T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DNA DİZİ ANALİZİNDE HİZALAMASIZ KARŞILAŞTIRMA İÇİN YENİ YAKLAŞIMLAR

Emre DELİBAŞ

DOKTORA TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Temmuz-2020
KONYA

TEZ KABUL VE ONAYI

Emre DELİBAŞ tarafından hazırlanan "**DNA Dizi Analizinde Hizalamasız Karşılaştırma İçin Yeni Yaklaşımlar**" adlı tez çalışması 27/08/2020 tarihinde aşağıdaki jüri tarafından oy birliği ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda DOKTORA TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof. Dr. Harun UĞUZ	……..
Danışman Prof. Dr. Ahmet ARSLAN	……..
Üye Doç. Dr. Halife KODAZ	……..
Üye Doç. Dr. Gülay TEZEL	……..
Üye Dr. Öğr. Üyesi Ayşe Merve ACILAR	……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Saadettin Erhan KESEN
Enstitü 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.

Emre DELİBAŞ

27.08.2020

ÖZET

DOKTORA TEZİ

DNA DİZİ ANALİZİNDE HİZALAMASIZ KARŞILAŞTIRMA İÇİN YENİ YAKLAŞIMLAR

Emre DELİBAŞ

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Ahmet ARSLAN

2020, 104 Sayfa

Jüri
Prof. Dr. Ahmet ARSLAN
Prof. Dr. Harun UĞUZ
Doç. Dr. Halife KODAZ
Doç. Dr. Gülay TEZEL
Dr. Öğr. Üyesi Ayşe Merve ACILAR

Teknolojinin hızlı gelişiminin bir sonucu olarak ortaya çıkan büyük boyutlardaki biyolojik verinin işlenmesinde geleneksel biyolojik ve istatistik metotlarının yetersiz kalması, konuyu bilgisayar biliminin çalışma alanları arasında odak noktalarından biri haline getirmiştir. Benzerlik, hesaplamalı biyoloji ve biyoinformatikte DNA dizi analizinin anahtar süreçlerinden biridir. Evrimsel ilişkileri, gen fonksiyon analizini, protein yapı tahminini ve dizi eşleşmelerini araştıran neredeyse tüm araştırmalarda, benzerlik hesaplamaları yapmak gerekmektedir. Yüksek hesaplama maliyetiyle sonuçlanan hizalamaya dayalı dizi karşılaştırma yöntemlerine bir alternatif olarak, diziyi farklı bir uzayda sayısallaştırarak, tanımlanan yeni matematiksel tanımlayıcılarla benzerliği hesaplayan hizalamasız yöntemler ortaya çıkmıştır. Bu tez çalışmasında hizalamasız DNA dizi benzerlik analizinde yeni yaklaşımlar önerilmiştir. Önerilen metotlarla DNA dizileri arasındaki benzerliklerin tespitinde kullanılmak üzere özellik çıkarım süreçleri tasarlanmıştır. Uzunluktan bağımsız olarak çalışan metotların farklı uzunluktaki DNA dizileri üzerindeki etkisini test etmek üzere, literatürde kullanılmış üç farklı veri seti ele alınmıştır. Elde edilen özellik vektörlerinin benzerlik metrikleri ile ilişkileri tanımlanmış ve referans olarak kullanılan filogenetik ağaçlarla karşılaştırılmıştır. Testler neticesinde önerilen metotların elde ettiği sonuçların başarılı ve kabul edilebilir oldukları görülmüştür.

Anahtar Kelimeler: DNA dizi benzerliği, hizalamasız benzerlik analizi, özellik çıkarımı, sayısal karakterizasyon

ABSTRACT

PhD THESIS

NOVEL APPROACHES TO ALIGNMENT-FREE COMPARISON IN DNA SEQUENCE ANALYSIS

Emre DELİBAŞ

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Prof. Dr. Ahmet ARSLAN

2020, 104 Pages

Jury
Advisor Prof. Dr. Ahmet ARSLAN
Prof. Dr. Harun UĞUZ
Assoc. Prof. Dr. Halife KODAZ
Assoc. Prof. Dr. Gülay TEZEL
Asst. Prof. Dr. Ayşe Merve ACILAR

The fact that traditional biological and statistical methods are insufficient in the processing of large-scale biological data, which has emerged as a result of the rapid development of technology, has made the subject one of the focal points of the field of computer science. Similarity is one of the key processes of DNA sequence analysis in computational biology and bioinformatics. In nearly all research that explores evolutionary relationships, gene function analysis, protein structure prediction, and sequence retrieving, it is necessary to perform similarity calculations. As an alternative to alignment-based sequence comparison methods that have high computational costs, alignment-free methods have emerged that by digitizing the sequence in a different space, calculate the similarity with the new mathematical identifiers utilized. In this thesis, novel approaches are proposed in alignment-free DNA sequence similarity analysis. With the proposed methods, feature extraction processes are designed to be used to determine similarities between DNA sequences. Three different datasets used in the literature are discussed to test the effect of methods analysing independently of length on different sequences of DNA. The relations of the feature vectors obtained with similarity metrics are defined and compared with phylogenetic trees used as reference. As a result of the tests, it is seen that the results obtained by the proposed methods are successful and acceptable.

