DNA Dizilimlerindeki Nükleotit Çiftlerinin Frekans Değerlerine Göre Farklı Sınıflandırma Yöntemleri ile Karşılaştırılması

Bihter Daş[1], İbrahim Türkoğlu[2]

[1]Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği, 23119, Elazığ/Turkiye.
[2]Bingöl Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 12000, Bingöl/Türkiye.

bihtedas@gmail.com iturkoglu@bingol.edu.tr

Özet

DNA dizilim hizalama, biyoinformatiğin en temel problemlerinden biridir. Son yıllarda DNA üzerine yapılan çalışmalarda, DNA'daki nükleotidlerin dizilişlerinin birbiri ardı sıra tek, ikili, üçlü olarak tekrar ettiği belirlenmiştir. Kısa ard arda tekrar eden (STR) olarak adlandırılan ve çeşitli alanlarda kullanılan bu belirteçler genetik hastalıklarda, adli amaçlı kimlik tespitinde, babalık araştırmaları ve tümör biyokimyasal araştırılmaları gibi birçok farklı amaçlar için kullanılmaktadır.

Bu makale çalışmasında, bakteri türlerinin farklı uzunluklardaki DNA dizilimleri alınarak, bu dizilimlerde tekrar eden nükleotid çiftlerin frekansı bulunmuş ve elde edilen bu frekans değerlerine YSA, DVM ve KNN yöntemleri uygulanarak bir sınıflandırılma yapılmıştır. Sınıflandırma sonucunda KNN yönteminin, YSA ve DVM yöntemlerinden başarımının daha yüksek olduğu görülmüştür.

Anahtar Kelimeler: *DNA dizilimi, özellik çıkarımı, sınıflandırma, STR*

Abstract

DNA sequence alignment is one of the major problems of bioinformatics. In recent year's studies on DNA, it is determined that sequences of nucleotide in DNA are consecutive as single, double or triple. These markers called STR are used in various fields such as genetic diseases, forensic identification, fatherhood research, biochemical tumor research.

In this paper, it was found the frequency of repetitive nucleotide pairs taking DNA sequences of different lengths of the bacterial species and a classification was made to obtained frequency values by using ANN, SVM and KNN methods. Classification results show performance of KNN method is higher than performance of SVM and ANN methods.

Keywords: *DNA sequence, feature extraction, classification, STR*

1. Giriş

Biyoinformatik, biyolojik problemlerin çözümünde bilişim teknolojilerinin kullanılması esasına dayanan bilimsel bir disiplindir. DNA'daki nükleotid diziliminden yola çıkarak hastalıklara sebep olan biyolojik verilerin derlenmesi, analiz edilmesi bu hastalıkların tedavi yöntemlerinin geliştirilmesi biyoinformatiğin en temel amacıdır.

DNA molekülü birbiri etrafında kıvrılmış iki iplikten oluşur. Her bir iplik nükleotid adını verdiğimiz (şeker+fosfat+baz) ünitelerinin tekrarlarından ibaret olup polinükleotid adını alır. DNA'nın yapısında bulunan bazlar adenin(A), timin(T), guanin(G) ve sitozin(C) olmak üzere 4 çeşittir.

Son yıllarda DNA üzerinde yapılan araştırmalarda; DNA'daki bazların ve baz dizilimlerin birbiri ardı sıra tekrar ettiği belirlenmiştir. STR olarak adlandırılan bu tekrarlar genetik kökenli haritalamada, tümör biyokimyasal araştırmalarda, adli bireysel kimlik tespitinde, babalık ve nüfus genetik analizlerde yaygın bir şekilde kullanılmaktadır [4] Çeşitli veri madenciliği tekniklerinin kullanılması DNA dizilimlerinin sınıflandırılması açısından büyük öneme sahiptir.

Bu makalede DNA dizilimlerini sınıflandırmada Yapay Sinir Ağı(YSA), Destek Vektör Makineleri(DVM) ve K-En Yakın Komşu (KNN) metotları kullanılmıştır.

