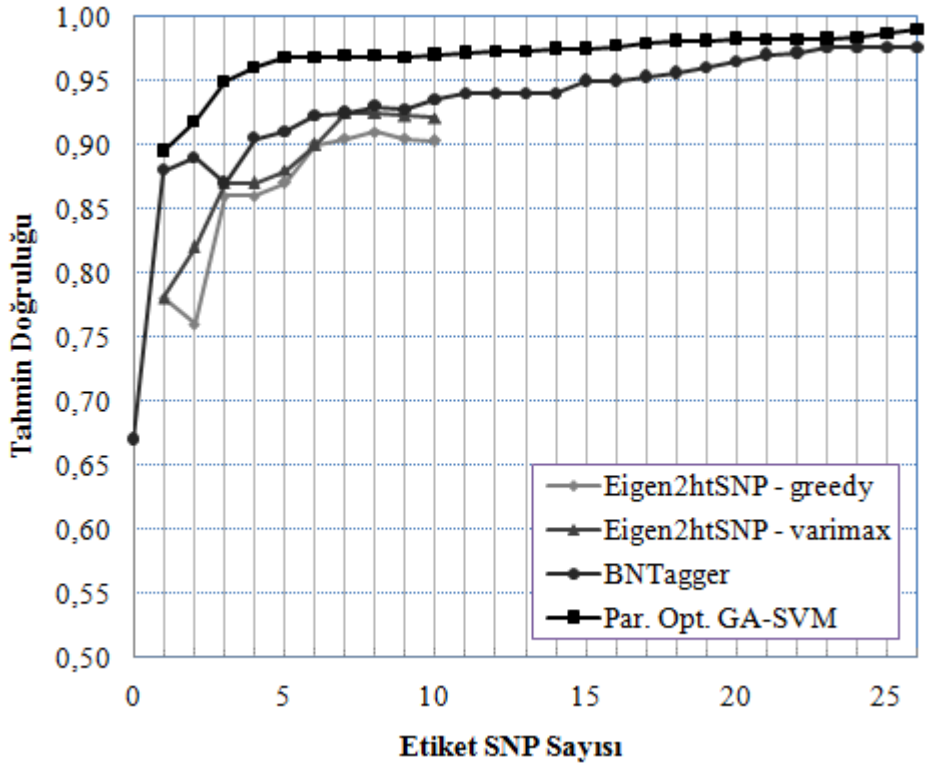
Şekil 6.12. Parametre optimizasyonlu GA-SVM yönteminin Matlab'da oluşturulan grafiksel arayüzü

Grid search ile karşılaştırıldığında PSO kullanmanın faydalarını göstermek için C ve γ parametrelerinin farklı aralık değerleri için çeşitli deneyler yapılmıştır. Bu deneylerin ilkinde, bu parametrelerin [0.1, 5] aralık değerleri için grid search (grid nokta sayısı 50^2=2500'dür) ve PSO ile aynı tahmin doğrulukları elde edilirken PSO'nun diğerine göre 250 kat daha hızlı çalıştığı görülmüştür. Yapılan ikinci deneyde ise bu parametrelerin [0.1, 10] aralık değerleri için grid search (grid nokta sayısı 100^2=10000'dir) ve PSO ile yine aynı tahmin doğrulukları elde edilirken PSO'nun diğerine göre 1000 kat daha hızlı çalıştığı belirlenmiştir. Yapılan bu deneyler C ve γ parametrelerinin optimizasyonu için kullanılan PSO'nun, grid serach algoritmasına göre çok daha hızlı çalıştığını göstermiştir. Ayrıca her iki algoritma ile aynı tahmin doğrulukları elde edilmiştir.

6.2.1. ACE veri kümesi

Önerilen parametre optimizasyonlu GA-SVM metodu, 52 SNP'li ACE veri kümesi için BNTagger metodu ve Eigen2htSNP metodları ile karşılaştırılmıştır. Şekil

6.13'den de görülebildiği gibi parametre optimizasyonlu GA-SVM metodu, bütün etiket SNP sayıları için BNTagger ve Eigen2htSNP metodlarından önemli oranda daha iyi bir performans sergilemiştir. Parametre optimizasyonlu GA-SVM metodunda etiket SNP sayısı artarken tahmin doğruluğu da düzenli olarak artmaktadır. Buna karşılık, diğer metotların tahmin doğruluklarında dalgalanmalar görülmektedir. Önerilen yöntem 1 ile 10 etiket SNP aralığında BNTagger ve Eigen2htSNP metotlarından ortalama olarak %4.39 ve %7.2 daha yüksek bir tahmin doğruluğu sergilemiştir.

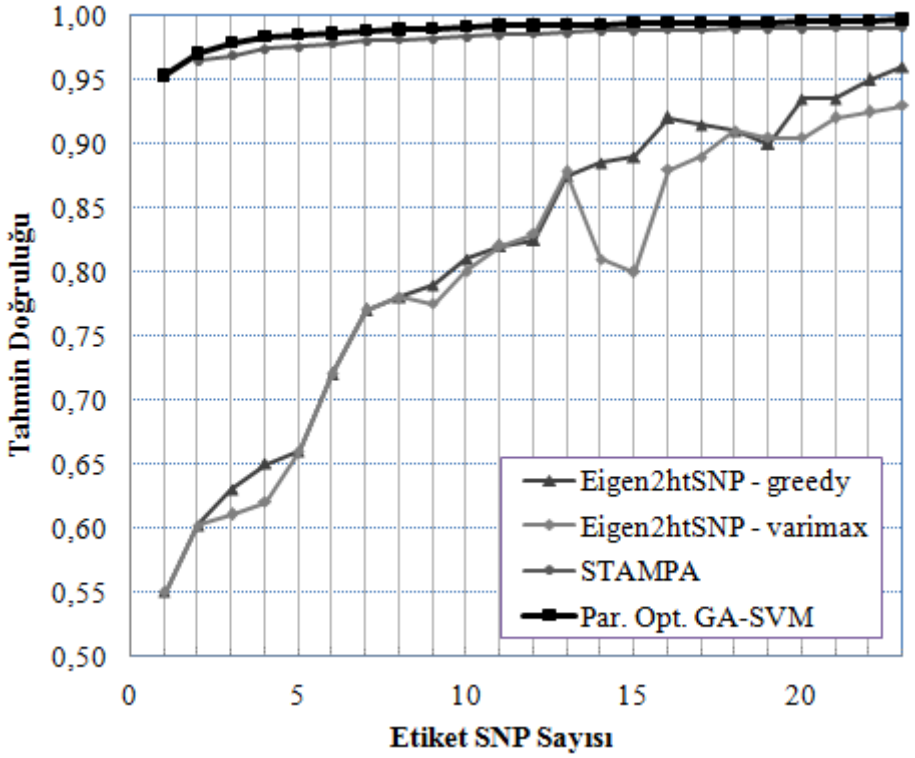


Şekil 6.13. ACE veri kümesi için parametre optimizasyonlu GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.2.2. ABCB1 veri kümesi

27 SNP'li ABCB1 veri kümesi için parametre optimizasyonlu GA-SVM metodu Eigen2htSNP ve STAMPA metodları ile karşılaştırılmıştır. Tek etiket SNP için önerilen parametre optimizasyonlu GA-SVM metodu, %95.3 tahmin doğruluğu sağlarken Eigen2htSNP metotları ise %55 tahmin doğruluğu sağlamıştır. Benzer şekilde iki etiket SNP için önerilen yöntem %97 tahmin doğruluğuna ulaşırken STAMPA %96.5 tahmin doğruluğuna ulaşmıştır. Aynı etiket SNP sayısı için Eigen2htSNP metotları ise sadece %60 tahmin doğruluğuna ulaşabilmiştir. Şekil 6.14'den de görülebileceği gibi 2 ile 23

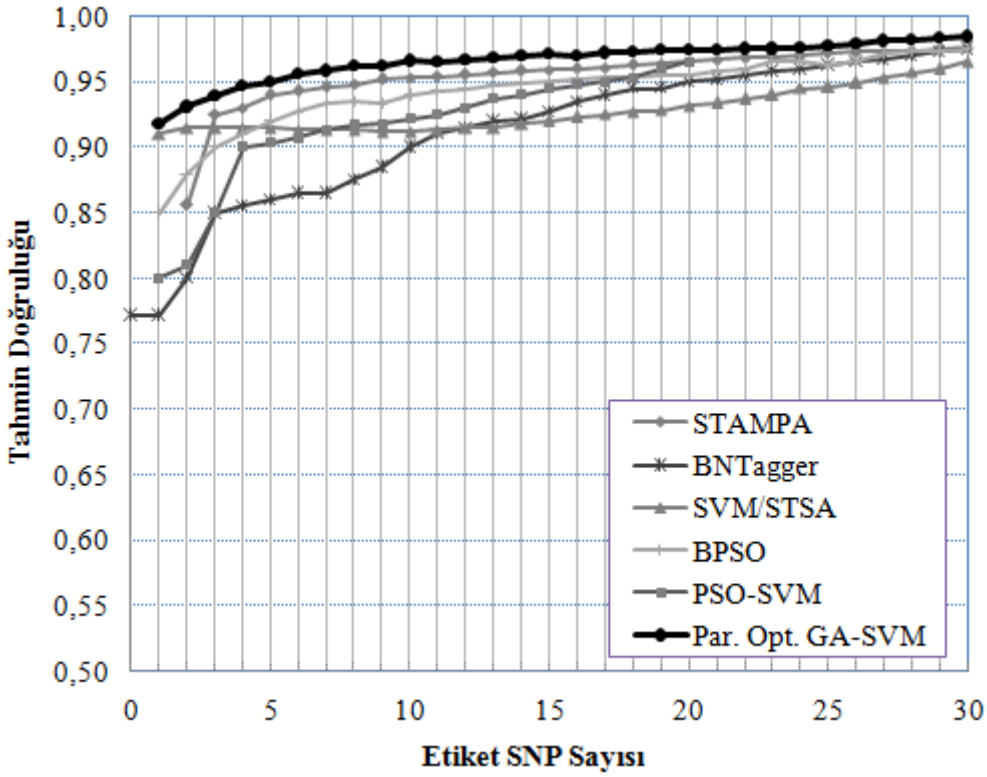
etiket SNP aralığında parametre optimizasyonlu GA-SVM metodu, Eigen2htSNP ve STAMPA metotlarından sırasıyla %16.6 ve %0.66 daha fazla tahmin doğruluğu elde etmiştir.



Şekil 6.14. ABCB1 veri kümesi için parametre optimizasyonlu GA-SVM, STAMPA ve Eigen2htSNP metotları tarafından elde edilen tahmin doğrulukları

6.2.3. LPL veri kümesi

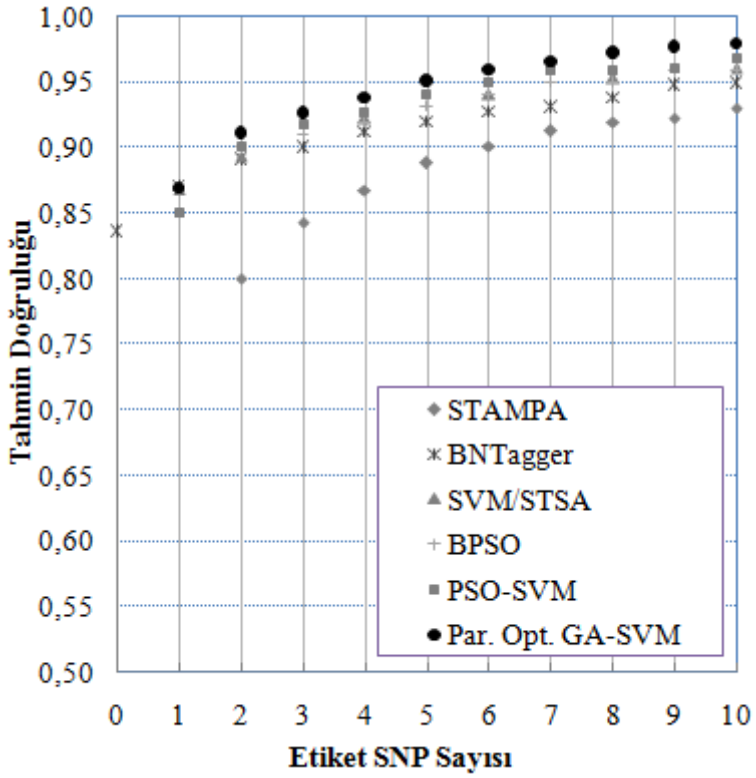
Önerilen parametre optimizasyonlu GA-SVM yöntemi, 88 benzersiz haplotip içeren 88 SNP'li LPL veri kümesi üzerinde STAMPA, BNTagger, SVM/STSA, BPSO ve PSO-SVM metotları ile karşılaştırılmıştır. Şekil 6.15'den de görülebileceği gibi 2 ile 20 etiket SNP aralığında parametre optimizasyonlu GA-SVM metodu PSO-SVM, BPSO, SVM/STSA, BNTagger ve STAMPA metodlarından sırasıyla ortalamada %4.11, %2.59, %4.40, %6.37 ve %1.50 oranında daha fazla tahmin doğruluğu sergilemiştir.



Şekil 6.15. 88 benzersiz haplotip içeren LPL veri kümesi üzerinde parametre optimizasyonlu GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.2.4. 5q31 veri kümesi

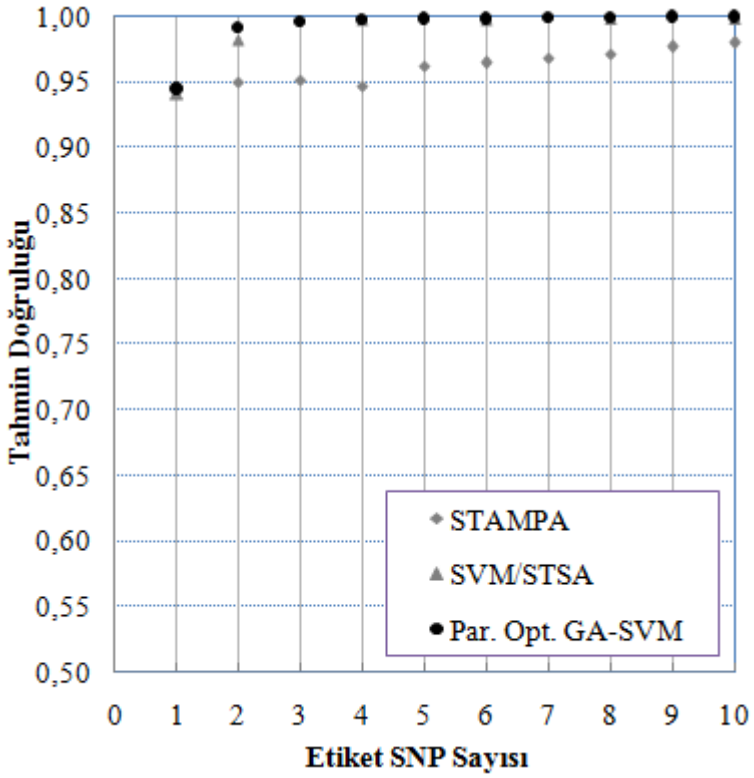
Önerilen parametre optimizasyonlu GA-SVM yöntemi, 103 biallelik SNP içeren 5q31 veri kümesi üzerinde STAMPA, BNTagger, SVM/STSA, BPSO ve PSO-SVM metotları ile karşılaştırılmıştır. Bu veri kümesi üzerinde 2 ile 10 etiket SNP aralığında önerilen yöntem, PSO-SVM, BPSO, SVM/STSA, BNTagger ve STAMPA yöntemlerine göre ortalamada %1.1, %1.8, %1.9, %2.9 ve %6.6 oranında daha iyi tahmin doğruluğu göstermiştir. Şekil 6.16, parametre optimizasyonlu GA-SVM ve diğer yöntemler tarafından elde edilen tahmin doğruluklarını göstermektedir.