Keywords: DNA sequence similarity, alignment-free similarity analysis, feature generation, numerical characterization

ÖNSÖZ

Doktora tez çalışmam boyunca fikirleri, desteği ve anlayışı ile katkısını esirgemeyen, rehberliği ve yol göstericiliğiyle ufkumu açan tez danışmanım ve çok değerli hocam Prof. Dr. Ahmet ARSLAN'a; tezin gelişiminde yönlendirici katkılarıyla destek olan tez izleme komitesi üyeleri Prof. Dr. Harun UĞUZ ve Doç. Dr. Halife KODAZ hocalarıma; bu süre boyunca destekleri, ilgileri ve yardımlarından dolayı Bilgisayar Mühendisliği bölümündeki kıymetli öğretim elemanlarına; tez içeriğindeki moleküler biyoloji ve genetik alanına ait teknik konuların üstesinden gelmeme katkı sağlayan KTÜ Fen Fakültesi Moleküler Biyoloji ve Genetik Bölümü öğretim üyesi, kıymetli dostum Dr. Öğr. Üyesi Halil İbrahim GÜLER'e; özellikle manevi desteğini hiçbir zaman esirgemeyen, her zaman ve her konuda yanımda olan eşime ve aileme teşekkürlerimi arz ederim.

Emre DELİBAŞ
KONYA-2020

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

2-B	: 2 Boyutlu
A	: Adenin
C	: Cytosyne (Sitozin)
ÇDH/MSA	: Çoklu Dizi Hizalama / Multiple Sequence Alignment
DNA	: Deoksiribo Nükleik Asit
FASTA	: FAST-All
G	: Guanin
HTU	: Hypothetical Taxonomic Units
K-LD	: Kullback–Leibler Divergence
LCA	: Last Common Anchestor
MATLAB	: Matrix Laboratory
MEGA	: Molecular Evolutionary Genetics Analysis
NADH	: Nicotinamide Adenine Dinucleotide (NAD) + Hydrogen (H).
NCBI	: National Center for Biotechnology Information
NIH	: National Institutes of Health
NLM	: National Library of Medicine
NLP	: Natural Language Processing
nRF	: Normalized Robinson-Foulds
N-W	: Needleman-Wunsch
OTU	: Observed Taxonomic Units
PCA	: Principal Component Analysis
PET/CT	: Positron Emission Tomography - Computed Tomography
RGB	: Red Green Blue
RNA	: Ribonükleik Asit
T	: Timin
UPGMA	: Unweighted Pair Group Method with Arithmetic Mean

1. GİRİŞ

Ayrık sembollerden oluşan veri dizileri bilgisayar bilimindeki temel veri gösterim biçimlerindendir. Arama motorlarından belge sınıflandırmalarına, gen bulmadan protein fonksiyon tahminlerine, ağ izleme araçlarından anti virüs yazılımlarına kadar birçok uygulama bu tür verilerin analizine bağlıdır. Bu alanlardaki veri serilerine arayüz sağlamak makine öğrenmesi uygulamaları için temel bir önkoşuldur. Makine öğrenmesi algoritmaları, iyi tanımlanmış hesaplama ve matematiksel analiz araçlarının da kullanılabilirliğini sağlayabilmek için geleneksel olarak vektörel yapıdaki veriler için tasarlanmıştır. Öğrenme algoritmalarının büyük bölümü, işlenebilecek veri türü konusunda fazla kısıtlamalar getirmeyen nesneler arasındaki ikili ilişkileri formüle edebilmektedir (Rieck ve Laskov, 2008). Nesneler arasındaki bu ilişki *benzerlik* olarak tanımlanmaktadır.

Verinin temsil ettiği nesneler arasındaki benzerliğin hesaplanması matematiksel karşılaştırmalarla gerçekleştirilir. Bilgisayar biliminde makine öğrenmesi ve veri madenciliği alanlarında önemli bir yere sahip olan benzerlik ve karşılaştırma, bu alanların müdahil olduğu tüm disiplinlerde yerini almıştır. İnsan Genom Projesi'nin (Robbins ve ark., 1995) ilan edilmesi ve teknolojinin hızlı gelişiminin bir sonucu olarak ortaya çıkan büyük boyutlardaki verinin işlenmesinde biyoloji ve istatistik tabanlı metotların yetersiz kalması, konuyu bilgisayar biliminin çalışma alanları arasında odak noktalarından biri haline getirmiştir.