2. Verilerin elde edilmesi

Bu çalışmada deneysel veriler için *National Center for Biotechnology (NCBI)* sitesi gen bankasından [5] gerçek veriler alınmıştır. Alınan Escherichia coli, Bacillus cereus, Buchnera aphidicola ve Enterobacter cloacae gibi 4 bakteri türü sınıflandırılma için kullanılmıştır. Tablo 1'de kullanılan bakteri türleri, erişim numaraları, sayıları ve ortalama uzunlukları verilmiştir.

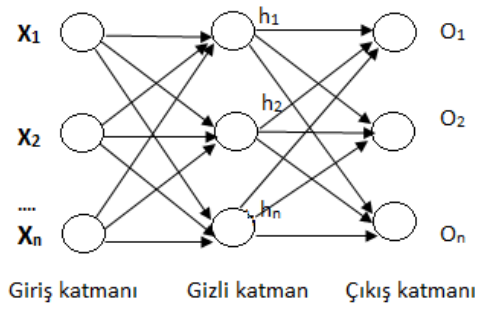
Tablo 1: Kullanılan veri türlerinin açıklaması

Sıra	Bakteri Türü	Erişim Numarası	Bakteri Örnek Sayısı	Dizilimdeki Nükleotid Sayısı
1	Escherichia coli	AE14075	31	102
2	Bacillus cereus	NC_016771	40	94
3	Buchnera aphidicola	AF012886	35	96
4	Enterobacter cloacae	EU606203	48	104

3. Kullanılan sınıflandırma yöntemleri

a. Yapay Sinir Ağları

Yapay Sinir Ağları insan beyninden esinlenerek geliştirilmiş bir bilgi işlem teknolojisidir. YSA ile basit biyolojik sinir sisteminin çalışma şekli benzetim yapılır. Yapay sinir ağları, insan beyninin özelliklerinden olan öğrenme yolu ile hafızaya alma, yeni bilgiler oluşturabilme ve keşfedebilme, veriler arasındaki ilişkileri ortaya koyma gibi yetenekleri otomatik olarak gerçekleştirmek amacı ile geliştirilmişlerdir [8]. Yapay sinir ağlarının iki türlü çalışma şekli vardır. Biri eğitim (öğrenme) diğeri test (kullanma) aşamasıdır. Bir yapay sinir ağının kullanılabilmesi için önce eğitilmesi gerekir. Bir YSA'nın birim elemanı nörondur (düğüm). Yapay sinir ağları Şekil 1'de görüldüğü üzere temel olarak girdi, gizli ve çıktı katman olmak üzere üç katmandan oluşmakta ve her katmanda birçok nöron (düğüm) bulunmaktadır.

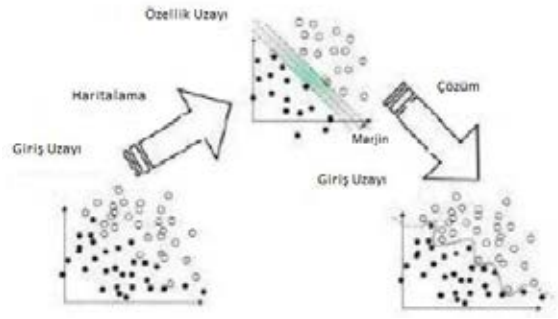


Şekil 1: Üç Katmanlı bir Yapay Sinir Ağı

Bu çalışmada kullanılan YSA modelinde giriş katmanı 16 düğümden oluşmaktadır. Gizli katmandaki düğümlerin sayısının eğitim süresince tespit edilmesi gerekmektedir. Çıkış katmanı ise bakteri türlerini temsil eden 1,2,3,4 şeklinde 4 düğümden oluşmaktadır.

b. Destek vektör makinesi

Son yıllarda, sınıflandırma problemlerinin çözümü için geliştirilmiş en başarılı makine öğrenimi algoritmalarından biri Destek Vektör Makineleri'dir. Destek Vektör Makineleri, değişkenler arasındaki örüntülerin bilinmediği veri setlerindeki birçok sınıflandırma probleminin çözümünde başarıyla uygulanmış, performansı yüksek ve etkin makine öğrenimi algoritmalarından biri olarak veri madenciliği uygulamalarındaki yerini almıştır [9]. Bu yöntem, sınıflandırmayı bir doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla yerine getirir. Doğrusal olmayan dönüşümlerde kernel fonksiyonu kullanılmakta ve verilerin daha yüksek boyutta doğrusal olarak ayrımına imkân sağlanmaktadır ve Şekil 2' de destek vektör algoritmasının genel yapısı görülmektedir. [7]