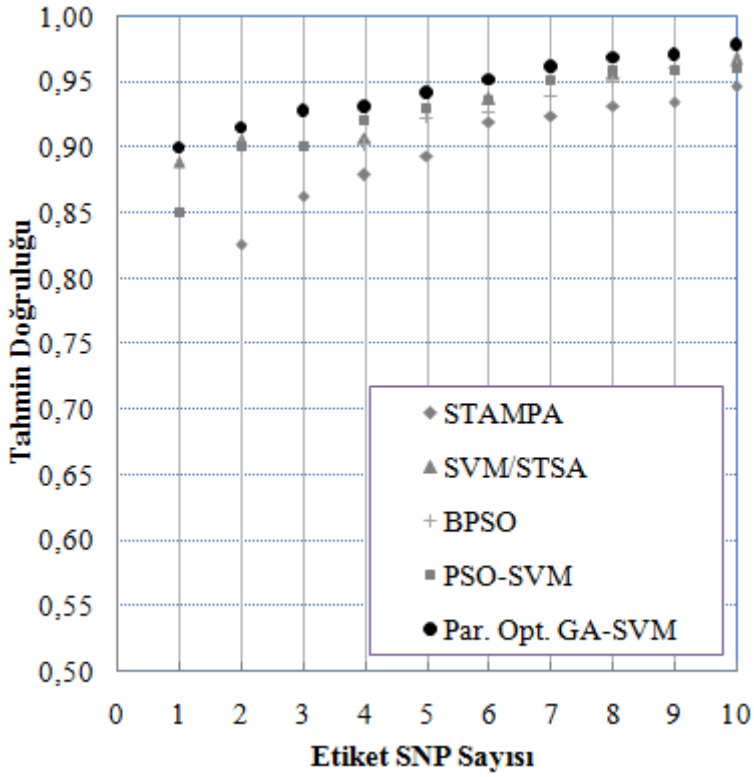
Şekil 6.16. 5q31 veri kümesi üzerinde parametre optimizasyonlu GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.2.5. STEAP ve TRPM8 veri kümeleri

Önerilen parametre optimizasyonlu GA-SVM yöntemi, 22 SNP'li STEAP veri kümesi üzerinde STAMPA ve SVM/STSA yöntemleri ile 101 SNP'li TRPM8 veri kümesi üzerinde de STAMPA, SVM/STSA, BPSO ve PSO-SVM yöntemleri ile karşılaştırılmıştır. STEAP veri kümesi üzerinde elde edilen sonuçlar Şekil 6.17, TRPM8 veri kümesi üzerinde elde edilen sonuçlar ise Şekil 6.18'de gösterilmektedir.



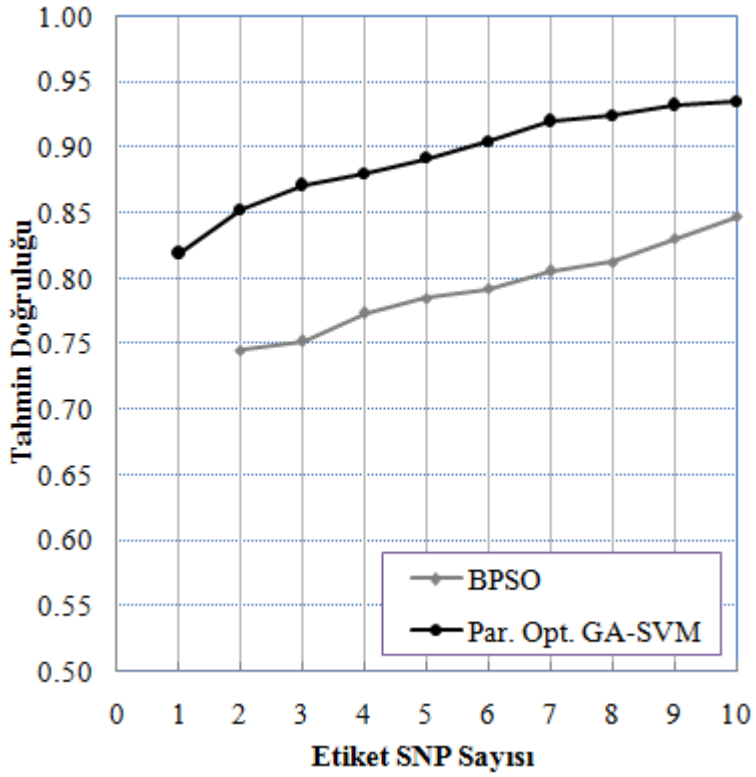
Şekil 6.17. STEAP veri kümesi üzerinde parametre optimizasyonlu GA-SVM, STAMPA ve SVM/STSA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.18. TRPM8 veri kümesi üzerinde parmetre optimizasyonlu GA-SVM ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.2.6. Popülasyon D'nin D9 veri kümesi

Önerilen yöntem, 180 haplotip içeren 49 SNP'li popülasyon D'nin D9 veri kümesi üzerinde BPSO yöntemi ile karşılaştırılmıştır. 2 ile 10 etiket SNP aralığında önerilen yöntem, BPSO yöntemine göre %10.7 oranında daha iyi bir tahmin doğruluğu göstermiştir. Şekil 6.19, D9 veri kümesi üzerinde her iki yöntem tarafından elde edilen tahmin doğruluklarını göstermektedir.

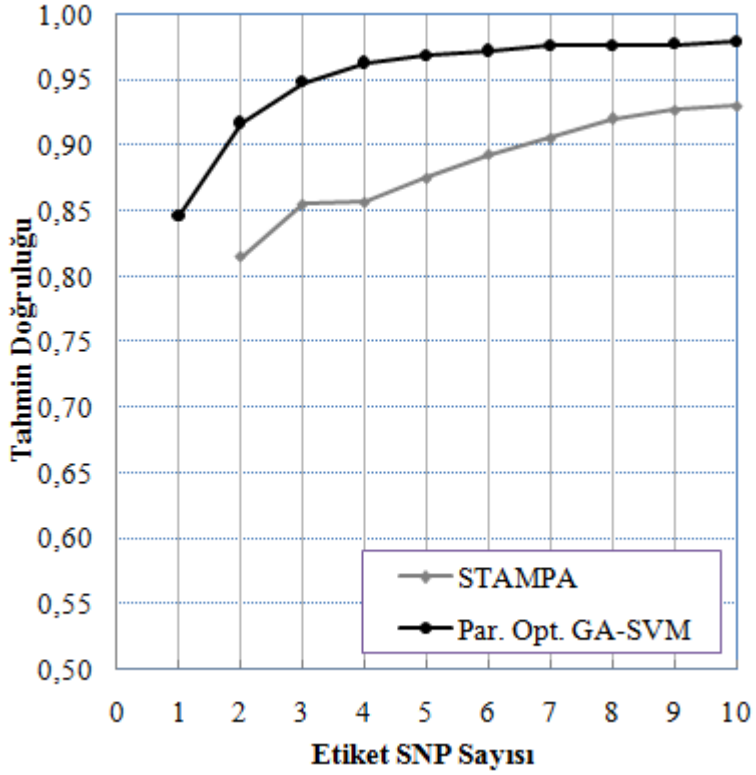


Şekil 6.19. D9 veri kümesi üzerinde parametre optimizasyonlu GA-SVM ve BPSO metotları tarafından elde edilen tahmin doğrulukları

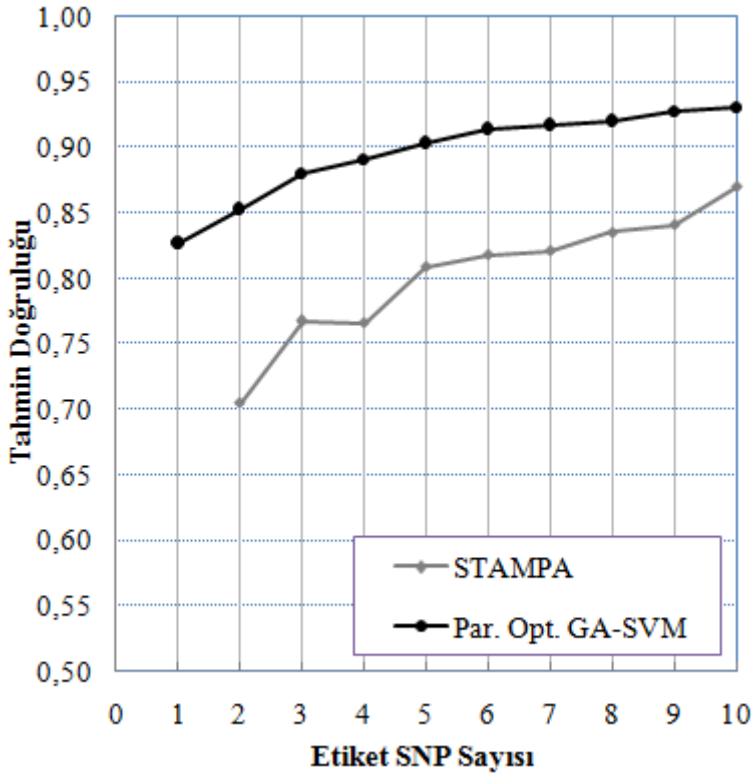
6.2.7. ENm013, ENr112 ve ENr113 veri kümeleri

Parametre optimizasyonlu GA-SVM yönteminin daha büyük veri kümeleri üzerindeki performansını değerlendirmek için ENm013, ENr112 ve ENr113 veri kümeleri kullanılmıştır. Bu veri kümeleri üzerinde elde edilen sonuçlar STAMPA metoduyla karşılaştırılmış ve aşağıdaki şekillerde verilmiştir. Şekil 6.20, Şekil 6.21 ve Şekil 6.22'den de görülebildiği gibi önerilen parametre optimizasyonlu GA-SVM yöntemi ENm013, ENr112 ve ENr113 veri kümeleri üzerinde yapılan deneylerde 2 ile

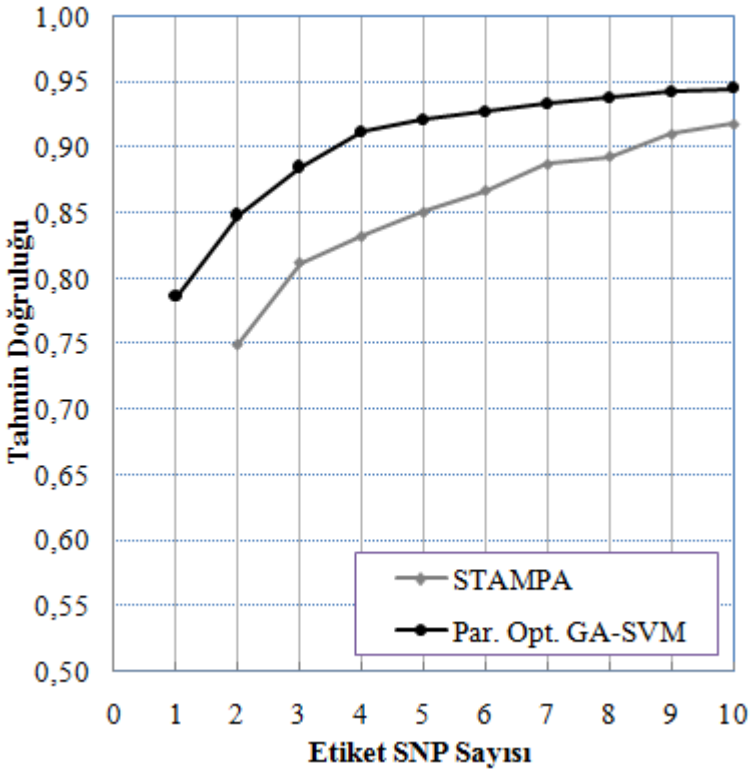
10 etiket SNP aralığında STAMPA metoduna göre ortalamada %7.8, %10 ve %5.9 daha yüksek tahmin doğruluğu sergilemiştir.



Şekil 6.20. ENm013 veri kümesi üzerinde parametre optimizasyonlu GA-SVM ve STAMPA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.21. ENr112 veri kümesi üzerinde parametre optimizasyonlu GA-SVM ve STAMPA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.22. ENr113 veri kümesi üzerinde parametre optimizasyonlu GA-SVM ve STAMPA metotları tarafından elde edilen tahmin doğrulukları

6.2.8. Parametre optimizasyonlu GA-SVM yönteminin çalışma süresi

Önerilen yöntemin farklı veri kümeleri üzerinde farklı sayıdaki etiket SNP'ler için elde edilen çalışma süreleri Çizelge 6.2'de verilmiştir. SVM/STSA yönteminin 5q31, TRPM8 ve STEAP veri kümeleri üzerindeki çalışma süreleri ise orjinal yayınlarından alınmıştır. STAMPA metodunun 2 ile 10 etiket SNP aralığında tüm veri kümeleri üzerindeki çalışma süreleri deneyler yapılarak belirlenmiş ancak bu aralıkta süre değişimi çok az olduğu için tabloda belirtilmemiştir. Örneğin STAMPA metodu, kullandığımız en küçük veri kümesi olan ACE için 2'den 10'a kadar olan tüm etiket SNP'leri yaklaşık 2 saniyede belirlemektedir. Buna karşılık önerilen metot, 2 etiket SNP'i 2 dakika 16 saniyede belirlerken 10 etiket SNP'i ise 6 dakika 41 saniyede tespit etmektedir. 5q31 veri kümesi için ise 2'den 10'a kadar olan tüm etiket SNP'leri STAMPA metodu 7 saniyede seçmektedir. Buna karşılık SVM/STSA metodu, 2 etiket SNP'i 5 saatte 10 etiket SNP'i ise 24 saatte seçmektedir. Önerilen metot ise 2 etiket SNP'i 2 saat 54 dakikada 10 etiket SNP'i ise 8 saat 30 dakikada belirlemiştir. En büyük veri kümesi olan ENr113 üzerinde yapılan deneyler, STAMPA metodunun 2'den 10'a kadar olan tüm etiket SNP'leri yaklaşık 6 dakikada seçtiğini göstermiştir. Önerilen metot ise 2 etiket SNP'i 7 saat 12 dakikada 10 etiket SNP'i ise 29 saat 19 dakikada belirlemiştir. Yapılan deneylerden de anlaşıldığı gibi parametre optimizasyonlu GA-SVM metodu, SVM/STSA metoduna göre daha hızlı çalışırken STAMPA medoduna göre daha yavaş çalışmaktadır.

Çizelge 6.2. Kullanılan veri kümeleri üzerinde farklı etiket SNP sayıları için parametre optimizasyonlu GA-SVM ve SVM/STSA yöntemlerinin çalışma süreleri

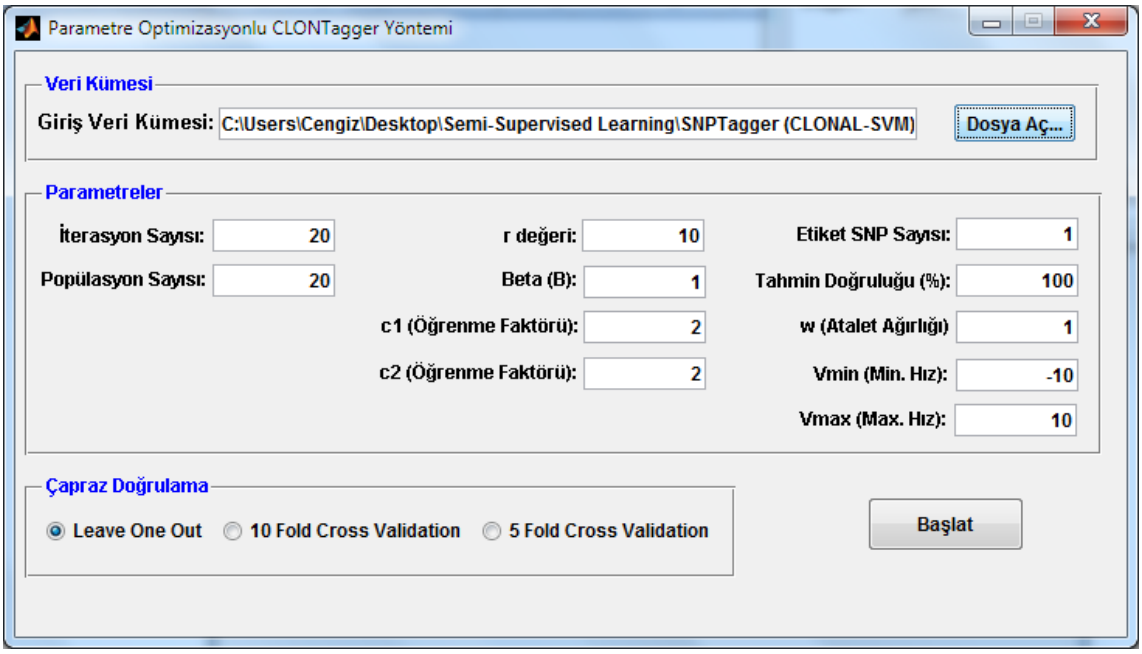
Veri kümeleri (Hap, SNP)	Yöntemler	Etiket SNP Sayısı									
		1	2	3	4	5	6	7	8	9	10
ACE (22, 52)	Par. Opt.	1 d	2 d	3 d	4 d	4 d	4 d	5 d	6 d	6 d	6 d
	GA-SVM	39 sn	16 sn	25 sn	17 sn	58 sn	44 sn	52 sn		25 sn	41 sn
ABCB1 (494, 27)	Par. Opt.	1 s	1 s 7	58 d	57 d	51 d	53 d	47 d	43 d	40 d	37 d
	GA-SVM	15 d	d	43 sn	23 sn	23 sn	11 sn	54 sn	41 sn	10 sn	30 sn
LPL (88, 88)	Par. Opt.	7 d	11 d	15 d	19 d	22 d	26 d	30 d	35 d	38 d	44 d
	GA-SVM	35 sn	13 sn	41 sn	23 sn	4 sn	19 sn	58 sn	58 sn	55 sn	24 sn
D9 (180, 49)	Par. Opt.	14 d	19 d	23 d	31 d	34 d	42 d	52 d	53 d	1 s 4	1 s
	GA-SVM	8 sn	5 sn	3 sn	23 sn	3 sn	40 sn	57 sn	42 d	d	17 d
5q31 (258, 103)	SVM/STSA	3 s	5 s	-	11 s	-	16 s	-	18 s	-	24 s
	Par. Opt.	2 s	2 s	3 s	4 s	4 s	5 s	6 s	6 s	7 s	8 s
	GA-SVM	32 d	54 d	45 d	5 d	11 d	21 d	10 d	55 d	21 d	30 d
TRPM8 (120, 101)	SVM/STSA	1 s	2 s	-	5 s	-	9 s	-	16 s	-	23 s
	Par. Opt.	19 d	23 d	29 d	28 d	37 d	33 d	47 d	46 d	48 d	53 d
	GA-SVM	25 sn	31 sn	31 sn		27 sn	47 sn	24 sn	13 sn	8 sn	34 sn
STEAP (120, 22)	SVM/STSA	14 d	27 d	-	1 s	-	2 s	-	3 s	-	4 s
	Par. Opt.	2 d	2 d	2 d	2 d	2 d	2 d	2 d	2 d	2 d	2 d
	GA-SVM	35 sn	30 sn	44 sn	37 sn	58 sn	58 sn	57 sn	52 sn	43 sn	46 sn
ENm013 (120, 361)	Par. Opt.	2 s	3 s	4 s	5 s	6 s	6 s	7 s	10 s	10 s	15 s
	GA-SVM	40 d	32 d	27 d	43 d	12 d	54 d	45 d	32 d	43 d	27 d
ENr112 (120, 412)	Par. Opt.	3 s	4 s	5 s	6 s	10 s	11 s	13 s	16 s	20 s	24 s
	GA-SVM	4 d	12 d	35 d	47 d	32 d	20 d		55 d	26 d	53 d
ENr113 (120, 515)	Par. Opt.	5 s	7 s	10 s	10 s	13 s	17 s	19 s	22 s	25 s	29 s
	GA-SVM	18 d	12 d	21 d	54 d	52 d	15 d	59 d	21 d	37 d	19 d

Hap: Haplotip sayısı, SNP: SNP sayısı, s: saat, d: dakika, sn: saniye

6.3. Parametre Optimizasyonlu CLONTagger Yönteminin Uygulama Sonuçları

Parametre optimizasyonlu CLONTagger yönteminin Matlab programında geliştirilen grafiksel arayüzü Şekil 6.23'de verilmiştir. Şekilden de görülebildiği gibi veri kümesi seçimi ve parametre girişleri bu arayüz aracılığıyla yapılmaktadır.