DNA dizi analizi genetik araştırmalar ve analizler içerisinde anahtar rol oynama noktasında önemli bir potansiyele sahiptir. DNA dizi analizi, büyük genomik veri havuzlarındaki benzer nükleotid dizileri, birçok evrimsel ve akrabalık ilişkisini, patofizyolojik süreçleri vb. tanımlamak için ilk adımdır. DNA dizilerinde benzerlik analizi, bilinen dizilerle bilinmeyen dizileri, bilinmeyen dizilerin fonksiyonlarını elde etmek için karşılaştıran veya bu diziler arasındaki ilişkiyi ortaya koyan çıkarım sürecidir (Wang ve ark., 2010). Modern biyolojide veriye dayalı yaklaşımların yaygınlaşması bu araştırma sürecini önemli ölçüde hızlandırmış ve öne çıkarmıştır. Ancak kullanılan algoritmalar ve sistemler, üretilen büyük miktardaki genom verisi için asla yeterince hızlı olamamaktadır (Zielezinski ve ark., 2017).

Araştırmacılar tarafından genetik bilgileri doğru ve hızlı bir şekilde kodlayabilmek adına DNA dizileri arasındaki benzerlikleri analiz etmek için birçok metot

geliştirilmiştir (Jin ve ark., 2017). Bu metotlar genel özellikleri açısından iki gruba ayrılır: *Hizalamaya dayalı analiz metotları* ve *hizalamasız analiz metotları*.

1.1. Hizalamaya Dayalı Benzerlik Analizi

DNA dizileri arasındaki işlevsel, yapısal ve evrimsel ilişkiler için anlamlı sonuçları olabilecek benzerlik bölgelerini tanımlamak için biyolojik dizilerin yapı taşlarını konumlandıran yöntemler olan hizalamaya dayalı metotlar benzerlik analizinin geleneksel metotları olarak kabul edilebilir (Zielezinski ve ark., 2017). Dizi benzerliği arama araçları (ör: BLAST (Altschul ve ark., 1997), FASTA (Pearson ve Lipman, 1988)), çoklu dizi hizalayıcıları (ör: ClustalW (Thompson ve ark., 1994), MUSCLE (Edgar, 2004), MAFFT (Katoh ve ark., 2002)), dizilerin profillerini arama programları (ör: PSI-BLAST (Altschul ve ark., 1997), HMMER/Pfam (Finn ve ark., 2013)) ve tam genom hizalayıcıları (ör: progressiveMauve (Darling ve ark., 2010), BLASTZ (Schwartz ve ark., 2003) dahil olmak üzere birçok başarılı hizalama tabanlı metot önerilmiştir. Bu araçlar genlerin ve proteinlerin fonksiyonlarını değerlendirmek isteyen herkes için geniş çalışma alanları oluşturmuştur. Dizi hizalama genel olarak mümkün olan en fazla sayıda elemanı eşleştirmek için dizilere boşluk ekleyerek yapılır. Eşleşen bölgelerin sayısı tam hizalama için en uygun çözüm olarak kabul edilebilecek seviyede ise bazı konumlardaki uyumsuzluklar göz ardı edilir. Burada amaç sembollerin en uygun eşleşme seviyesini yakalamasıdır (Compeau ve Pevzner, 2015).

Hizalama işlemi *a* ve *b* dizilerini karşılaştırırken, *a* dizisinin *b* dizisine dönüştüğü varsayımıyla bir teorik senaryo sunmaktadır. Her iki dizide de aynı simgeye sahip konumlarda *eşleşme*ler (match) görülür ve bu eşleşmeler korunmuş nükleotidleri temsil eder. *Uyuşmazlık*lar (mismatch) ise söz konusu dizilerdeki tekil nükleotid substitüsyonlarını (substitution) ifade eder. Bir hizalama işleminde her iki diziye de yerine göre *boşluk*lar (gap) eklenmektedir. Bu boşluklara *indel* denir ve dizide oluşan *ekleme* (insersiyon) ve *silinme*leri (delesyon) ifade eder (Mjelva, 2018). Şekil 1.1'de *a* ve *b* dizileri arasında gerçekleştirilmiş bir hizalama görülmektedir.

A T - G T T A T A
A T C **G T** - C - C

Şekil 1.1. Örnek hizalama işlemi

Hizalama işlemi yapılırken hizalama skorlanmaktadır. Hizalama işlemlerinde iki temel yaklaşım vardır: *global hizalama* ve *lokal hizalama*. Global hizalamada *a* ve *b* dizilerinin tamamı için mümkün olan en yüksek skor hedeflenmektedir. Lokal hizalamada bunun yerine, bu diziler arasında en yüksek skoru alan alt diziler aranmaktadır. Genel olarak global dizi hizalama, dizilerin oldukça benzer ve yaklaşık aynı uzunlukta olduğu durumlar için faydalıdır. Öte yandan lokal hizalama ise, benzer bölgeler olduğundan şüphelenilen birbirinden farklı diziler için kullanılır (Polyanovsky ve ark., 2011). Global hizalamada bilinen en yaygın algoritma dinamik programlama tabanlı Needleman & Wunsch (Needleman ve Wunsch, 1970), lokal hizalamada ise Smith& Waterman (Smith ve Waterman, 1981) algoritmalarıdır.