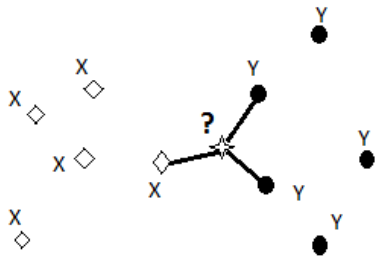


Şekil 2: Destek Vektör Makine Algoritması [7]

Bu çalışmada destek vektör makineleri (DVM) ile gerçekleştirilecek dört sınıflı bir sınıflandırma için radyal tabanlı fonksiyon (RBF) kerneli kullanılmıştır. Uygulamada kernel parametresi (RBF kerneli için band genişliği değeri) 2, düzenleme parametresi (C) ise 100000 olarak belirlenmiştir.

c. K-En yakın komşu algoritması

K-En Yakın Komşu Algoritması(K-Nearest Neighbor Algorithm), yeni bir veri geldiğinde varolan öğrenme verisi üzerinde sınıflandırma yapan eğitimli öğrenme algoritmasıdır. Algoritma, yeni bir veri geldiğinde, onun en yakın K komşusuna bakarak bu verinin sınıfına karar verir. Her sınıfın özelliklerinin önceden belirlenmiş olması çok önemlidir. Yeni gelen verinin daha önceki verilerden k tanesine yakınlığına bakılır. Bu iki veri arasındaki mesafe çeşitli uzaklık fonksiyonları kullanılarak hesaplanır. Manhattan Uzaklık Fonksiyonu, Minkowski Uzaklık Fonksiyonu, Öklid Uzaklık Fonksiyonu içerisinden en çok tercih edilen fonksiyon oklid uzaklık fonksiyonudur. En yakın mesafe neresi ise yeni veri o sınıfa atanır. Bu çalışmada kullanılan KNN algoritmasında k değeri 5 olarak alınmıştır.



Şekil 3: K-En Yakın Komşu Algoritması

4. Deneysel çıktılar

Bu çalışmada yer alan uygulama, Gen bankasından alınan Escherichia coli, Bacillus cereus, Buchnera aphidicola ve Enterobacter cloacae gibi 4 bakteri türden oluşan toplam 154 örnek üzerinde gerçekleştirilmiştir. Makalede kullanılan Yapay Sinir Ağı, çok sınıflı Destek Vektör Makinesi ve K- En Yakın Komşu sınıflandırma metotları için 154 satır 17 sütun içeren bir giriş matrisi oluşturulmuştur. Bu matrisin ilk 16 sütununu DNA dizilimde bulunan nükleotid çitlerinin frekans dağılım değerleri oluşturmaktadır. Bu türlerin DNA dizilimlerinin karakteristik matrisini oluşturmak için DNA dizilimlerinden özellik çıkarımı yapılmıştır. Bu özellik çıkarımı için Tablo 1'de verilen her bir türün ortalama uzunluğuna göre dizilimde yer alan nükleotid çiftlerin frekans dağılımı bulunmuştur. Bir DNA diziliminde yer alabilecek nükleotid çiftleri AG, AC, AT, AA, GA, GC, GT, GG, CA, CG, CT, CC, TA, TG, TC, TT şeklindedir. Uygulamada her bir örnek için bu nükleotid çiftlerinin dağılım frekansı denklem (1) 'e göre hesaplanmıştır [4].

$$f_{SiSj}=\frac{S_iS_j}{|S|-1} \quad (1)$$

Denkleme göre S_iS_j nükleotid çiftlerinin DNA diziliminde bulunma sayısı, $|S|$ ise DNA diziliminin uzunluğudur. Örneğin 'GATCAACATATTGTGACGCACGT' şeklinde verilen DNA diziliminde nükleotid çitlerinin frekanslarının denklem (1)'e göre hesaplanması tablo 2'de verilmiştir.