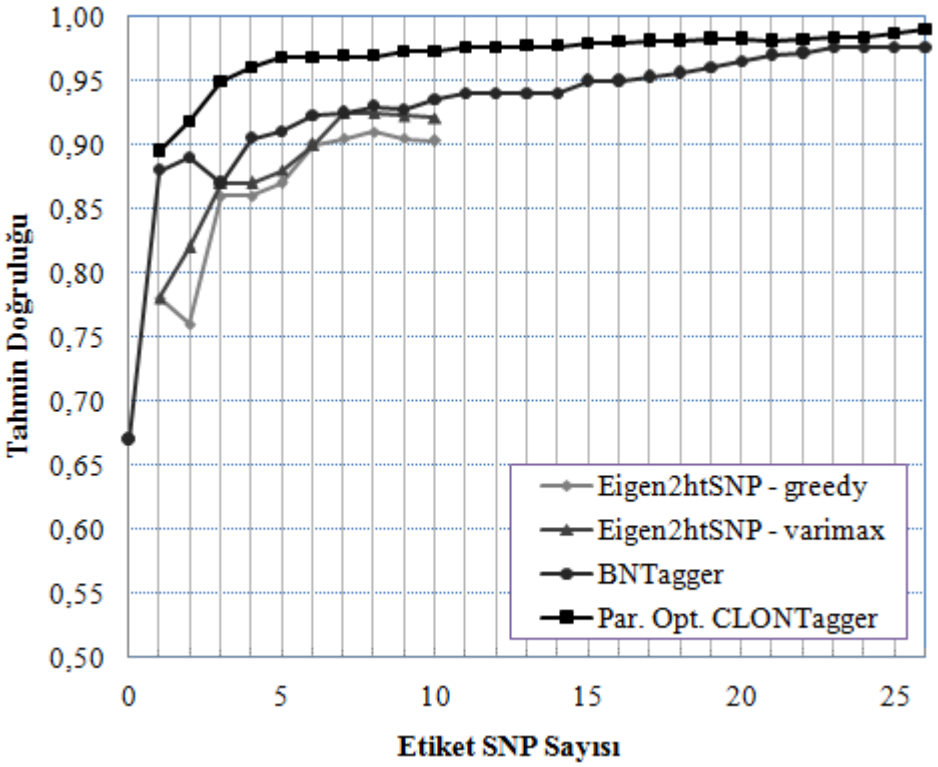
Parametre optimizasyonlu CLONTagger yönteminde, CLONALG ve PSO için iterasyon sayısı ve popülasyon büyüklüğü *20* olarak alınmıştır. Ayrıca PSO için her iki öğrenme faktörü *2* ve atelet ağırlığı da *1* olarak kabul edilmiştir. C ve γ parametrelerinin her ikisi için de $[10^{-2}, 10^{2}]$ arama aralığı kullanılmış, V_{min} değeri *-10* ve V_{max} değeri de *10* olarak kabul edilmiştir. Böylece bu parametreler, parametre optimizasyonlu CLONTagger yönteminin bütün veri kümeleri üzerindeki uygulama sonuçları için kullanılmıştır.



Şekil 6.23. Parametre optimizasyonlu CLONTagger yönteminin Matlab'da oluşturulan grafiksel arayüzü

6.3.1. ACE veri kümesi

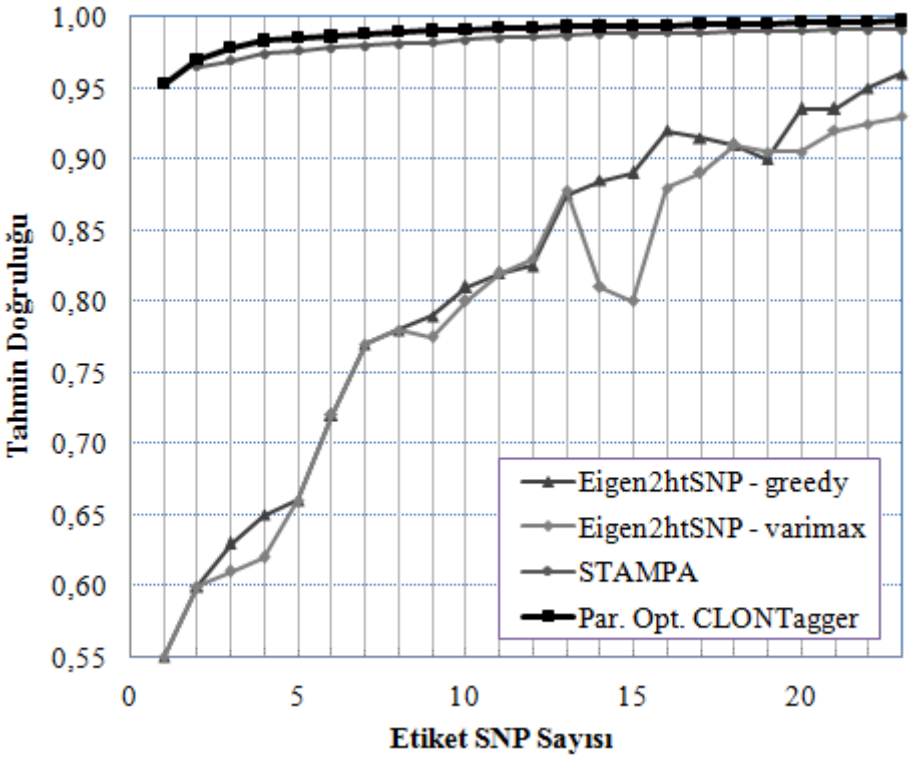
ACE veri kümesi için, önerilen parametre optimizasyonlu CLONTagger metodu, BNTagger ve Eigen2htSNP metodları ile karşılaştırılmıştır. Şekil 6.24'den de görülebildiği gibi parametre optimizasyonlu CLONTagger metodu, bütün etiket SNP sayıları için BNTagger ve Eigen2htSNP metodlarından önemli oranda daha iyi bir performans sergilemiştir. Önerilen yöntem ile 1 ile 10 etiket SNP aralığında BNTagger ve Eigen2htSNP metotlarından ortalama olarak %4.4 ve %7.3 oranında daha yüksek tahmin doğruluğu elde edilmiştir.



Şekil 6.24. ACE veri kümesi için parametre optimizasyonlu CLONTagger ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.3.2. ABCB1 veri kümesi

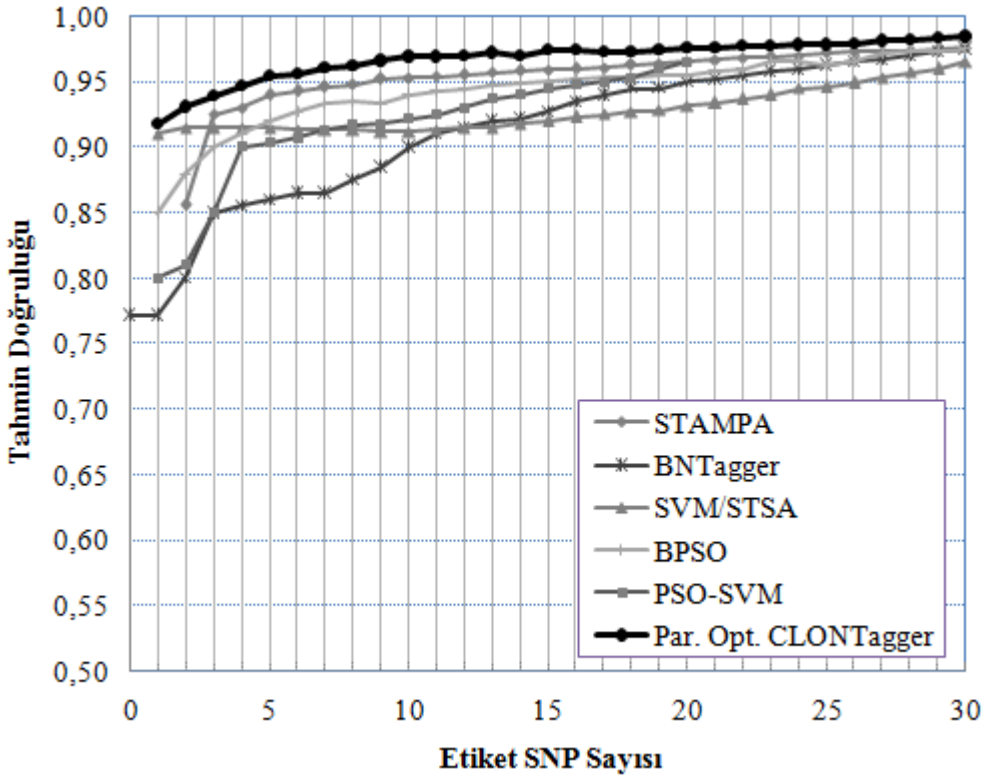
ABCB1 veri kümesi için parametre optimizasyonlu CLONTagger metodu, Eigen2htSNP ve STAMPA metodları ile karşılaştırılmıştır. İki etiket SNP için önerilen yöntem, %97 tahmin doğruluğuna ulaşırken STAMPA, %96.5 tahmin doğruluğuna ulaşmıştır. Aynı etiket SNP sayısı için Eigen2htSNP metotları ise sadece %60 tahmin doğruluğuna ulaşabilmiştir. Şekil 6.25'den de görülebileceği gibi 2 ile 23 etiket SNP aralığında parametre optimizasyonlu CLONTagger metodu, Eigen2htSNP ve STAMPA metotlarından sırasıyla %16.8 ve %0.80 oranında daha fazla tahmin doğruluğu elde etmiştir.



Şekil 6.25. ABCB1 veri kümesi için parametre optimizasyonlu CLONTagger, STAMPA ve Eigen2htSNP metotları tarafından elde edilen tahmin doğrulukları

6.3.3. LPL veri kümesi

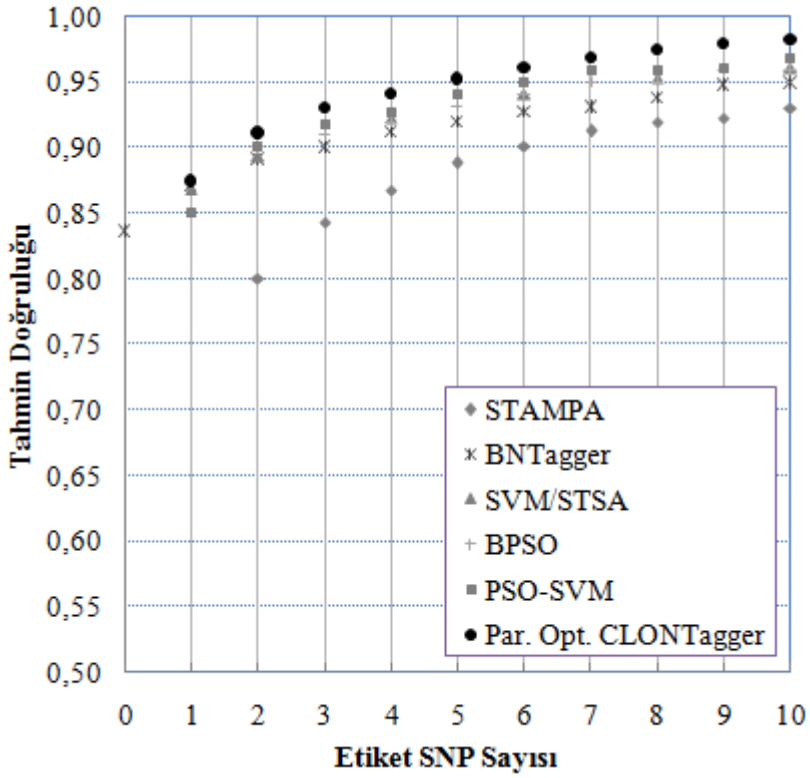
Önerilen parametre optimizasyonlu CLONTagger yöntemi, LPL veri kümesi üzerinde STAMPA, BNTagger, SVM/STSA, BPSO ve PSO-SVM metotları ile karşılaştırılmıştır. Örneğin, iki etiket SNP için önerilen yöntem, %93.1 tahmin doğruluğu üretirken diğer metotlar içerisinde en iyi sonucu üreten SVM/STSA metodu ancak %91.5 tahmin doğruluğu üretmektedir. Şekil 6.26'dan da görülebileceği gibi 2 ile 20 etiket SNP aralığında parametre optimizasyonlu CLONTagger metodu, PSO-SVM, BPSO, SVM/STSA, BNTagger ve STAMPA metodlarından sırasıyla ortalamada %4.3, %2.8, %4.6, %6.6 ve %1.7 daha fazla oranda tahmin doğruluğu sergilemiştir.



Şekil 6.26. LPL veri kümesi üzerinde parametre optimizasyonlu CLONTagger ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.3.4. 5q31 veri kümesi

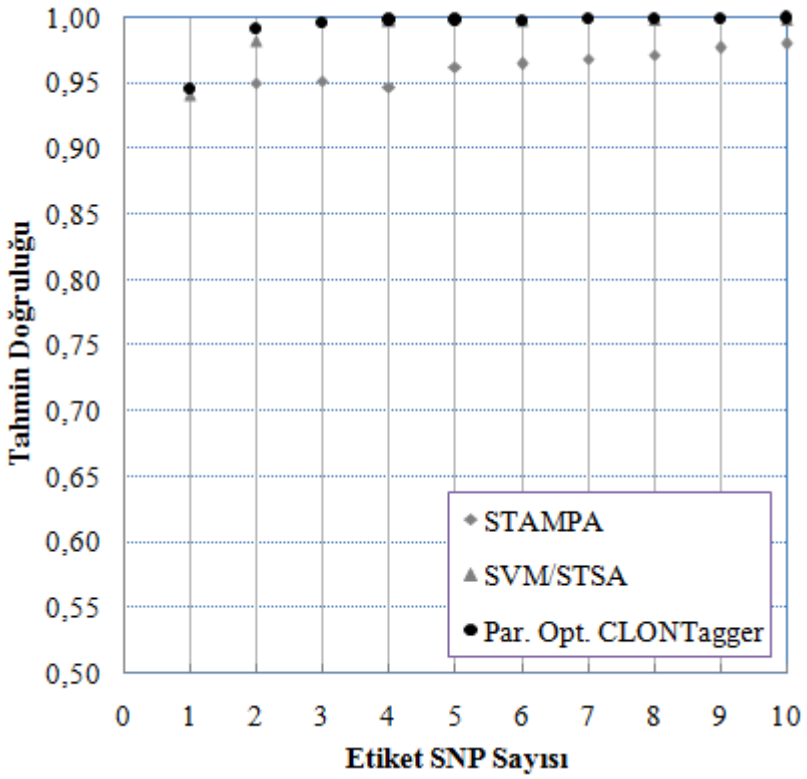
Önerilen parametre optimizasyonlu CLONTagger yöntemi, 5q31 veri kümesi üzerinde STAMPA, BNTagger, SVM/STSA, BPSO ve PSO-SVM metotları ile karşılaştırılmıştır. Bu veri kümesi üzerinde 2 ile 10 etiket SNP aralığında önerilen yöntem PSO-SVM, BPSO, SVM/STSA, BNTagger ve STAMPA yöntemlerine göre ortalamada %1.3, %2.0, %2.1, %3.1 ve %6.8 oranında daha iyi bir tahmin doğruluğu göstermiştir. Şekil 6.27, parametre optimizasyonlu CLONTagger ve diğer yöntemler tarafından elde edilen tahmin doğruluklarını göstermektedir.