Karmaşık evrimsel senaryoları anlamlandırma seviyemiz ve biyolojik dizilerdeki desenlerin özellikleri hakkındaki bilgilerimiz genişledikçe hizalamaya dayalı metotların bazı kısıtları gün yüzüne çıkmıştır (Zielezinski ve ark., 2017). Bu kısıtların başında hizalamaya dayalı metotların bellek ve zaman açısından oldukça yüksek hesaplama maliyetine sahip olmaları gelmektedir. Bu nedenle multi-genom ölçekli dizi verileriyle kullanımında sınırlamalar söz konusudur. İki diziye ait olası hizalamalar, dizi boyutunun büyümesiyle hızlı bir şekilde yüksek işlem maliyetlerine ulaşır. Dinamik programlama adı verilen ve olası tüm çözümleri listelemeden matematiksel olarak optimum bir hizalama elde etmeyi garanti eden bir yöntem olmasına rağmen, girdi dizilerinin uzunluklarının çarpım sırasına göre değişmesinden dolayı zaman karmaşıklığının artması sebebiyle hesaplama açısından yüksek maliyetlere sahiptir (Eddy, 2004). Bu nedenle hizalama araçlarının zenginliğine ve bu alanda yapılan birçok çalışmaya rağmen uzun dizilerde hizalama sorunu tam olarak çözülememiştir (Earl ve ark., 2014). Öte yandan hizalama işlemlerinde kullanılan skorlama metotları söz konusu olduğunda bir fikir birliği yoktur ve puanlama mekanizmasındaki bir sistemden diğerine metotların değişimi, döndürülen son hizalama sonucunu büyük ölçüde etkileyebilmektedir. Zielezinski ve ark. (2017)'na göre hizalamaya dayalı metotların bir diğer çıkmazı, bir hizalama işleminin düşük özdeşliğe sahip diziler için rastgele dizilerle karıştırılmış uzak homologlarını bulmaktaki başarısızlığıdır. Bu problem daha çok ikili hizalama işlemleri için bir başlıktır ve bu tür diziler için doğru hizalamaların yapılabilmesi için daha karmaşık yöntemlerin uygulanması gerekmektedir. Bahsettiğimiz bu kısıtlar ve dezavantajlar ışığında biyolojik dizilerin benzerliklerinin tespitinde farklı yaklaşımlara olan ihtiyaç gün yüzüne çıkmıştır.

1.2. Hizalamasız Benzerlik Analizi

"Hizalamasız" terimi ilk kez Vinga ve Almeida (Vinga ve Almeida, 2003) tarafından kullanılmıştır. Bu alandaki ilk kapsamlı inceleme makalesi olan çalışmada yazarlar, çeşitli farklı yöntemleri ortak bir çerçevede sistematize etmiş, düzenlemiş ve sınıflandırmıştır. Biyolojik dizilerin analizi ve karşılaştırması için hizalamasız analiz metotları, bu dizilerin özellikleri ve içerdikleri veri desenlerini anlama konusundaki zorlukların üstesinden gelmek için doğal bir çalışma alanı oluşturmuştur. Bu metotlar, DNA, RNA ve proteinleri tanımlayan sembolik dizilerin, daha etkili analizler sunmak amacıyla vektör uzaylarına dönüştürülmesine dayanır. İşlemlerin özü, dizileri sayısal gerçek-değerli vektörler olarak temsil etmek ve bu alana filtreleme teknikleri, normalizasyon, benzerlik tahmini ve kümeleme gibi mevcut araçları uygulamaktır. Güçlü ve kapsamlı bir teorik ve hesaplama altyapısı sağlamak için olasılık, istatistik, doğrusal cebir ve çeşitli makine öğrenmesi uygulamaları gibi geniş bir sistem mevcuttur. Bu durum ise son yıllarda başarılı uygulamaların kapsamını genişletmeyi sağlayan hizalamasız benzerlik analiz metotlarının yüksek hesaplama verimliliğini açıklamaktadır. Hizalamasız analiz metotlarının avantajı, düşük hesaplama maliyetine sahip olmalarının yanı sıra tüm genomlar üzerinde işlem yapabilme kapasitesiyle büyük veya tam dizi analizine imkân sağlama yeteneğidir (Vinga, 2014b).

DNA dizilerinin benzerliklerini doğru ve etkili bir şekilde analiz etmek için, tasarlanan metodun, aşağıdaki kritik problemlerin üstesinden gelebiliyor olması beklenmektedir (Randić ve ark., 2001; Liu ve Wang, 2005):

i. DNA dizilerinin sayısal dizilere dönüştürülmüş halinin, bu dizileri etkili bir şekilde temsil edebilmesi.
ii. Sayısal karakterizasyon için DNA dizilerinin öznitelikleri olarak kabul edilebilecek tanımlayıcılarının (veya sabitlerinin) seçilmesi ve elde edilmesi.
iii. Farklı uzunluktaki DNA dizilerinin etkili bir şekilde işlenebilmesi ve tutarlılığının korunması.