Tablo 2: Örnek dizilimin matris sütunları

Nükleotid Çifti	Frekans Değeri	Nükleotid Çifti	Frekans Değeri
AA	0.0526	TG	0.0434
AT	0.157	CA	0.1304
TA	0.0434	AC	0.0526
TT	0.434	TC	0.0434
GG	0	CT	0
GA	0.0869	CC	0
AG	0	GC	0.0434
GT	0.0869	CG	0.0869

Giriş matrisimizin 17. sütununu ise 4 türü temsil edecek sınıflarımız oluşturmaktadır. Bu giriş matrisi MATLAB yazılım ortamında DVM, YSA ve K-En Yakın Komşu algoritmasına göre sınıflandırılmış ve sınıflandırma sonuçları tablo (3) de gösterilmiştir.

Tablo 3: Bakteri Türleri için Algoritmaların Sınıflandırma Başarılarının Karşılaştırılması

	Örnek sayısı	YSA (%)	DVM(%)	KNN (%)
Escherichia coli	31	45.3	85.85	92.5
Bacillus cereus	40	66.6	96.75	98.48
Buchnera aphidicola	35	97.1	92.72	97.4
Enterobacter cloacae	48	86.56	87.77	96.02
Toplam	154	73.89	90.77	96.1

Tablo 3'e bakıldığında KNN sınıflandırma metodunun başarısının toplamda %96.1 oranıyla diğer metotlardan daha başarılı bir sınıflandırma yaptığı görülmüştür. Buna rağmen YSA metodunun ise DNA dizilimlerini sınıflandırmada başarısının toplamda %73.89 oranıyla düşük olduğu gözlemlenmiştir. Uygulamalarda örnek sayısı arttıkça ve örnek verilerinin DNA dizilim uzunluğu fazla alındıkça algoritmaların sınıflandırma performansının arttığı gözlemlenmiştir.

5. Sonuç

DNA STR'lerin tekrarlanan ünite sayılarındaki varyasyonları birçok genetik çalışmalarda belirteç olarak kullanılmaktadır. [1] Çeşitli alanlarda kullanılan STR belirteçleri genetik hastalıklarda, adli amaçlı kimlik tespitinde, babalık araştırmaları ve tümör biyokimyasal araştırılmaları gibi birçok farklı amaçlar için kullanılmaktadır.

Bu makale çalışmasında, bakteri türlerinin farklı uzunluklardaki DNA dizilimleri alınarak, bu dizilimlerde tekrar eden nükleotid çiftlerin frekansı bulunmuş ve elde edilen bu frekans değerlerine YSA, DVM ve KNN yöntemleri uygulanarak bir sınıflandırılma yapılmıştır. Elde edilen sınıflandırma sonucunda KNN yönteminin, YSA ve DVM yöntemlerinden başarımının daha yüksek olduğu görülmüştür.

6. Kaynaklar

[1] Dönbak L. "Kısa ardarda tekrar eden dna dizilerinin adli amaçlı dna çalışmalarındaki yeri " *Türkiye Klinikleri Tıp Bilimleri* 22(2):233-8,2002.

[2] Allendorf F. W., Luikart G. H. ve Aitken S. N.,Conservation and the genetics of populations.(2nd Edition) , WileyBlackwell, USA, 2013.

[3] Butler J.M., Forensic dna typing: biology, technology, and genetics of str markers. Elsevier academic press, New york,2005.

[4] Zhou Q., Jiang Q. Ve Wei D. ,"A new method for classification in dna sequence", the 6th international conference on computer science&education (ıccse 2011) august 3-5, 2011.
[5] http://www.ncbi.nlm.nih.gov/Genbank/genomes/bacteria.
[6] Ayhan S., Erdoğmuş Ş., "Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi ",iibf dergisi, 9(1), 175- 198, 2014.
[7] Karagülle F., "Destek vektör makinelerin kullanarak yüz bulma" ,yüksek lisans tezi, 2008
[8] A.şengür, i.türkoğlu ve m.c.ince, wavelet packet neural networks for texture classification, expert systems with applications, 32(2), mart 2007.
[9] Sengur, a., "multiclass least-squares support vector machines for analog modulation classification", expert systems with applications, 36(3), 6681-6685 (2009).