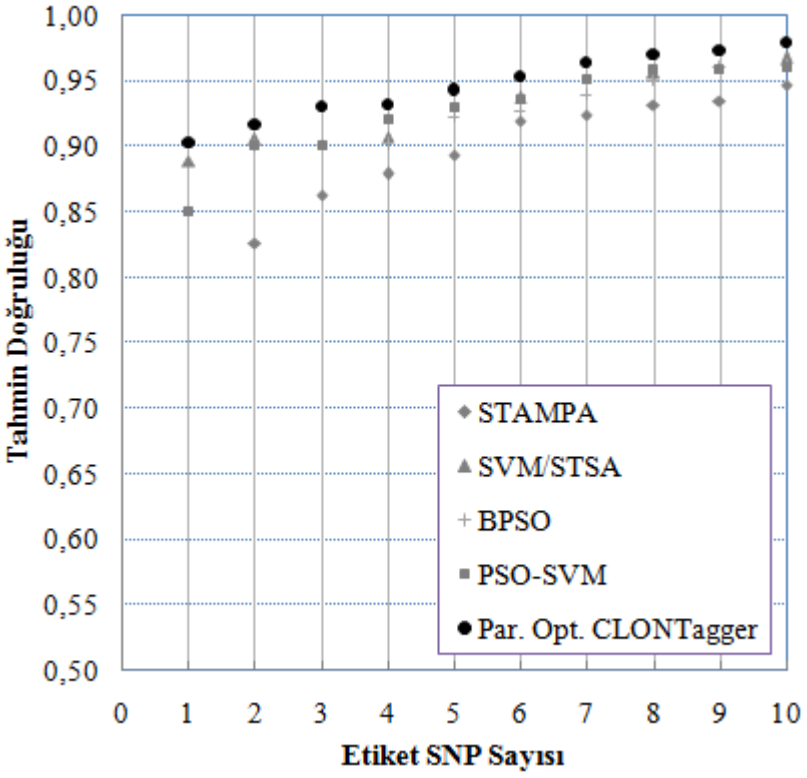
Şekil 6.27. 5q31 veri kümesi üzerinde parametre optimizasyonlu CLONTagger ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.3.5. STEAP ve TRPM8 veri kümeleri

Önerilen parametre optimizasyonlu CLONTagger yöntemi, STEAP veri kümesi üzerinde STAMPA ve SVM/STSA yöntemleri ile TRPM8 veri kümesi üzerinde de STAMPA, SVM/STSA, BPSO ve PSO-SVM yöntemleri ile karşılaştırılmıştır. STEAP veri kümesi üzerinde elde edilen sonuçlar Şekil 6.28, TRPM8 veri kümesi üzerinde elde edilen sonuçlar ise Şekil 6.29'da gösterilmektedir.



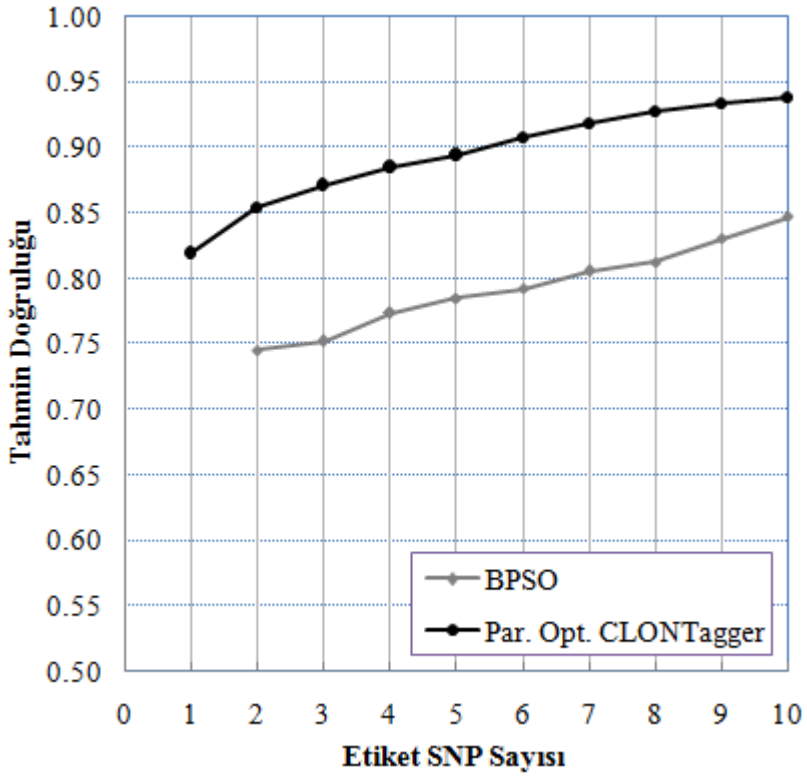
Şekil 6.28. STEAP veri kümesi üzerinde parametre optimizasyonlu CLONTagger, STAMPA ve SVM/STSA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.29. TRPM8 veri kümesi üzerinde parmetre optimizasyonlu CLONTagger ve diğer metotlar tarafından elde edilen tahmin doğrulukları

6.3.6. Popülasyon D'nin D9 veri kümesi

Parametre optimizasyonlu CLONTagger yöntemi, popülasyon D'nin D9 veri kümesi üzerinde BPSO yöntemi ile karşılaştırılmıştır. 2 ile 10 etiket SNP aralığında önerilen yöntem, BPSO yöntemine göre %11.3 oranında daha iyi bir tahmin doğruluğu göstermiştir. Şekil 6.30, D9 veri kümesi üzerinde her iki yöntem tarafından elde edilen tahmin doğruluklarını göstermektedir.

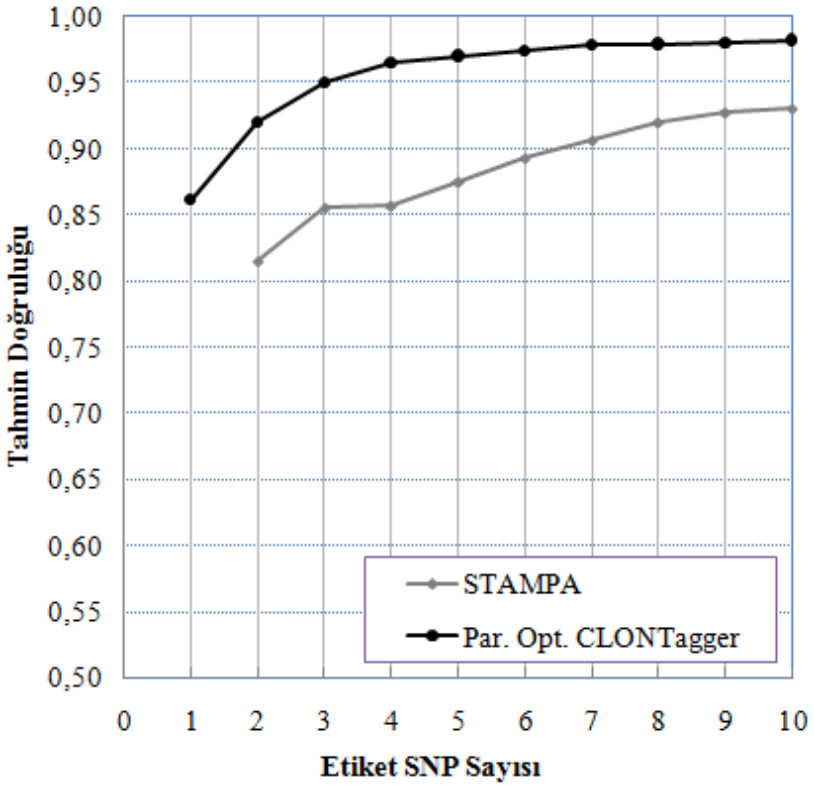


Şekil 6.30. D9 veri kümesi üzerinde parametre optimizasyonlu CLONTagger ve BPSO metotları tarafından elde edilen tahmin doğrulukları

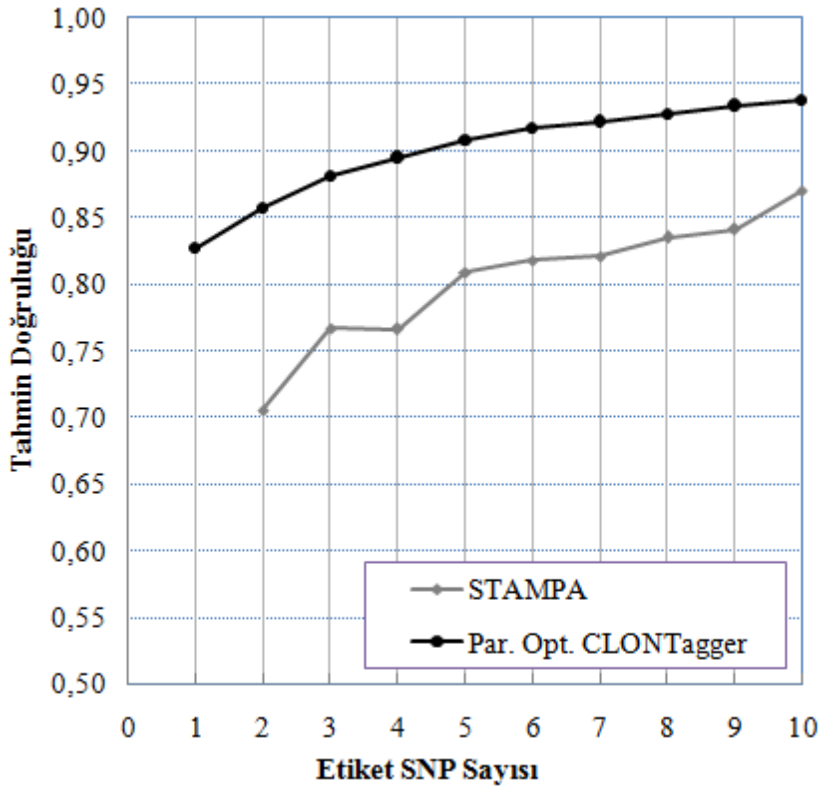
6.3.7. ENm013, ENr112 ve ENr113 veri kümeleri

Parametre optimizasyonlu CLONTagger yönteminin daha büyük veri kümeleri üzerindeki performansını değerlendirmek için ENm013, ENr112 ve ENr113 veri kümeleri kullanılmıştır. Bu veri kümeleri üzerinde elde edilen sonuçlar STAMPA metoduyla karşılaştırılmış ve aşağıdaki şekillerde verilmiştir. Şekil 6.31, Şekil 6.32 ve Şekil 6.33'den de görülebildiği gibi önerilen parametre optimizasyonlu CLONTagger yöntemi, ENm013, ENr112 ve ENr113 veri kümeleri üzerinde yapılan deneylerde 2 ile

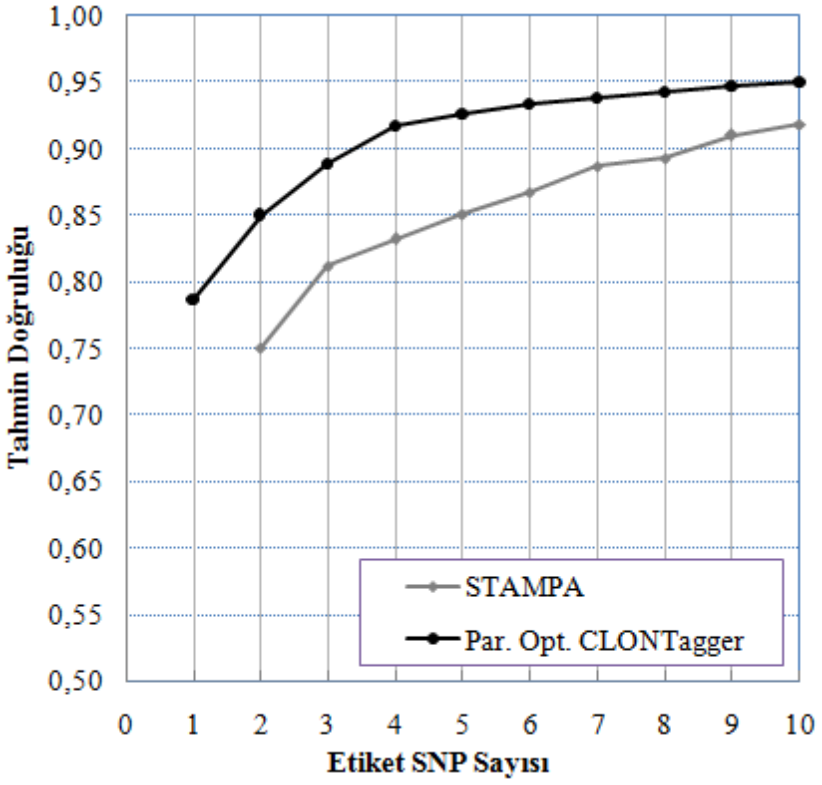
10 etiket SNP aralığında STAMPA metoduna göre ortalamada %8.0, %10.5 ve %6.3 daha yüksek bir tahmin doğruluğu sergilemiştir.



Şekil 6.31. ENm013 veri kümesi üzerinde parametre optimizasyonlu CLONTagger ve STAMPA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.32. ENr112 veri kümesi üzerinde parametre optimizasyonlu CLONTagger ve STAMPA metotları tarafından elde edilen tahmin doğrulukları



Şekil 6.33. ENr113 veri kümesi üzerinde parametre optimizasyonlu CLONTagger ve STAMPA metotları tarafından elde edilen tahmin doğrulukları

6.3.8. Parametre optimizasyonlu CLONTagger yönteminin çalışma süresi

Önerilen yöntemin farklı sayıdaki etiket SNP'ler için elde edilen çalışma süreleri Çizelge 6.3'de verilmiştir. SVM/STSA yönteminin 5q31, TRPM8 ve STEAP veri kümeleri üzerindeki çalışma süreleri ise orjinal yayınlarından alınmıştır. STAMPA metodunun 2 ile 10 etiket SNP aralığında tüm veri kümeleri üzerindeki çalışma süreleri deneyler yapılarak belirlenmiş ancak bu aralıkta süre değişimi çok az olduğu için tabloda belirtilmemiştir. Örneğin STAMPA metodu kullandığımız en küçük veri kümesi olan ACE için 2'den 10'a kadar olan tüm etiket SNP'leri yaklaşık 2 saniyede belirlemektedir. Buna karşılık önerilen metot, 2 etiket SNP'i 5 dakika 6 saniyede belirlerken 10 etiket SNP'i ise 19 dakika 7 saniyede tespit etmektedir. 5q31 veri kümesi için 2'den 10'a kadar olan tüm etiket SNP'leri STAMPA metodu, 7 saniyede seçmektedir. Buna karşılık SVM/STSA metodu, 2 etiket SNP'i 5 saatte 10 etiket SNP'i ise 24 saatte seçmektedir. Önerilen metot ise 2 etiket SNP'i 2 saat 54 dakikada 10 etiket SNP'i ise 8 saat 30 dakikada belirlemiştir. En büyük veri kümesi olan ENr113 üzerinde yapılan deneyler, STAMPA metodunun 2'den 10'a kadar olan tüm etiket SNP'leri yaklaşık 6 dakikada seçtiğini göstermiştir. Önerilen metot ise 2 etiket SNP'i 7 saat 20 dakikada 10 etiket SNP'i ise 52 saat 2 dakikada belirlemiştir. Yapılan deneylerden de anlaşıldığı gibi parametre optimizasyonlu CLONTagger metodu, SVM/STSA metoduna göre daha hızlı çalışırken (5q31 veri kümesinin bazı etiket SNP sayıları hariç) STAMPA medoduna göre daha yavaş çalışmaktadır.