Yukarıdaki problemlere uygun çözüm sunabilmesi için tasarlanan metotların aşağıdaki özelliklere sahip olması beklenmektedir (Jin ve ark., 2017):

i. Metot DNA dizisinin sayısal karakterizasyonu sürecinde veri kaybına neden olmamalıdır.

ii. DNA dizisindeki sayısal karakterizasyonda kullanılacak tanımlayıcılar farklı türlerin genetik bilgilerini etkili bir şekilde temsil etmeli ve benzerliklerini etkili bir şekilde niceliksel olarak hesaplamalıdır.

iii. DNA dizilerinin sayısal karakterizasyonu uzunluktan bağımsız olarak tasarlanmalıdır.

1.3. Tezin Önemi ve Amacı

DNA dizilerinde çeşitli uzunluklar söz konusudur. Farklı DNA dizilerinin değişen uzunlukları hizalamaya dayalı benzerlik analizleri için önemli bir problemdir. DNA dizilerinin benzerlikleri, dizilerin biyolojik işlevlerini ve evrimsel bilgilerini öğrenmek için kullanılmaktadır. Buna göre de hesaplanan benzerlikler biyolojik mekanizmaları ile temsil edilecektir. Bununla birlikte biyolojik özelliklerin ve kimyasal özelliklerin makul bir şekilde nasıl dikkate alınacağı göz ardı edilmemesi gereken önemli başlıklar arasında yer almaktadır (Jin ve ark., 2017). Biyologlar tarafından sıklıkla kullanılan DNA dizi benzerlik analiz araçları hala hassasiyet ve zamansal maliyet açısından kısıtlara sahiptir (Kumar ve ark., 2016b). Daha doğru ve kullanışlı benzerlik analizi yöntemleri, bilinmeyen DNA dizilerinin fonksiyon bilgilerini ve evrim bilgilerini bulmamıza etkili bir şekilde yardımcı olabilecektir (Wang ve ark., 2010). Bu çerçevede hizalama gerektirmeyen ve düşük maliyetli benzerlik analiz metotlarının kullanımına duyulan ihtiyaç, bu alandaki yapılmış birçok çalışmanın üzerine gelecekte de yoğunlaşılması gerektiğini göstermektedir.

Hazırlanan bu tezde, DNA dizilerinin benzerlik analizlerinin yapılabilmesi için hizalamasız metotlar üzerinde çalışılmıştır. Bu amaçla, bilgisayar mühendisliği alanında kullanılan ve içeriğinde “benzerlik” hesaplaması bulunan çeşitli metotların DNA benzerliğinde de kullanılabileceği öngörüsüyle algoritma tasarımları yapılmıştır. Önerilen hizalamasız metotlarda, dizileri karakterize eden metriklerin oluşturduğu vektörel dönüşüm, dizilerin uzunluklarından bağımsız hesaplanması, küçük boyutlu vektörler ve hesaplama yöntemlerinin düşük hesaplama maliyetlerine sahip olması, ihtiyaç duyulan metotlarda aranan özellikleri karşılayabilmektedir.

Tez çalışmasında önerilen ilk metotta dizilerin biyolojik ve kimyasal özelliklerinin istatistiksel hesaplamaları ile yapılan özellik çıkarımı neticesinde 7-boyutlu bir vektör elde edilerek bu vektörler arasında benzerlik hesaplaması yapılmıştır. İkinci metot, dijital görüntü analizinin bir alt çalışma konusu olan doku analizinin DNA dizi

benzerliğine uyarlanması şeklinde tasarlanmıştır. DNA dizileri gri-seviye görüntülere dönüştürülerek elde edilen dokulardan özellik çıkarımı yapılmış ve doku histogramının istatistiksel metriklerinden 4-boyutlu vektörler elde edilerek benzerlik hesaplaması yapılmıştır. Önerilen son metotta ise kelime tabanlı algoritmalardan *n*-gram modeli dizi benzerliğinde kullanılmıştır. Metotta dizilerin *n*-gram frekansları çıkarılmış, bu frekanslara göre azalan sıra ile belirli oranda üst bölge alınarak üst-*k* *n*-gramların diğer dizilerle eşleşme skorlarından özellik vektörleri çıkarılmıştır. Bu vektörler arasında da yine benzerlik hesaplamaları yapılarak DNA dizileri arasındaki benzerlikler çıkarılmıştır.

1.4. Tezin Organizasyonu

Tez çalışması altı ana bölümden oluşmaktadır.

İlk bölüm "giriş" bölümüdür. Burada tezle ilgili özet bilgiler, DNA dizi analizinde hizalama ve hizalamasız analiz kavramları hakkında kısa tanımlamalar yapılmıştır.

İkinci bölümde DNA dizi analizinde hizalamasız metotlar hakkında literatür çalışmaları hakkında bilgi verilmiştir. Bu çalışmaların hizalamasız analiz işlemlerine katkılarından bahsedilmiştir. Mevcut metotların özellikleri ve üstün/kısıtlı yönleri vurgulanmıştır. Bu çalışmada önerilen hizalamasız analiz metotlarının amaçları anlatılmış ve üstünlüklerinden bahsedilmiştir.