Çizelge 6.3. Kullanılan veri kümeleri üzerinde farklı etiket SNP sayıları için parametre optimizasyonlu CLONTagger ve SVM/STSA yöntemlerinin çalışma süreleri

Veri kümeleri (Hap, SNP)	Yöntemler	Etiket SNP Sayısı									
		1	2	3	4	5	6	7	8	9	10
ACE (22, 52)	Par. Opt.	3 d	5 d	7 d	8 d	12 d	11 d	14 d	16 d	19 d	19 d
	CLONTagger	41 sn	6 sn	31 sn	46 sn	35 sn	59 sn	46 sn	6 sn	48 sn	7 sn
ABCB1 (494, 27)	Par. Opt.	4 s	4 s	3 s	3 s	3 s	3 s	3 s	3 s	3 s	3 s
	CLONTagger	48 d	24 d	56 d	46 d	31 d	26 d	22 d	20 d	21 d	10 d
LPL (88, 88)	Par. Opt.	18 d	26 d	33 d	41 d	43 d	1 s	1 s	1 s	2 s	2 s
	CLONTagger	24 sn	4 sn		56 sn	8 sn	2 d	30 d	32 d	5 d	3 d
D9 (180, 49)	Par. Opt.	54 d	1 s	1 s	1 s	1 s	2 s	2 s	2 s	3 s	3 s
	CLONTagger	4 sn	1 d	13 d	20 d	46 d	10 d	21 d	53 d	21 d	41 d
5q31 (258, 103)	SVM/STSA	3 s	5 s	-	11 s	-	16 s	-	18 s	-	24 s
	Par. Opt.	5 s	7 s	7 s	8 s	9 s	15 s	19 s	21 s	26 s	29 s
	CLONTagger	59 d	16 d	2 d	47 d	32 d	10 d	5 d	19 d	28 d	11 d
TRPM8 (120, 101)	SVM/STSA	1 s	2 s	-	5 s	-	9 s	-	16 s	-	23 s
	Par. Opt.	44 d	48 d	57 d	1 s	1 s	1 s	1 s	2 s	2 s	2 s
	CLONTagger	21 sn	44 sn	22 sn	1 d	14 d	20 d	26 d	22 d	32 d	38 d
STEAP (120, 22)	SVM/STSA	14 d	27 d	-	1 s	-	2 s	-	3 s	-	4 s
	Par. Opt.	9 d	9 d	9 d	8 d	9 d	10 d	9 d	10 d	10 d	9 d
	CLONTagger	8 sn	20 sn	11 sn	56 sn	47 sn	29 sn	56 sn		9 sn	51 sn
ENm013 (120, 361)	Par. Opt.	2 s	3 s	4 s	6 s	8 s	10 s	13 s	18 s	24 s	27 s
	CLONTagger	38 d	18 d	2 d	14 d	10 d	30 d	48 d	56 d	15 d	33 d
ENr112 (120, 412)	Par. Opt.	3 s	4 s	5 s	8 s	12 s	15 s	20 s	28 s	37 s	44 s
	CLONTagger	19 d	16 d	46 d	40 d	22 d	16 d	2 d	17 d	49 d	11 d
ENr113 (120, 515)	Par. Opt.	5 s	7 s	10 s	13 s	16 s	23 s	30 s	37 s	44 s	52 s
	CLONTagger	43 d	20 d	41 d	57 d	18 d	15 d	48 d	23 d	35 d	2 d

Hap: Haplotip sayısı, SNP: SNP sayısı, s: saat, d: dakika, sn: saniye

6.4. Geliştirilen Yöntemlerin Tahmin Doğruluklarının ve Çalışma Sürelerinin Karşılaştırılması

Bu tez çalışmasında geliştirilen üç farklı yöntemin her biri farklı büyüklük ve özelliklere sahip 10 farklı veri kümesi üzerinde çalıştırılmış ve farklı etiket SNP sayıları için her birinin tahmin doğrulukları ve çalışma süreleri elde edilmiştir. Geliştirilen yöntemlerin, 1 ile 10 etiket SNP aralığında kullanılan bütün veri kümeleri üzerinde elde edilen tahmin doğrulukları Çizelge 6.4'de ve çalışma süreleri de Çizelge 6.5'de toplu olarak verilmiştir. Çizelgelerden de görülebileceği gibi parametre optimizasyonlu CLONTagger yöntemi, kullanılan bütün veri kümeleri için 1 ile 10 arasındaki farklı etiket SNP sayıları için diğer iki yönteme göre daha yavaş çalışmaktadır. Buna karşılık aynı yöntem, diğer iki yönteme göre daha yüksek bir tahmin doğruluğu göstermektedir.

Çizelge 6.4. Geliştirilen yöntemlerin farklı veri kümeleri üzerinde farklı etiket SNP sayıları için elde edilen tahmin doğrulukları

Veri kümeleri (Hap, SNP)	Yöntemler	Etiket SNP Sayısı									
		1	2	3	4	5	6	7	8	9	10
ACE (22, 52)	GA-SVM	87.3	90.4	91.5	92.8	93.5	94.4	94.8	95.4	95.3	96.2
	Par. Opt. GA-SVM	89.5	91.8	94.9	96.0	96.8	96.8	96.9	96.9	96.8	97.0
	Par. Opt. CLONTagger	89.5	91.8	94.9	96.0	96.8	96.8	96.9	96.9	97.1	97.3
ABCB1 (494, 27)	GA-SVM	95.3	97.0	97.8	98.3	98.4	98.6	98.7	98.8	98.9	99.1
	Par. Opt. GA-SVM	95.3	97.0	97.8	98.3	98.6	98.6	98.8	98.9	99.0	99.1
	Par. Opt. CLONTagger	95.3	97.0	98.0	98.5	98.6	98.8	99.0	99.1	99.2	99.3
LPL (88, 88)	GA-SVM	91.8	92.9	93.6	94.2	94.5	94.8	95.1	95.3	95.5	95.8
	Par. Opt. GA-SVM	91.8	93.1	93.9	94.7	95.0	95.6	95.8	96.2	96.2	96.6
	Par. Opt. CLONTagger	91.8	93.1	93.9	94.7	95.4	95.6	96.1	96.2	96.6	96.9
D9 (180, 49)	GA-SVM	81.9	84.4	86.2	87.2	88.4	89.1	89.9	91.0	91.9	91.8
	Par. Opt. GA-SVM	81.9	85.2	87.1	87.9	89.1	90.4	92.0	92.4	93.2	93.4
	Par. Opt. CLONTagger	81.9	85.4	87.1	88.5	89.4	90.7	91.8	92.7	93.3	93.8
5q31 (258, 103)	GA-SVM	86.9	90.2	92.2	92.9	94.5	95.2	96.3	96.7	97.5	97.9
	Par. Opt. GA-SVM	86.8	91.1	92.6	93.8	95.0	95.9	96.5	97.2	97.6	97.9
	Par. Opt. CLONTagger	87.4	91.1	93.0	94.1	95.2	96.1	96.8	97.4	97.9	98.2
TRPM8 (120, 101)	GA-SVM	90.0	90.6	92.2	93.0	93.3	94.5	95.3	96.2	96.6	97.2
	Par. Opt. GA-SVM	90.0	91.4	92.8	93.1	94.1	95.1	96.1	96.9	97.1	97.8
	Par. Opt. CLONTagger	90.2	91.6	93.0	93.2	94.3	95.3	96.4	97.0	97.3	97.9
STEAP (120, 22)	GA-SVM	94.4	98.3	99.5	99.7	99.7	99.8	99.8	99.9	99.9	99.9
	Par. Opt. GA-SVM	94.5	99.1	99.6	99.8	99.8	99.8	99.9	99.9	99.9	99.9
	Par. Opt. CLONTagger	94.5	99.1	99.6	99.8	99.8	99.8	99.9	99.9	99.9	99.9
ENm013 (120, 361)	GA-SVM	86.1	91.4	94.3	96.0	96.4	96.9	97.3	97.4	97.5	97.6
	Par. Opt. GA-SVM	84.6	91.7	94.8	96.3	96.8	97.2	97.6	97.6	97.7	97.9
	Par. Opt. CLONTagger	86.1	92.0	95.0	96.5	97.0	97.4	97.8	97.9	98.0	98.2
ENr112 (120, 412)	GA-SVM	82.4	84.8	86.8	88.2	89.8	90.3	91.6	92	92.8	93.0
	Par. Opt. GA-SVM	82.7	85.3	88.0	89.0	90.3	91.4	91.7	92.0	92.7	93.0
	Par. Opt. CLONTagger	82.7	85.7	88.1	89.5	90.8	91.7	92.2	92.7	93.4	93.8

ENr113 (120, 515)	GA-SVM	78.5	84.3	88	90.9	91.6	92.3	92.7	93.5	93.7	94.1
	Par. Opt. GA-SVM	78.6	84.8	88.5	91.2	92.1	92.7	93.3	93.8	94.2	94.5
	Par. Opt. CLONTagger	78.6	85.0	88.8	91.7	92.6	93.3	93.8	94.2	94.7	95.0

Hap: Haplotip sayısı, SNP: SNP sayısı

Çizelge 6.5. Geliştirilen yöntemlerin farklı veri kümeleri üzerinde farklı etiket SNP sayıları için elde edilen çalışma süreleri

Veri kümeleri (Hap, SNP)	Yöntemler	Etiket SNP Sayısı									
		1	2	3	4	5	6	7	8	9	10
ACE (22, 52)	GA-SVM	1 d 56 sn	2 d 45 sn	3 d 28 sn	3 d 38 sn	5 d 26 sn	5 d 23 sn	5 d 45 sn	5 d 23 sn	6 d 44 sn	6 d 52 sn
	Par. Opt. GA-SVM	1 d 39 sn	2 d 16 sn	3 d 25 sn	4 d 17 sn	4 d 58 sn	4 d 44 sn	5 d 52 sn	6 d	6 d 25 sn	6 d 41 sn
	Par. Opt. CLONTagger	3 d 41 sn	5 d 6 sn	7 d 31 sn	8 d 46 sn	12 d 35 sn	11 d 59 sn	14 d 46 sn	16 d 6 sn	19 d 48 sn	19 d 7 sn
ABCB1 (494, 27)	GA-SVM	1 s 19 d	1 s 8 d	1 s 3 d	51 d 58 sn	57 d 14 sn	52 d 4 sn	50 d 59 sn	44 d 32 sn	41 d 49 sn	37 d 35 sn
	Par. Opt. GA-SVM	1 s 15 d	1 s 7 d	58 d 43 sn	57 d 23 sn	51 d 23 sn	53 d 11 sn	47 d 54 sn	43 d 41 sn	40 d 10 sn	37 d 30 sn
	Par. Opt. CLONTagger	4 s 48 d	4 s 24 d	3 s 56 d	3 s 46 d	3 s 31 d	3 s 26 d	3 s 22 d	3 s 20 d	3 s 21 d	3 s 10 d
LPL (88, 88)	GA-SVM	7 d 38 sn	12 d 50 sn	13 d 56 sn	21 d 40 sn	22 d 54 sn	23 d 54 sn	29 d 45 sn	36 d 6 sn	35 d 48 sn	40 d 45 sn
	Par. Opt. GA-SVM	7 d 35 sn	11 d 13 sn	15 d 41 sn	19 d 23 sn	22 d 4 sn	26 d 19 sn	30 d 58 sn	35 d 58 sn	38 d 55 sn	44 d 24 sn
	Par. Opt. CLONTagger	18 d 24 sn	26 d 4 sn	33 d	41 d 56 sn	43 d 8 sn	1 s 2 d	1 s 30 d	1 s 32 d	2 s 5 d	2 s 3 d
D9 (180, 49)	GA-SVM	17 d 29 sn	21 d 2 sn	25 d 44 sn	29 d 56 sn	35 d 56 sn	43 d 47 sn	54 d 45 sn	57 d 4 sn	57 d 59 sn	1 s 3 d
	Par. Opt. GA-SVM	14 d 8 sn	19 d 5 sn	23 d 3 sn	31 d 23 sn	34 d 3 sn	42 d 40 sn	52 d 57 sn	53 d 42 d	1 s 4 d	1 s 17 d
	Par. Opt. CLONTagger	54 d 4 sn	1 s 1 d	1 s 13 d	1 s 20 d	1 s 46 d	2 s 10 d	2 s 21 d	2 s 53 d	3 s 21 d	3 s 41 d
5q31 (258, 103)	GA-SVM	2 s 12 d	3 s	3 s 43 d	4 s 5 d	4 s 49 d	5 s 36 d	6 s 59 d	6 s 30 d	8 s 13 d	8 s 55 d
	Par. Opt. GA-SVM	2 s 32 d	2 s 54 d	3 s 45 d	4 s 5 d	4 s 11 d	5 s 21 d	6 s 10 d	6 s 55 d	7 s 21 d	8 s 30 d
	Par. Opt. CLONTagger	5 s 59 d	7 s 16 d	7 s 2 d	8 s 47 d	9 s 32 d	15 s 10 d	19 s 5 d	21 s 19 d	26 s 28 d	29 s 11 d
TRPM8 (120, 101)	GA-SVM	20 d 51 sn	24 d 3 sn	24 d 17 sn	26 d 21 sn	33 d 6 sn	41 d 36 sn	42 d 37 sn	47 d 17 sn	49 d 20 sn	53 d 7 sn
	Par. Opt. GA-SVM	19 d 25 sn	23 d 31 sn	29 d 31 sn	28 d	37 d 27 sn	33 d 47 sn	47 d 24 sn	46 d 13 sn	48 d 8 sn	53 d 34 sn
	Par. Opt. CLONTagger	44 d 21 sn	48 d 44 sn	57 d 22 sn	1 s 1 d	1 s 14 d	1 s 20 d	1 s 26 d	2 s 22 d	2 s 32 d	2 s 38 d
STEAP (120, 22)	GA-SVM	2 d 41 sn	2 d 37 sn	2 d 18 sn	2 d 40 sn	2 d 45 sn	3 d 1 sn	2 d 40 sn	2 d 24 sn	2 d 27 sn	2 d 20 sn
	Par. Opt. GA-SVM	2 d 35 sn	2 d 30 sn	2 d 44 sn	2 d 37 sn	2 d 58 sn	2 d 58 sn	2 d 57 sn	2 d 52 sn	2 d 43 sn	2 d 46 sn
	Par. Opt. CLONTagger	9 d 8 sn	9 d 20 sn	9 d 11 sn	8 d 56 sn	9 d 47 sn	10 d 29 sn	9 d 56 sn	10 d	10 d 9 sn	9 d 51 sn

ENm013 (120, 361)	GA-SVM	2 s 10 d	2 s 55 d	3 s 51 d	4 s 43 d	5 s 41 d	7 s 13 d	8 s 45 d	9 s 10 d	9 s 43 d	11 s 23 d
	Par. Opt. GA-SVM	2 s 40 d	3 s 32 d	4 s 27 d	5 s 43 d	6 s 12 d	6 s 54 d	7 s 45 d	10 s 32 d	10 s 43 d	15 s 27 d
	Par. Opt. CLONTagger	2 s 38 d	3 s 18 d	4 s 2 d	6 s 14 d	8 s 10 d	10 s 30 d	13 s 48 d	18 s 56 d	24 s 15 d	27 s 33 d
ENr112 (120, 412)	GA-SVM	3 s 11 d	3 s 45 d	5 s 18 d	6 s 23 d	8 s 36 d	10 s	11 s 53 d	15 s 16 d	21 s 6 d	23 s 36 d
	Par. Opt. GA-SVM	3 s 4 d	4 s 12 d	5 s 35 d	6 s 47 d	10 s 32 d	11 s 20 d	13 s	16 s 55 d	20 s 26 d	24 s 53 d
	Par. Opt. CLONTagger	3 s 19 d	4 s 16 d	5 s 46 d	8 s 40 d	12 s 22 d	15 s 16 d	20 s 2 d	28 s 17 d	37 s 49 d	44 s 11 d
ENr113 (120, 515)	GA-SVM	5 s 8 d	6 s 40 d	8 s 26 d	10 s 33 d	13 s 53 d	16 s 41 d	19 s 26 d	21 s 40 d	24 s 43 d	28 s 3 d
	Par. Opt. GA-SVM	5 s 18 d	7 s 12 d	10 s 21 d	10 s 54 d	13 s 52 d	17 s 15 d	19 s 59 d	22 s 21 d	25 s 37 d	29 s 19 d
	Par. Opt. CLONTagger	5 s 43 d	7 s 20 d	10 s 41 d	13 s 57 d	16 s 18 d	23 s 15 d	30 s 48 d	37 s 23 d	44 s 35 d	52 s 2 d

Hap: Haplotip sayısı, SNP: SNP sayısı, s: saat, d: dakika, sn: saniye

Geliştirilen üç yöntem içerisinde etiket SNP'leri temsil eden 1'lerin sayısının algoritmaya giriş olarak verilen sayıya eşitlenmesi için ayarlama prosedürü kullanılmıştır. Ayarlama prosedürü içerisinde ise lokal arama algoritması kullanılmıştır. Böylelikle, aday kromozom veya antikorlar içerisinde en iyi tahmin doğruluklu olanlar belirlenmiş ve yeni popülasyona dahil edilmiştir. Bu nedenle, farklı veri kümeleri üzerinde yapılan deneylerde geliştirilen her üç yöntem ile diğer yöntemlere göre daha yüksek tahmin doğrulukları elde edilmiştir.

GA-SVM, parametre optimizasyonlu GA-SVM ve parametre optimizasyonlu CLONTagger yöntemlerinde rastgele arama algoritması kullanan ayarlama prosedürü kullanılarak da çeşitli deneyler yapılmıştır. Farklı veri kümeleri üzerinde farklı etiket SNP sayıları için elde edilen sonuçlar Çizelge 6.6 ve Çizelge 6.7'de verilmiştir. Rastgele arama algoritmasını kullanan üç yöntemin tahmin doğrulukları Çizelge 6.6'da ve çalışma süreleri ise Çizelge 6.7'de verilmiştir.

Çizelge 6.4 ve Çizelge 6.6'dan da görülebildiği gibi lokal arama algoritmasını kullanan yöntemlerin farklı veri kümeleri üzerinde farklı sayıdaki etiket SNP'ler için elde edilen tahmin doğrulukları rastgele arama algoritmasını kullanan yöntemlere göre genellikle daha yüksektir. Bu farkın diğer iki yönteme göre GA-SVM yönteminde daha fazla olduğu görülmektedir. Benzer şekilde Çizelge 6.5 ve Çizelge 6.7 incelendiğinde ise farklı veri kümeleri üzerinde farklı sayıdaki etiket SNP'ler için rastgele arama algoritmasını kullanan yöntemler için elde edilen çalışma süreleri lokal arama

algoritmasını kullanan yöntemlere göre daha düşüktür. Her üç yöntem içinde, etiket SNP'lerin sayısının artmasıyla birlikte çalışma süreleri arasındaki farkda artmaktadır. Bu durum rastgele arama algoritmasını kullanan yöntemlerin 1 etiket SNP'i belirlerken harcadıkları çalışma süresi ile 10 etiket SNP'i belirlerken harcadıkları çalışma süresi arasındaki farkın azlığından kaynaklanmaktadır.

Çizelge 6.6. Rastgele arama algoritması kullanılarak geliştirilen yöntemler için farklı veri kümeleri üzerinde farklı etiket SNP sayıları için elde edilen tahmin doğrulukları

Veri kümeleri (Hap, SNP)	Yöntemler	Etiket SNP Sayısı									
		1	2	3	4	5	6	7	8	9	10
ACE (22, 52)	GA-SVM	87.3	90.4	91.5	92.8	92.6	93.1	92.9	93.7	94.1	93.9
	Par. Opt. GA-SVM	89.5	91.6	94.6	94.9	96.8	96.3	96.4	96.2	96.3	96.6
	Par. Opt. CLONTagger	89.5	91.8	94.9	96.0	96.8	96.8	96.9	96.9	97.0	97.1
ABCB1 (494, 27)	GA-SVM	95.3	96.5	97.7	97.6	97.8	98.4	98.0	98.1	98.7	98.3
	Par. Opt. GA-SVM	95.3	96.8	97.8	98.1	98.3	98.5	98.6	98.8	98.9	99.0
	Par. Opt. CLONTagger	95.3	97.0	97.8	98.3	98.5	98.7	98.8	98.9	99.0	99.1
LPL (88, 88)	GA-SVM	91.8	92.4	92.8	93.1	93.3	93.3	93.5	94.0	94.2	94.6
	Par. Opt. GA-SVM	91.8	92.7	93.2	93.9	93.9	93.7	94.1	94.5	94.6	94.8
	Par. Opt. CLONTagger	91.8	93.1	93.8	94.4	94.8	95.2	95.9	95.8	96.1	96.6
D9 (180, 49)	GA-SVM	81.9	84.1	85.5	85.9	86.5	87.6	88.1	88.9	89.6	90.1
	Par. Opt. GA-SVM	81.4	84.4	85.9	87.8	87.3	88.4	89.6	90.0	91.6	91.4
	Par. Opt. CLONTagger	81.9	84.8	86.6	87.5	88.7	89.7	91.5	92.0	92.6	93.1
5q31 (258, 103)	GA-SVM	86.8	90.1	91.3	92.6	93.5	94.3	94.9	95.4	95.8	96.3
	Par. Opt. GA-SVM	86.8	90.3	91.6	93.5	94.3	95.6	95.9	96.3	96.8	97.3
	Par. Opt. CLONTagger	86.8	90.3	92.5	94.0	95.1	96.0	96.3	96.9	97.3	97.7
TRPM8 (120, 101)	GA-SVM	89.5	90.2	91.2	92.0	92.8	93.5	94.3	95.4	95.6	96.3
	Par. Opt. GA-SVM	89.5	91.0	92.1	92.4	93.7	94.6	95.8	96.2	96.6	97.1
	Par. Opt. CLONTagger	89.5	91.2	92.4	92.7	93.9	95.1	96.0	96.5	96.9	97.3
STEAP (120, 22)	GA-SVM	93.4	97.3	98.8	99.1	99.2	99.2	99.3	99.4	99.5	99.5
	Par. Opt. GA-SVM	93.5	98.1	99.0	99.2	99.3	99.4	99.4	99.6	99.7	99.8
	Par. Opt. CLONTagger	93.7	98.3	99.2	99.4	99.4	99.6	99.6	99.7	99.7	99.8
ENm013 (120, 361)	GA-SVM	84.5	89.9	92.8	93.3	93.5	94.5	95.3	95.9	96.1	96.4
	Par. Opt. GA-SVM	84.6	91.4	93.2	94.5	96.2	96.8	97.1	97.4	97.6	97.9
	Par. Opt. CLONTagger	85.2	91.7	94.1	95.9	97.0	97.5	98.0	97.9	98.3	98.2
ENr112 (120, 412)	GA-SVM	83.1	84.9	86.8	87.6	88.5	89.4	89.8	90.3	90.5	91.0
	Par. Opt. GA-SVM	83.0	85.3	86.6	89.0	89.7	90.1	90.5	91.1	91.5	92.8
	Par. Opt. CLONTagger	83.1	86.2	87.4	89.3	90.7	91.7	92.3	92.5	93.2	93.7

ENr113 (120, 515)	GA-SVM	80.5	84.6	88.4	90.3	91.1	91.7	92.3	92.6	93.3	93.6
	Par. Opt. GA-SVM	80.7	85.5	88.6	91.0	91.3	92.4	93.6	94.2	94.5	95.1
	Par. Opt. CLONTagger	80.7	86.2	89.4	92.1	92.9	93.7	94.3	94.8	95.3	95.7

Hap: Haplotip sayısı, SNP: SNP sayısı

Çizelge 6.7. Rastgele arama algoritması kullanılarak geliştirilen yöntemler için farklı veri kümeleri üzerinde farklı etiket SNP sayıları için elde edilen çalışma süreleri

Veri kümeleri (Hap, SNP)	Yöntemler	Etiket SNP Sayısı									
		1	2	3	4	5	6	7	8	9	10
ACE (22, 52)	GA-SVM	23 sn	22 sn	23 sn	22 sn	22 sn	22 sn	22 sn	21 sn	23 sn	22 sn
	Par. Opt. GA-SVM	22 sn	23 sn	23 sn	22 sn	23 sn	23 sn	22 sn	21 sn	23 sn	23 sn
	Par. Opt.	1 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d
	CLONTagger	27 sn	27 sn	26 sn	28 sn	28 sn	31 sn	28 sn	30 sn	28 sn	33 sn
ABCB1 (494, 27)	GA-SVM	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s	56 d	58 d
		48 d	40 d	30 d	30 d	18 d	14 d	15 d	18 d	29 sn	6 sn
	Par. Opt.	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s	56 d	54 d
	GA-SVM	44 d	41 d	37 d	26 d	16 d	10 d	4 d	1 d	55 sn	46 sn
	Par. Opt.	4 s	4 s	3 s	3 s	3 s	3 s	2 s	2 s	2 s	2 s
	CLONTagger	43 d	18 d	48 d	31 d	21 d	13 d	56 d	48 d	34 d	32 d
LPL (88, 88)	GA-SVM	4 d	5 d	5 d	5 d	5 d	5 d	5 d	5 d	5 d	5 d
		57 sn	2 sn	6 sn	15 sn	22 sn	29 sn	35 sn	37 sn	43 sn	51 sn
	Par. Opt.	4 d	4 d	4 d	5 d	5 d	5 d	5 d	5 d	5 d	5 d
	GA-SVM	50 sn	54 sn	52 sn		2 sn	10 sn	19 sn	18 sn	32 sn	55 sn
	Par. Opt.	17 d	17 d	19 d	19 d	20 d	23 d	23 d	22 d	28 d	25 d
	CLONTagger	57 sn	58 sn	21 sn	52 sn	1 sn	44 sn	33 sn	53 sn	2 sn	43 sn
D9 (180, 49)	GA-SVM	11 d	12 d	13 d	15 d	16 d	17 d	19 d	20 d	21 d	22 d
		5 sn	20 sn	39 sn	25 sn	10 sn	55 sn	11 sn	19 sn	28 sn	30 sn
	Par. Opt.	12 d	12 d	13 d	13 d	15 d	16 d	17 d	19 d	20 d	22 d
	GA-SVM	27 sn	7 sn	23 sn	46 sn	2 sn	12 sn	36 sn	47 sn	42 sn	12 sn
	Par. Opt.	15 d	16 d	18 d	21 d	25 d	32 d	38 d	42 d	50 d	55 d
	CLONTagger	4 sn	1 sn	2 sn	10 sn	46 sn	10 sn	21 sn	55 sn	21 sn	41 sn
5q31 (258, 103)	GA-SVM	2 s	2 s	2 s	2 s	2 s	2 s	2 s	2 s	2 s	3 s
		16 d	19 d	29 d	35 d	37 d	45 d	52 d	55 d	57 d	5 d
	Par. Opt.	2 s	2 s	2 s	2 s	2 s	2 s	2 s	2 s	3 s	3 s
	GA-SVM	15 d	16 d	11 d	10 d	13 d	20 d	33 d	55 d	8 d	2 d
	Par. Opt.	5 s	6 s	7 s	7 s	9 s	9 s	10 s	10 s	11 s	11 s
	CLONTagger	50 d	32 d	4 d	51 d	7 d	37 d	18 d	59 d	30 d	55 d
TRPM8 (120, 101)	GA-SVM	10 d	9 d	9 d	8 d	8 d	8 d	9 d	9 d	9 d	8 d
		21 sn	53 sn	10 sn	55 sn	48 sn	45 sn	10 sn	19 sn	15 sn	40 sn
	Par. Opt.	10 d	9 d	8 d	8 d	9 d	9 d	9 d	9 d	8 d	8 d
	GA-SVM	2 sn	5 sn	40 sn	48 sn	8 sn		1 sn	13 sn	55 sn	45 sn
	Par. Opt.	37 d	36 d	33 d	31 d	35 d	35 d	35 d	37 d	39 d	38 d
	CLONTagger	41 sn	20 sn		19 sn	44 sn	43 sn	36 sn	15 sn	11 sn	26 sn
STEAP (120, 22)	GA-SVM	2 d	2 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d
		3 sn	8 sn	55 sn	32 sn	48 sn	43 sn	35 sn	33 sn	28 sn	25 sn
	Par. Opt.	2 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d	1 d
	GA-SVM	7 sn	58 sn	49 sn	42 sn	38 sn	38 sn	25 sn	23 sn	24 sn	15 sn
	Par. Opt.	8 d	7 d	6 d	6 d	6 d	6 d	6 d	5 d	5 d	5 d
	CLONTagger	21 sn	4 sn	55 sn	7 sn	13 sn	43 sn	2 sn	41 sn	25 sn	6 sn
ENm013 (120, 361)	GA-SVM	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s
		7 d	12 d	16 d	19 d	23 d	25 d	29 d	32 d	36 d	41 d
	Par. Opt.	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s	1 s
	GA-SVM	3 d	6 d	8 d	11 d	10 d	14 d	18 d	24 d	28 d	36 d
	Par. Opt.	2 s	3 s	3 s	4 s	4 s	5 s	5 s	6 s	6 s	7 s
	CLONTagger	37 d	19 d	47 d	24 d	38 d	31 d	39 d	15 d	36 d	27 d

ENr112 (120, 412)	GA-SVM	1 s 17 d	1 s 25 d	1 s 36 d	1 s 43 d	1 s 58 d	2 s 13 d	2 s 25 d	2 s 35 d	2 s 47 d	2 s 58 d
	Par. Opt. GA-SVM	1 s 14 d	1 s 23 d	1 s 29 d	1 s 41 d	1 s 49 d	2 s 3 d	2 s 16 d	2 s 25 d	2 s 36 d	2 s 48 d
	Par. Opt. CLONTagger	3 s 16 d	4 s 11 d	5 s 25 d	6 s 15 d	7 s 30 d	8 s 43 d	10 s 1 d	11 s 22 d	13 s 10 d	14 s 22 d
ENr113 (120, 515)	GA-SVM	1 s 50 d	1 s 58 d	2 s 10 d	2 s 28 d	2 s 42 d	2 s 51 d	2 s 57 d	3 s 6 d	3 s 11 d	3 s 20 d
	Par. Opt. GA-SVM	1 s 43 d	1 s 48 d	1 s 58 d	2 s 8 d	2 s 22 d	2 s 40 d	2 s 48 d	2 s 58 d	3 s 10 d	3 s 21 d
	Par. Opt. CLONTagger	5 s 35 d	6 s 17 d	7 s 2 d	7 s 56 d	8 s 37 d	10 s 14 d	12 s 3 d	13 s 35 d	14 s 53 d	16 s 12 d

Hap: Haplotip sayısı, SNP: SNP sayısı, s: saat, d: dakika, sn: saniye

7. SONUÇLAR VE ÖNERİLER

7.1. Sonuçlar

Genetik hastalıklar, bir bireyin genetik materyali yani genomundaki bozukluklar sonucu ortaya çıkan hastalıklardır. Bu hastalıklar içerisinde en yaygın olarak karşılaşılanlar karmaşık hastalıklardır ve bu hastalıklarla ilişkili genetik değişimlerin araştırılması insan genomu üzerindeki güncel araştırma konularından bir tanesidir. Birçok genom çaplı ilişki çalışması ile karmaşık hastalıklarla ilişkili olabilecek genetik değişimler belirlenmeye çalışılmaktadır. Bu genetik değişimlerin büyük çoğunluğunu SNP'ler oluşturduğu için bu çalışmalarda öncelikli olarak kullanılmaktadır (Crawford ve Nickerson, 2005; Halldorsson ve ark., 2004a). Bir genom çaplı ilişki çalışmasının istatistiksel önemi, bireylerin ve SNP'lerin sayısı ile doğrudan ilgilidir. Ancak çok büyük çaplı ilişki çalışmalarında, çok sayıdaki bireyler için aday bölge içindeki bütün SNP'leri genotiplemek hala oldukça maliyetli ve zaman alıcıdır. Bu nedenle küçük bir hata ile SNP'lerin geriye kalanını temsil edecek bütün SNP'lerin uygun bir alt kümesi olan etiket SNP'lerin seçilmesi gerekir. Etiket SNP seçiminde, çok iyi bir tahmin doğruluğuna sahip minimum büyüklükteki etiket SNP kümesinin bulunması esastır (Halperin ve ark., 2005).

Etiket SNP seçim problemi üzerine yapılan araştırmalarda, etiketlenen SNP tahmin tabanlı yaklaşımların diğerlerine göre ön blok bölme işleminin gerekmemesi, fenotip bilgisine ihtiyaç duyulmaması ve çoklu SNP ilişkilerinin kullanılabilmesi gibi avantajlara sahip olduğu görülmektedir. Ancak etiketlenen SNP tahmin tabanlı yöntemlerde, genellikle sayısal karmaşıklığı yüksek olan dinamik programlama prosedürleri kullanılmasına rağmen tahmin performanslarının geliştirilmesi gerekmektedir. Bu nedenle bu tez çalışmasında etiket SNP seçim problemi için yapay zekâ teknikleri kullanılarak üç farklı yöntem geliştirilmiştir.

Bu metotlardan ilkinde, etiket SNP seçim yöntemi olarak Genetik Algoritma (Genetic Algorithm - GA) ve geriye kalan SNP'lerin (etiket SNP haricindeki SNP'ler) tahmini için ise Destek Vektör Makinesi (Support Vector Machine - SVM) kullanılmış ve GA-SVM olarak adlandırılmıştır. GA-SVM metodunun tahmin doğruluğunun değerlendirilmesi için Birini Dışarıda Bırak Çapraz Doğrulama (Leave One-Out Cross Validation - LOOCV) yöntemi kullanılmıştır. Bu yöntemde, çaprazlama ve mutasyon işlemlerinden sonra kromozomlardaki değişen etiket SNP'lerin sayısının algoritmaya

giriş olarak verilen değere eşitlenmesi için bir ayarlama prosedürü kullanılmıştır. Buradaki amaç, aday kromozomlar içerisinde en iyi uygunluk değerine sahip olan kromozomu seçmektir. Aday kromozomların uygunluk değerlerinin hesaplanması işleminde ise 10 kat çapraz doğrulama yöntemi kullanılmıştır. Böylece, bu prosedürde LOOCV yöntemini kullanan metodlarla yaklaşık aynı sonuçlar elde edilirken daha hızlı çalışma sağlanmıştır.