Üçüncü bölümde tez çalışmasında kullanılan materyal ve yöntemlerden bahsedilmiştir. Bu bölümde DNA yapısı, dizi özellikleri, hizalama, özellik çıkarımı ve özellik vektörü, görüntü doku analizi, kelime tabanlı metin benzerliği, hiyerarşik kümeleme, çıktılar, değerlendirme ölçütleri ve tez çalışmasında kullanılan veri kümeleri hakkında detaylı bilgiler verilmiştir.

Dördüncü bölümde DNA dizilerinde hizalamasız benzerlik analizi yapmak amacıyla üç yeni metot önerilmiştir. Önerilen "Grup Mutasyonlarının Ortalama Pozisyonu ve Nükleotid Bazlarının Frekansları Tabanlı Hizalamasız DNA Dizi Analiz Metodu", "Birinci Dereceden İstatistik Tabanlı Görüntü Doku Analizi Kullanarak DNA Dizi Benzerlik Analizi" ve "Üst-*k* *n*-gram Eşleşmesi Tabanlı DNA Dizi Benzerlik Analizi Yaklaşımı" yöntemleri açıklanmış ve yöntemlerin aşamaları anlatılmıştır.

Beşinci bölümde önerilen metotlardan elde edilen sonuçlar listelenmiştir. Bu sonuçlar literatürdeki diğer çalışmaların sonuçları ile farklı ölçütler üzerinden karşılaştırılarak incelenmiştir.

Son bölümde ise tez çalışması değerlendirilmiş ve elde edilen sonuçlar genel olarak değerlendirilmiştir. Tez çalışmasının genel katkıları anlatılmıştır. Gelecekte yapılabilecek çalışmalar hakkında önerilerde bulunulmuştur.

2. KAYNAK ARAŞTIRMASI

Bu bölümde hizalamasız DNA dizi analizini ele alan metotlara ilişkin geniş bir inceleme sunulmaktadır. Bu amaçla hizalamasız analiz metotlarının benzerlik analizi yaklaşımlarına göre gruplandırma yapılmış, bu gruplara göre metotlar ele alınmıştır.

2.1. Grafiksel Gösterim Tabanlı Metotlar

Grafiksel gösterim tabanlı metotlar DNA dizi benzerlik analizi için, dizilerin görsel olarak incelenmesine imkân sağlayan yaygın sayısallaştırma metotlarındandır. Ayrıca verinin görselleştirilmesi ile diziler arasındaki farklılıkların da incelenmesini sağlamaktadır. Dizilerin temsil edildiği uzayın boyutuna göre isimlendirilmek üzere 2-Boyutlu (2-B), 3-B ve daha fazla boyutlu gösterim olarak üç grupta incelenebilir.

2-B gösterim tabanlı metotlar DNA dizi benzerlik analizinde sıklıkla kullanılan şemalardır. Dört tip nükleotid bazın (A, G, C, T) koordinat ekseninin dört yönü veya dört simetri eşdeğeri olmayan yatay çizgi gibi çeşitli geometrik alternatiflerle spesifik bir eşleştirmesini içermektedir (Randic ve ark., 2003). Bu haritalama metodu doğrudan bazların dört farklı sayıyla gösterimi, bazların birleşimlerinin gösterimi ya da bazların kimyasal özelliklerine göre gösterimleri şeklinde olabilmektedir.

Randic ve ark. (2003) dört farklı nükleotid bazın koordinat sisteminde rastgele yatay çizgiler şeklinde atanmasını içeren bir 2-B zikzak grafik gösterimi önermiştir ve grafik çizgisindeki her bir birim bir nükleotid baza denk gelecek şekilde tasarlamışlardır. Bu gösterime bağlı bir *L/L* matris aracılığıyla DNA dizileri arasındaki benzerlik hesaplamasını yapmışlardır. Metot, farklı DNA dizileri ile birleştirilmiş *L/L* matrislerinin baskın özdeğerleri olan bileşenlerden oluşan bir 12-bileşenli vektörden oluşturmaktadır. Guo ve Wang (2008) ise bir DNA dizisine karşılık gelen 2-B Kartezyen koordinat sisteminde ayarlanan 24 farklı eğriye karşılık gelen 2-B grafik gösterim tabanlı analiz metodu önermiştir. Metotlarında grafik gösterim için matrisin baskın özdeğerlerini hesaplamak yerine, zikzak eğrisini yumuşatarak bu eğriliği benzerlik hesaplaması için DNA dizilerinin tanımlayıcısı olarak kullanmışlardır.

Dört bazın kimyasal özellikleri, DNA yapısını oluşturmak için anahtar öneme sahip olduğundan benzerlik analizlerinde sıklıkla kullanılmıştır. Wang ve Zhang (2006) iki boyutlu uzayda üç kimyasal yapıya dayalı bir yöntem önermiştir. DNA'nın kimyasal yapısına göre *non-A = G, C, T*; *non-G = A, C, T* ve *non-C = A, G, T* tanımlanmış, buna

göre de *non-A, non-g* ve *non-C* değerleri 1 ile; *A, G* ve *C* ise 0 ile belirtilmiştir. Bu şekilde ${}^{inf}L$ / ${}^{inf}L$ matrisine dönüştürülmek üzere üç adet (0,1)-dizi elde edilmiştir. Bu matrislerin maksimum ve minimum özdeğerlerinin toplamı, DNA dizilerini tanımlamak için bir matematiksel sabit olarak hesaplanmıştır. Benzerlikler Öklid uzaklığı ve farklı DNA dizilerinin vektörlerinin uç noktaları arasındaki korelasyon açısı ile hesaplanmıştır.

Dai ve ark. (2006) yaptıkları çalışmada ise iki boyutlu dairelere gömülmüş olan *W*-eğrisi adı verilen 2-B uzayda bazlar arasındaki iki harita ile DNA dizisini temsil etmişlerdir. Bu yöntemde, girişlere sahip 8 bileşenli bir vektör, üç kimyasal özelliğin apsis ve ordinatın ortalama toplamları ile elde edilmiştir. Farklı DNA dizilerinin benzerliklerini hesaplamak için Öklid uzaklığı ve korelasyon açısı kullanılmıştır. Yao ve ark. (2006)'nın çalışması incelendiğinde de *W-S* eğrisi *M-K* eğrisi ve *R-Y* eğrisi adını verdikleri Kartezyen koordinat sistemindeki üç karakteristik eğriden oluşan kimyasal özelliklere dayanan başka bir 2-D grafik gösterim yöntemi önerilmiştir. Metotta daha sonra sırasıyla DNA eğrisi ile ilişkili grafik yarıçapı, açısı ve bağıl hareketine dayanan benzerlik analizi yöntemi sunulmuştur. Genel olarak bu iki çalışmada dört bazın kimyasal özellikleri, 2-B uzayda DNA dizilerini sayısallaştırmak için kategorileri ile birleştirilmiştir. Böylece benzerlikler hakkında yararlı bilgiler yansıtılabilmektedir. Her iki yayında da temsil yöntemlerinin DNA dizilerinin bilgi kaybını önleyebileceğini, ancak uygun değerlendirme yöntemlerinin bulunmaması nedeniyle yine de bu bakış açısının kanıtlanmasında sorunlar olduğu beyan edilmiştir. Yukarıda bahsedilen üç çalışmada da sadece çok kısa (yaklaşık 100 baz uzunluktaki) diziler kullanılmıştır ve bu da bu tür analizlerin gücünü ispatlama noktasında oldukça yetersiz görülmektedir (Jin ve ark., 2017).

DNA dizisindeki bazların birleşimi lokal permütasyonunu yansıtabilir ve kısa bir parçanın biyolojik anlamını belirleyebilir. Buna göre Liu ve ark. (2006) komşu nükleotid çiftlerinin 16 farklı çiftine ve PNN ile ifade edilen kimyasal özelliklerine dayalı 2-B grafiksel gösterim metodu önermiştir. DNA'ya ait sayısal dizileri elde etmek için 4x4 hücre ve sistemler tasarlamıştır. Daha sonra PNN'lerin dağılımları ve bir y koordinat matrisi (y-M) elde ederek DNA dizilerinin benzerliklerini elde etmek için iki y-M'nin korelasyonu tanımlanmıştır. Benzer şekilde Yu ve ark. (2010) komşu nükleotid çiftlerinin 16 çeşidine ve kimyasal özelliklere dayanan DNA dizilerini sayısallaştırmak için 2 genişlikli bir kayan pencere kullanmıştır. Konumu ile koordine edilen 2 boyut, bir DNA dizisini (x, y, n) temelli döngü içermeyen tekil bir 3-B eğrisine dönüştürebilen bir D eğrisi

olarak kabul edilebilir. Bu metottan DNA dizilerinin niceliksel tanımlayıcıları olarak 6 bileşen kullanılmıştır.

Jafarzadeh ve Iranmanesh (2012) yaptıkları çalışmada DNA dizileri için her bir kodon türüne dayanan bir 2-B grafiksel gösterim yöntemi önermişlerdir. Grafiksel temsil verileri, hesaplanan D/D matrisini ve nicel karşılaştırmaları kolaylaştırmak için bir matrise dönüştürülmüştür. Daha sonra bu matrisin "ALE-endeksi" olarak adlandırılan bir sabiti hesaplanmıştır.

3-B grafiksel gösterim DNA dizi benzerlik analizinde yine sıklıkla kullanılan metotlar arasındadır. Bu metotlar DNA dizilerini bazı kurallara göre belirli bir 3-B uzaya eşitlemektedir. 3-B uzaydaki haritalama yöntemi bazların veya baz birleşimlerinin kimyasal özelliklerine dayanabilmektedir. Haritalamada kullanılan kuralların, benzerlik analizi yönteminin, özellikle de bellek alanının ve hesaplama hızının üzerinde büyük bir etki yaratacağı belirtilmiştir (Jin ve ark., 2017).