GA-SVM yönteminin performansını değerlendirmek için farklı büyüklük ve özelliklere sahip 10 adet veri kümesi üzerinde çeşitli deneyler yapılmıştır. Örneğin, 22 haplotip ve 52 SNP'ten oluşan ACE veri kümesi için GA-SVM yöntemi, 1 ile 10 etiket SNP aralığında diğer yöntemler içerisinde en iyi performansa sahip BNTagger metoduna göre ortalamada %2.2 oranında daha yüksek bir tahmin doğruluğu sergilemiştir. 258 haplotip ve 103 SNP içeren 5q31 çocuk popülasyonu veri kümesi için GA-SVM yöntemi, 2 ile 10 etiket SNP aralığında karşılaştırılan metodlardan en iyi tahmin doğruluğuna sahip PSO-SVM yöntemine göre ortalamada %0.6 oranında daha yüksek bir tahmin doğruluğu üretmiştir. Kullanılan en büyük veri kümesi olan ve 120 haplotip ve 515 SNP içeren ENr113 veri kümesi için GA-SVM yöntemi, 2 ile 10 etiket SNP aralığında STAMPA metoduna göre ortalamada %5.4 oranında daha yüksek bir tahmin doğruluğu sergilemiştir.

Geliştirilen ikinci metotta ise GA-SVM metodunda sabit değerler olarak kullanılan SVM'nin C ve γ parametrelerinin optimizasyonu için Parçacık Sürü Optimizasyon algoritması kullanılmış ve bu metot parametre optimizasyonlu GA-SVM yöntemi olarak adlandırılmıştır. GA-SVM yönteminde olduğu gibi parametre optimizasyonlu GA-SVM yönteminde de etiket SNP seçim yöntemi olarak Genetik Algoritma ve geriye kalan SNP'lerin tahmini için ise SVM kullanılmıştır. Bu metodun tahmin doğruluğunun değerlendirilmesi LOOCV yöntemi ile yapılmıştır. Bu yöntemde de, çaprazlama ve mutasyon işlemlerinden sonra bir ayarlama prosedürü kullanılmış ve aday kromozomların uygunluk değerleri yine 10 kat çapraz doğrulama yöntemi ile hesaplanmıştır.

Parametre optimizasyonlu GA-SVM yönteminin performansını değerlendirmek için GA-SVM yönteminde kullanılan aynı veri kümeleri üzerinde çeşitli deneyler yapılmıştır. Örneğin ACE veri kümesi için parametre optimizasyonlu GA-SVM yöntemi, 1 ile 10 etiket SNP aralığında diğer yöntemler içerisinde en iyi performansa sahip BNTagger metoduna göre ortalamada %4.39 oranında daha yüksek tahmin doğruluğu sergilemiştir. 5q31 veri kümesi için parametre optimizasyonlu GA-SVM

yöntemi, 2 ile 10 etiket SNP aralığında karşılaştırılan metodlardan en iyi tahmin doğruluğuna sahip PSO-SVM yöntemine göre ortalamada %1.1 oranında daha yüksek tahmin doğruluğu üretmiştir. Kullanılan en büyük veri kümesi olan ENr113 veri kümesi için parametre optimizasyonlu GA-SVM yöntemi, 2 ile 10 etiket SNP aralığında STAMPA metoduna göre ortalamada %5.9 oranında daha yüksek tahmin doğruluğu sergilemiştir.

Bu tez çalışmasında geliştirilen diğer bir yöntemde ise parametre optimizasyonlu GA-SVM yöntemindeki etiket SNP seçim metodu olarak kullanılan GA yerine Klonal Seçim Algoritması (Clonal Selection Algorithm - CLONALG) kullanılmış ve Parametre Optimizasyonlu CLONTagger yöntemi olarak adlandırılmıştır. Yine bu yöntemde de geriye kalan SNP'lerin tahmini için ise SVM ve SVM'nin C ve γ parametrelerinin optimizasyonu için de PSO kullanılmıştır. Yine bu metodunun tahmin doğruluğunun değerlendirilmesi LOOCV yöntemi ile yapılmıştır. Bu yöntemde farklı bir yaklaşım olarak, performans artımına katkıda bulunmak amacıyla antikorun duyarlılığını temel alan bir mutasyon mekanizması kullanılmıştır. Yani, duyarlılık değeri yüksek olan antikorların daha fazla sayıda biti, duyarlılık değeri düşük olan antikorların ise daha az sayıda biti üzerinde mutasyon işlemi uygulanmıştır. Bu yaklaşım, düşük duyarlılıklı hücrelerin hemen çevresindeki arama alanının araştırılmasına izin vermiştir. Diğer taraftan, daha yüksek duyarlılıklı hücreler için daha yüksek mutasyon oranı, daha az verimli alanlardan uzaklaşmayı kolaylaştırdığı için arama uzayında daha büyük atlamalara izin vermiştir. Parametre optimizasyonlu CLONTagger yönteminde, ayarlama prosedürü mutasyon işleminden sonra kullanılmış ve bu prosedürde aday kromozomların uygunluk değerlerinin hesaplanmasında 10 kat çapraz doğrulama yöntemi kullanılmıştır.

Parametre optimizasyonlu CLONTagger yönteminin performansını değerlendirmek için diğer iki yöntemde kullanılan veri kümeleri kullanılmıştır. ACE veri kümesi için bu yöntem, 1 ile 10 etiket SNP aralığında diğer yöntemler içerisinde en iyi performansa sahip BNTagger metoduna göre ortalamada %4.4 oranında daha yüksek bir tahmin doğruluğu sergilemiştir. 5q31 veri kümesi için parametre optimizasyonlu CLONTagger yöntemi, 2 ile 10 etiket SNP aralığında karşılaştırılan metodlardan en iyi tahmin doğruluğuna sahip PSO-SVM yöntemine göre ortalamada %1.3 oranında daha yüksek bir tahmin doğruluğu üretmiştir. Kullanılan en büyük veri kümesi olan ENr113 veri kümesi için parametre optimizasyonlu CLONTagger yöntemi, 2 ile 10 etiket SNP

aralığında STAMPA metoduna göre ortalamada %6.3 oranında daha yüksek bir tahmin doğruluğu sergilemiştir.

Geliştirilen her üç yöntemde de kullanılan ayarlama prosedürü içinde lokal arama algoritması kullanılmıştır. Böylelikle farklı etiket SNP sayılarında farklı veri kümeleri için yapılan deneyler sonucunda diğer yöntemlere göre daha yüksek tahmin doğrulukları elde edilmiştir. Ayrıca geliştirilen bu üç yöntem ile rastgele arama algoritması kullanılarak da çeşitli deneyler yapılmıştır. Yapılan bu deneyler neticesinde istenilen sayıdaki etiket SNP'ler belirlenirken her üç yöntemin de çalışma süreleri azalırken tahmin doğruluklarında ise belirli miktarlarda azalma gözlemlenmiştir.

Bu üç yöntemi kendi aralarında kıyasladığımızda, parametre optimizasyonlu CLONTagger yöntemi farklı etiket SNP sayılarında kullanılan bütün veri kümeleri için diğer yöntemlere göre daha yüksek oranda tahmin doğruluğu sergilemiştir. Çalışma sürelerine baktığımızda ise diğer iki yönteme göre daha düşük tahmin doğrulukları sergileyen GA-SVM yöntemi daha hızlı çalışmaktadır.

7.2. Öneriler

Etiket SNP seçiminin uygulanabilirliği deneysel olarak simülasyon çalışmaları ile gösterilmiştir (Burkett ve ark., 2005; Goldstein ve ark., 2003; Ke ve ark., 2004; Ke ve ark., 2005; Zhang ve ark., 2002a). Sonuçlar etiket SNP seçiminin genotipleme çabalarında, 2 ile 5 kat arasında tasarruf sağladığını göstermiştir. Ayrıca, Zhang ve ark. (2002a) 1000 simüle edilmiş veri kümesi üzerinde, SNP'lerin bütün kümesi kullanılarak yapılan ilişki çalışmaları ile orijinal SNP kümesinin ¼ oranındaki etiket SNP kümesi kullanılarak yapılan çalışmalar arasındaki farkın ortalama %4 olduğunu göstermişlerdir.

Ancak hala etiket SNP seçimi konusunda birtakım eksiklikler mevcuttur. Bugüne kadar geliştirilmiş olan çoğu etiket SNP seçim algoritması nadir haplotip veya SNP'lerden ziyade yaygın olanları dikkate almıştır (Zhao ve ark., 2003). Birçok yaygın hastalık nadir DNA değişimlerinden daha ziyade yaygın olanlar ile açıklandığı için çalışmaların çoğu yaygın değişimler üzerine yapılmıştır (Doris, 2002; Ke ve ark., 2005; Pritchard ve Cox, 2002). Ayrıca nadir haplotipleri tanımlamak için çok daha fazla örnek veri kümesine ihtiyaç duyulur (Kruglyak ve Nickerson, 2001). Ancak, yaygın ve karmaşık hastalıklara nadir veya yaygın değişimlerin etki edip etmediği hala üzerinde çalışılması gereken bir konudur.

Algoritmaların çoğu genotip verisinden daha ziyade haplotip verisini kullanmıştır. Sadece genotip verisi mevcut iken haplotip fazlama işlemi genotip verisine uygulanır ve elde edilen haplotip verisi kullanılır. Ancak, haplotip fazlama işlemi yanlış çözümler üretebilir. Bu problemi çözmek için bazı istatistiksel algoritmalar yardımıyla çoklu çözümler üretilmiştir (Lu ve ark., 2003; Zhao ve ark., 2003). Araştırmacılar tarafından bu belirsizliği dikkate alan çalışmaların yapılması gerekmektedir.

Etiket SNP seçimi konusunda geliştirilen algoritmaların tamamı verilen bir örnek üzerinden seçilen etiket SNP'lerin aynı popülasyondan diğer örnekler için de iyi çalışacağını kabul etmişlerdir. Ancak genel bir performans değerlendirmesi yapılabilmesi için yeterli sayıda bireyin örneklenmesi gerekir. Goldstein ve ark. (2003) SNP sayısı 20 civarında iken en az 200 haplotipe ihtiyaç duyulduğunu göstermişlerdir. Bu nedenle, etiket SNP seçim işlemini örneklenmiş olan yeterli sayıda bireye uygulayan çalışmaların yapılmasına dikkat edilmelidir.

Etiket SNP seçim yaklaşımlarının eksik ve üstün olan yönlerini anlamak için birçok karşılaştırma çalışması yapılmıştır. Burkett ve ark. (2005) 5 metodu (2 haplotip farklılığı tabanlı, 2 ilişki katsayısı tabanlı ve 1 eşit aralıklı) karşılaştırmış ve aynı yaklaşımı kullanan iki farklı metodun bile farklı sayıda etiket SNP üretebildiğini ve ortak olarak seçilen etiket SNP'lerin oranının şaşırtıcı bir şekilde sadece %30 civarında olduğunu göstermişlerdir. Duggal ve ark. (2005) ise aksine 6 haplotip farklılığı tabanlı metodlar üzerinde yaptıkları deneylerde ortak olarak seçilen etiket SNP'lerin oranının yaklaşık %50 ile %95 arasında olduğunu göstermişlerdir. İlaveten Ke ve ark. (2004, 2005), haplotip farklılığı ve ilişki katsayısı tabanlı metotlar için seçilen etiket SNP'lerin tahmin yeteneğinin oldukça uyumlu olduğunu göstermişlerdir. Farklı veri kümelerine ve hatta farklı değerlendirme ölçülerine dayanan bu sonuçların etiket SNP seçim problemiyle ilgili çalışmalar yapılırken dikkate alınması gerekir.

Etiket SNP seçim problemi aslında, etiket SNP seçimi ve geri kalan SNP'lerin (etiket SNP haricindeki) tahmini olmak üzere iki aşamadan oluşmaktadır. Bu konuyla ilgili çalışmalar yapılırken her iki aşama içinde farklı algoritmaların kullanılmasına ve geliştirilmesine özellikle özen gösterilmelidir. Bu tez çalışmasında, geliştirilen üç yöntem tarafından elde edilen sonuçların farklılığı bunu doğrulamaktadır. Parametre optimizasyonlu GA-SVM yönteminde önceki GA-SVM yöntemindeki etiket SNP seçim yöntemi aynı kalmış ve geri kalan SNP'lerin tahmini için kullanılan yönteme parametre optimizasyonu eklenerek önceki yönteme göre daha yüksek tahmin doğrulukları elde edilmiştir. Parametre optimizasyonlu CLONTagger yönteminde ise parametre

optimizasyonlu GA-SVM yöntemindeki geri kalan SNP'lerin tahmini için kullanılan yöntem aynı kalmış ve etiket SNP seçim yöntemi değiştirilerek bir önceki yönteme göre daha yüksek tahmin doğrulukları elde edilmiştir. Geliştilen bu üç yöntemin tahmin doğruluklarına bakıldığında geri kalan SNP'lerin tahmini için kullanılan yöntemin geliştirilmesinin diğerine göre daha önemli olduğu gözlemlenmiştir. Bu konuda çalışmak isteyen araştırmacıların bu noktaya özellikle dikkat etmeleri gerekmektedir.

Yapılan araştırmalara göre etiket SNP seçimi konusunda Türkiye'de bir ilk olan bu tez çalışmasının, genom çaplı ilişki çalışmalarında kullanılacağı ve lisansüstü öğrencilerinin konuyla ilgili araştırmalarına bir ışık olacağı düşünülmektedir.

Genetik alanındaki gelişmeleri takiben, insan genom dizisinin karakterizasyonuyla başlayan süreçte, yaşam bilimleri alanında daha önce benzeri görülmemiş oranda bir veri birikimi başlamıştır. Teknolojik gelişmelerin de etkisiyle bu bilgi birikimi her geçen gün katlanarak artmaktadır. Biyoinformatik bilim dalı, bu biyolojik verilerin anlamlandırılması, saklanması, görsellenmesi ve bu devasa bilgi birikiminden azami ölçüde yararlanabilmek amacıyla, matematik, istatistik, bilgisayar bilimleri, moleküler biyoloji ve genetik alanlarının sentezi olarak doğmuştur. Birçok gelişmiş ülkede, bu dalın lisans ve lisansüstü seviyede eğitimi verilmekte olup, ülkemizde sadece bir üniversitede Genetik ve Biyonformatik adı altında lisans düzeyinde okutulmaktadır. Önümüzdeki dönemde, ülkemizde, Biyoinformatik bilim dalı ile ilgili hem lisans hem de lisansüstü seviyede eğitim veren üniversitelerin sayısında bir artış olacağına inanılmaktadır.

KAYNAKLAR

Ackerman, H., Usen, S., Mott, R., Richardson, A., Sisay-Joof, F., Katundu, P., Taylor, T., Ward, R., Molyneux, M., Pinder, M., Kwiatkowski, D.P., 2003, Haplotypic analysis of the TNF locus by association efficiency and entropy, *Genome Biology*, 4(4), 1-13.

Akey, J., Jin, L., Xiong, M., 2001, Haplotypes vs single marker linkage disequilibrium tests: What do we gain?, *European Journal of Human Genetics*, 9, 291-300.

Angeline, P. J., 1995, Evolution revolution: An introduction to the special track on genetic and evolutionary programming, *IEEE Expert Intelligent Systems and their Applications*, 10(3), 6-10.

Ao, S. I., Yip, K., Ng, M., Cheung, D., Fong, P., Melhado, I., Sham, P. C., 2005, CLUSTAG: Hierarchical clustering and graph methods for selecting tag SNPs, *Bioinformatics*, 21, 1735-1736.

Arlot, S. ve Celisse, A., 2010, A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, 40-79.

Avi-Itzhak, H. I., Su, X., De La Vega, F. M., 2003, Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity, *Proceeding of Pac. Symp. Biocomputing*, 466-477.

Bafna, V., Halldorsson, B. V., Schwartz, R., Clark, A. G., Istrail, S., 2003, Haplotypes and informative SNP selection algorithms: Don't block out information, *In Proceedings of the 7th International Conference on Computational Molecular Biology*, 19-26.

Burkett, K. M., Ghadessi, M., McNeney, B., Graham, J., Daley, D, 2005, A comparision of five methods for selecting tagging single-nucleotide polymorphisms, *BMC Genetics*, 6(1), 71-75.

Byng, M. C., Whittaker, J. C., Cuthbert, A., Mathew, C. G., Lewis, C. M., 2003, SNP subset selection for genetic association studies, *Annals of Human Genetics*, 67, 543-556.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., Nickerson, D. A., 2004, Selecting a maximally informative set of single nucleotide polymorphisms for association analyses using linkage disequilirium, *American Journal of Human Genetics*, 74, 106-120.

Chanda, P., Zhang, A. ve Ramanathan, M., 2009, Machine learning in bioinformatics, Zhang, Y. and Rajapakse, J., *John Wiley & Sons*, New Jersey - Canada, 389-412.

Chang, C. C. ve Lin, C. J., 2011, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2(3), 1-27.

Clark, A. G., Weiss, K., Nickerson, D., Taylor, S., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., Sing, C. F., 1998, Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase, *Am. J. Hum. Genet*., 63, 595-612.

Claverie, J. ve Notredame, C., 2007, Bioinformatics for dummies, *Wiley Publishing*, Indiana - USA, 17-21.

Cores, C. ve Vapnik, V. N., 1995, Support Vector Networks, *Machine Learning*, 20(3), 273-297.

Crawford, D. ve Nickerson, D. A., 2005, Definition and clinical importance of haplotypes, *Annu Rev Med.*, 56, 303-320.

Çomak, E., 2008, Destek vektör makinelerinin etkin eğitimi için yeni yaklaşımlar, Doktora Tezi, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*, Konya, 35-57.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., Lander, E. S., 2001, High-resolution haplotype strucutre in the human genome, *Nature Genetics*, 29(2), 229-232.

De Bakker, P.I.W., Graham, R. R., Altshuler, D., Henderson, B.E., Haiman, C.A., 2006, Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple population, *In Proceedings of Pacific Symposium on Biocomputing*, 35.

De Castro, L. N. ve Von Zuben, F. J., 1999, Artificial immune systems: Part I- Basic theory and applications, Technical Report - DCA-RT 02/00.

Ding, K., Zhang, J., Zhou, K., Shen, Y., Zhang, X., 2005a, htSNPer1.0: Software for haplotype block partition and htSNPs selection, *BMC Bioinformatics*, 6(38),1-7.

Ding, K., Zhou, K., Zhang, J., Knight, J., Zhang, X., Shen, Y., 2005b, The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection, *Molecular Biology and Evolution*, 22(1), 148-159.

Doris, P. A., 2002, Hypertension genetics, single nucleotide polymorphisms, and the common disease: Common variant hypothesis, *Hypertension*, 39, 323-331.

Duggal, P., Gillanders, E. M., Mathias, R. A., Ibay, G. P., Klein, A. P., Baffoe-Bonnie, A. B., Ou, L., Dusenberry, I. P., Tsai, Y. Y., Chines, P. S., Doan, B. Q., Bailey-Wilson, J. E., 2005, Identification of tag single-nucleotide polymorphisms in regions with varying linkage disequilibrium, *BMC Genetics*, 6(1), S73.

Efron, B., 1982, The jackknife, the bootstrap, and other resampling plans, *Society for Industrial and Applied Mathematics*. 27-35.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, F., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C.,

Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M., Altshuler, D., 2002, The structure of haplotype blocks in the human genome, *Science,* 296, 2225-2229.

Goldberg, D. E., 1989, Genetic algorithms in search, optimization, and machine learning, Addison Wesley, New York.

Goldberg, D. E. ve Deb, K., 1991, A comparative analysis of selection schemes used in genetic algorithms, *Foundations of Genetic Algorithms*, 1, 69-93.

Goldstein, D.B., 2001, Islands of linkage disequilibrium, *Nature Genetics*, 29, 109-211.

Goldstein, D. B., Ahmadi, K. R., Weale, M. E., Wood, N. W., 2003, Genome scans and candidate gene approaches in the study of common diseases and variable drug responses, *Trends in Genetics*, 19(11), 615-622.

Greenspan, G. ve Geiger, D., 2003, Model-based inference of haplotype block variation, *In Proceeding of Intl Conf Res Comp Mol Biol*ogy, 131-137.

Guo, Y., Cao, X., Yin, H., Tang, Z., 2007, Coevolutionary optimization algorithm with dynamic sub-population size, *International Journal Of Innovative Computing, Information And Control*, 3(2), 435-448.

Halldorsson, B. V., Bafna, V., Edwards, N., Lippert, R., Yooseph, S., Istrail, S., 2004a, A survey of computational methods for determininghaplotyes, *Lecture Notes in Computer Science*, 2983, 26-47.

Halldorsson, B. V., Bafna, V., Lippert, R., Schwatz, R., De La Vega, F. M., Clark, A. G., Istrail, S., 2004b, Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies, *Genome Research*, 14, 1633-1640.

Hampe, J., Schreiber, S., Krawczak, M., 2003, Entropy-based SNP selection for genetic association studies, *Human Genetics*, 114, 36-43.

Halperin, E., Kimmel, G., Shamir, R., 2005, Tag SNP selection in genotype data for maximizing SNP prediction accuracy, *Bioinformatics*, 21, 195-203.

He, J., ve Zelikovsky, A., 2006, MLR-tagging: Informative SNP selection for unphased genotypes based on multiple linear regression, *Bioinformatics*, 22, 2558-2561.

He, J. ve Zelikovsky, A., 2007, Informative SNP selection methods based on SNP prediction, *IEEE Trans. Nanobioscience*, 6(1), 60-67.

Hedrick, P.W., 2004, Genetics of pouplation, *Jones and Bartlett Publishers*, 3rd Edition.

Hoh, J., Wille, A., Zee, R., Cheng, S., Reynolds, R., Lindpaintner, K., Ott, J., 2000, Selecting SNPs in two-stage analysis of disease association data: A model-free approach, *Annual Human Genetics*, 64, 413-417.

Horne, B. ve Camp, N. J, 2004, Principal component analysis for selection of optimal SNP sets that capture intragenic genetic variation, *Genetic Epidemiology*, 26, 11-21.

Hsu, C. W., Chang, C. C., Lin, C. J., 2010, A practical guide to support vector classification, *Bioinformatics*, 11, 1-16.

International HapMap Consortium, 2003. The international HapMap project, *Nature,* 426(6968), 789-796.

İlhan, İ., Göktepe, Y. E., Kahramanlı, Ş., Özcan, C., 2011a, Tag SNP selection using clonal selection algorithm based on support vector machine, *3rd International Conference on Future Computer and Communication*, Iasi, Romania, 47-57.

İlhan, İ., Göktepe, Y. E., Kahramanlı, Ş., 2011b, Tag SNP selection using GA-SVM Approach, *IADIS European Conference on Data Mining 2011*, Rome, Italy, 27-34.

Johnson, G. C, Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., Todd, J.A., 2001, Haplotype tagging for the identification of common disease genes, *Nat Genet*, 29, 233-237.

Judson, R., Salisbury, B., Schneider, J., 2002, How many SNPs does a genome-wide haplotype map require?, *Parmacogenomics*, 3, 379-391.

Ke, X., Cardon, L. R., 2003, Efficient selective screening of haplotype tag SNPs, *Bioinformatics*, 19, 287-288.

Ke, X., Durrant, C., Morris, A. P., Hunt, S., Bentley, D. R., Deloukas, P., Cardon, L. R., 2004, Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples, *Human Molecular Genetics*, 13(21), 2557-2565.

Ke, X., Miretti, M. M., Broxholme, J., Hunt, S., Beck, S., Bentley, D. R., Deloukas, P., Cardon, L. R., 2005, A comparison of tagging methods and their tagging space, *Human Molecular Genetics*, 14(18), 2757-2767.

Kennedy, J. ve Eberhart, R., 1995, Particle swarm optimization, *IEEE International Conference on Neural Networks*, Path, Australia, 4, 1942-1948.

Kodaz, H., 2007, Bilgi kazancı tabanlı yapay bağışıklık tanıma sistemi, Doktora Tezi, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*, Konya, 30-40.

Kroetz, D. L., Pauli-Magnus, C., Hodges, L. M., Huang, C. C., Kawamoto, M., Johns, S. J., Stryke, D., Ferrin, T. E., DeYoung, J., Taylor, T., Carlson, E. J., Herskowitz, I., Giacomini, K. M., Clark, A. G., 2003, Sequence diversity and haplotype structure in the human ABCB1 (MDR1, multidrug resistance transporter) gene, *Pharmacogenetics*, 13, 481-494.

Kruglyak, L. ve Nickerson, D. A., 2001, Variation is the spice of life, *Nature Genetics*,27, 234-236.

Lee, P. H. ve Shatkay, H., 2006, BNTagger: Improved tagging SNP selection using Bayesian Networks, *Bioinformatics*, 22, 211-219.

Lee, P. H., 2009, Prioritizing SNPs for disease-gene association studies: Algorithms and systems, Doktora Tezi, *Queen's University School of Computing*, Kingston, Ontario, Canada, 10-30.

Lee, P. H. ve Shatkay, H., 2009, Machine learning in bioinformatics, Zhang, Y. and Rajapakse, J., *John Wiley & Sons*, New Jersey - Canada, 367-388.

Lin, W. Y., Lee, W. Y., Hong, T. P., 2003, Adapting crossover and mutation rates in genetic algorithms, *Journal of Informatin Science and Engineering*, 19(5), 889-903.

Lin, Z. ve Altman, R. B, 2004, Finding haplotype tagging SNPs by use of principal components analysis, *American Journal of Human Genetics*, 75, 850-861.

Lin, M. H., Leu, C. L., 2010, A hybrid PSO-SVM approach for haplotype tagging SNP selection problem, *International Journal of Computer Science and Information Security*, 8(6), 60-65.

Lu, X., Niu, T., Liu, J. S., 2003, Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms, *Genome Research*, 13, 2112-2117.

Mahdevar, G., Zahiri, J., Sadeghi, M., Dalini, A. N., Ahrabian, H., 2010, Tag SNP selection via a genetic algorithm, *Journal of Biomedical Informatics*, 43(5), 800-804.

Meng, Z., Zaykin, D. V., Xu, C., Wagner, M., Ehm, M. G., 2003, Selection of genetic mark-ers for association analyses using linkage disequilibrium and haplotypes, *American Journal of Human Genetics*, 73, 115-130.

Nothnagel, M., 2004, The definition of multilocus haplotype blocks and common diseases, PhD thesis, *University of Berlin*.

Özsağlam, M. Y. ve Çunkaş, M., 2008, Optimizasyon problemlerinin çözümü için parçacık sürü optimizasyonu algoritması, *Politeknik Dergisi*, 11(4), 299-305.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C., Lee, D., Marjoribanks, C., McDomough, D., Nguyen, B., Norris, M., Sheehan, B., Shen, N., Stern, N., Stokowski, R., Thomas, D., Trulson, M., Vyas, K., Frazer, K., Fodor, S., Cox, D., 2001, Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, *Science*, 294, 1719-1723.

Pritchard, J. K. ve Cox, N. J., 2002, The allelic architecture of human disease genes: Common disease-common variant... or not?, *Human Molecular Genetics*, 11(20), 2417-2423.

Prügel-Bennett, A., 2001, The mixing rate of different crossover operators, *Foundations of Genetic Algorithms*, 6, 261-274.

Rieder, M., Taylor, S.L., Clark, A.G., Nickerson, D.A., 1999, Sequence variation in the human angiotensin converting enzyme, *Nature Genetics*, 22, 59-62.

Sağ, T. ve Çunkaş, M., 2009, A tool for multiobjective evoluationary algorithms, *Advances in Engineering Software*, 40(9), 902-912.

Schulze, T.G., Zhang, K., Chen, Y., Akula, N., Sun, F., McMahon, F.J., 2004, Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome, *Human Molecular Genetics*, 13(3), 335-342.

Shah, S. C. ve Kusiak, A., 2004, Data mining and genetic algorithm based gene/SNP selection, *Artifical Intelligence in Medicine*, 31, 183-196.

Sheng, H., Zhang, P., Uehara, R., 2004, A double classification tree search algorithm for index SNP selection, *BMC Bioinformatics*, 5(89),1-6.

Song, Y. ve Eberhart, R., 1998, A modified particle swarm optimizer, *1998 IEEE World Congress on Computational Intelligence*, 63-67.

Sywerda, G., 1989, Uniform crossover in genetic algorithms, *Proceedings of the 3rd International Conference on Genetic Algorithms*, Los Altos, CA, 2-9.

Tekşen, F., 2006, Tıbbi biyoloji ve genetik ders kitabı, *Ankara Üniversitesi Basımevi*, Ankara, 29-34.

Tsalenko, A., Ben-Dor, A., Cox, N., ve Yakhini, Z., 2003, Methods for analysis and visualization of SNP genotype data for complex diseases, *In Proceedings of Pacific Symposium on Biocomputing*, 548-561.

Vapnik, V., 1998, Statistical learning theory. Wiley, New York.

Weale, M. E., Depondt, C., Macdonald, S. J., Smith, A., Lai, P. S., Shorvon, S. D., Wood, N. W., Goldstein, D. B., 2003, Selection and evaluation of tagging SNPs in the nueronal-sodium-channel gene scn1a: Implications for linkage-disequilibrium gene mapping, *American Journal of Human Genetics*, 73(3), 551-565.

Wu, X., Luke, A., Rieder, M., Lee, K., Toth, E. J., Nickerson, D., Zhu, X., Kan, D., Cooper, R. S., 2003, An association study of angiotensiongen polymorphisms with serum level and hypertension in an african-american population, *Journal of Hypertension*, 21(10), 1847-1852.

Xing, E. P., Sharan, R., Jordan, M. I., 2004, Bayesian haplotype inference via the Dirichlet process, *In Proceedings of the 21st International Conference on Machine Learning*, 879-886.

Yang, C. Y., Hou, C. H., Chuang, L. Y., 2008, Improved tag SNP selection using binary particle swarm optimization, *IEEE Congress on Evolutionary Computation (CEC 2008)*, 854-860.

Zhang, K., Calabrese, P., Nordborg, M., Sun, F., 2002a, Haplotype block structure and its application to association studies: Power and study designs, *American Journal of Human Genetics*, 71, 1386-1394.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., Sun, F., 2002b, A dynamic programming algorithm for haplotype block partitioning, *Proceedings of the National Academy of Sciences of the United States of America 99*, 7335-7339.

Zhang, K., Sun, F., Waterman, M. S., Chen, T., 2003, Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data, *American Journal of Human Genetics*, 73, 63-73.

Zhang, K., Qin, Z., Liu, J., Chen, T., Waterman, M., Sun, F., 2004, Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies, *Genome Research*, 14, 908-916.

Zhang, K., Qin, Z., Chen, T., Liu, J. S., Waterman, M. S., Sun, F., 2005, Hapblock: haplotype block partitioning and tag SNP selection software using a set of dynamic progr-mming algorithms, *Bioinformatics*, 21(1), 131-134.

Zhao, H., Pfeiffer, R., Gail, M. H., 2003, Haplotype analysis in population genetics and association studies, *Pharmacogenomics*, 4(2), 171-178.

Zou, S., Huang, Y., Wang, Y., Wang, J., Zhou, C., 2008, SVM learning from imbalanced data by GA sampling for protein domain predicting, *The 9th International Conference for Young Computer Scientists*, 982-987.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	İlhan İLHAN
Uyruğu	**:**	T.C.
Doğum Yeri ve Tarihi	**:**	Konya – 1974
Telefon	**:**	0 332 461 28 90 / 14
Faks	**:**	0 332 461 28 92
e-mail	**:**	ilhan@selcuk.edu.tr

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	Fatih Teknik Lisesi Bilgisayar Bölümü	1992
Üniversite	:	F.Ü. Teknik Eğitim Fakültesi Bilgisayar Öğr.	1996
Yüksek Lisans	:	S.Ü. Teknik Eğitim Fakültesi Bilgisayar Sist. Eğt.	2004
Doktora	:	S.Ü. Müh. Mim. Fak. Bilgisayar Müh.	

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
1996-2001	Milli Eğitim Bakanlığı - Mersin	Bilgisayar Öğr.
2001-	Selçuk Üniversitesi Akören ARE MYO	Öğr. Gör.

UZMANLIK ALANI

Etiket SNP Seçimi, Optimizasyon ve Sınıflama Algoritmaları.

YABANCI DİLLER

İngilizce

YAYINLAR

İlhan, İ., Tezel, G., "A Genetic Algorithm - Support Vector Machine Method With Parameter Optimization For Selecting The Tag SNPs", Journal of Biomedical Informatics, doi: 10.1016/j.jbi.2012.12.002, 2012 (Doktora tezinden yapılmıştır).

İlhan, İ., Ünsaçar, F., "Otomotiv Yedek Parça Sektörüne Elektronik Veri Değişimi (EDI) Uygulanması İçin Veritabanı ve Formların Oluşturulması", Politeknik Dergisi, 14(4), 2011 (Yüksek Lisans tezinden yapılmıştır).

Göktepe, Y. E., İlhan, İ., Kahramanlı, Ş., "A Weighted Amino-Acid Composition-Based Protein-Protein Interaction Prediction Method", International Journal of Innovative Computing, Information and Control, 9(7), 2013.

Göktepe, Y. E., İlhan, İ., Elitok, E., Kahramanlı, Ş., "Increasing The Performance of PSEAAC in Predicting Protein-Protein Interactions", IADIS International Conference IADIS Data Mining 2011, İtalya, 2011.

Göktepe, Y. E., İlhan, İ., Kahramanlı, Ş., "Boosting The Performance of Pseudo Amino Acid Composition", 3rd International Conference on Future Computer and Communication (ICFCC), Romanya, 2011.