Liao ve Wang (2005); Liao ve ark. (2005) çalışmalarında bazların kimyasal özelliklerine dayanarak 3-B uzayda DNA dizileri için benzerlik analiz metodu sunmuşlardır. Bu çalışmalarda, DNA dizisinin 3-B eğrisi, kimyasal özelliklerin üç karakteristik eğrisi ve bazların sırası ile temsil edilmiştir. İlk yapılan çalışma, bileşenleri DNA dizileri ile ilişkili L / L matrislerinin baskın özdeğerleri olan 3-bileşenli vektör oluşturmuştur. Diğer çalışmada ise tam DNA dizisi kullanılarak D / D matrislerinin ortalama bant genişliğinden oluşan 15 bileşenli bir vektör oluşturulmuştur. Bu çalışmalarda, her bir bazın kategorilerine göre belirlenen DNA'yı temsil yapısı, aynı zamanda kimyasal kategorilerine de karşılık gelerek bu durumu somutlaştırmıştır. Sonuç olarak bu yaklaşımlar sadece dizinin yapısını değil aynı zamanda DNA primer dizileri için de kimyasal yapıyı dikkate almıştır. Her iki yöntem de DNA dizilerini Öklid uzaklığı ve korelasyon açısı ile benzerliklerini hesaplamıştır.

Liu ve ark. (2006)'nın çalışması benzer olarak Qi ve ark. (2007) 16 farklı ikili nükleotid çiftine ve bunların kimyasal özelliklerine göre şekillendirilmiş bir 3-B grafiksel gösterim metodu önermişlerdir. Sayısal dizileri (*s*-vektör, *p*-vektör) elde etmek için bir 4x4 matris tasarlamışlardır: s-Matristeki 16 uzay-toplamından oluşan *s*-vektör, *p*-Matrisindeki 16 dağılımdan oluşan *p*-vektör. Bu vektöreler arasındaki benzerlikler iki şekilde hesaplanmıştır: (i) *s*-vektörlerin bitiş noktası arasındaki Öklid uzaklığı, (ii) *p*-vektörlerinin bitiş noktası arasındaki Öklid uzaklığı.

Jafarzadeh ve Iranmanesh (2013) DNA dizisindeki dört nükleotidin 64 çeşit kodonuna ve kimyasal özelliklerine dayanarak, DNA dizilerinin sayısal temsilini

gerçekleştirmek ve kantitatif karşılaştırmalarını kolaylaştırmak için iki matrise dönüştürülmüş başka bir 3-B grafiksel gösterim yöntemi (C-eğrisi) önermiştir. Daha sonra DNA dizilerinin analizi için uzay-toplam matrisi (*s*-M) ve dağılım matrisi (*p*-M) hesaplanmıştır. Benzer şekilde Yu ve ark. (2009) TN eğrisi olarak adlandırılan trinükleotidlere dayanan bir 3-B grafik gösterim yöntemi önermişlerdir. DNA dizisini sayısal olarak temsil etmek için altı tanımlayıcı kullanılan metotta, daha sonra farklı DNA dizilerinin benzerlik analizi için 64 ve 6 bileşenden oluşan iki vektör ile iki Öklid uzaklığı tabanlı yöntem hesaplanmıştır. Bu çalışmanın dışında Yao ve ark. (2005) DNA dizisinin sayısal karakteristik eğrisini elde etmek için DNA dizilerinin spesifik bir pozisyon korelasyonunu göz önünde bulundurarak 3-B grafiksel gösterim yöntemine dayanan başka bir benzerlik analiz yöntemi önermişlerdir. DNA dizi benzerlik ölçümü, bileşenleri geometrik merkez olan 3 bileşenli bir vektör ve DNA eğrileriyle ilişkili grafik yarıçapından oluşan 4 bileşenli bir vektörün oluşturulmasıyla gerçekleştirilmiştir.

2.2. Kelime Tabanlı Metotlar

Kelime tabanlı DNA dizi benzerlik analiz metotları literatürde sıklıkla kullanılan bir diğer metot grubudur. Bilgisayar biliminde doğal dil işleme alanının önemli algoritmalarından *n*-gram, çalışılan bir metindeki "*n*" uzunluklu kelimelerinin dizileridir. Bir metinde oluşan farklı *n*-gramlar ve her *n*-gramın görülme sıklığı belgenin konusunu veya yazar tarzını karakterize etmek için kullanılabilir. *N*-gram frekansları veya daha karmaşık *n*-gram istatistik modelleri, bilgi alma, dil tanımlama, otomatik metin kategorizasyonu ve yazarlık ilişkilendirmesi gibi metin işleme uygulamaları için yaygın olarak kullanılmaktadır. Biyolojik bir bağlamda, *n*-gram amino asitlerin veya nükleotidlerin dizileri olabilir. Doğal dil metinleri ve biyolojik diziler arasında bu benzetmeyi kullanarak, yani "*biyolojik dil modellemesi*" uygulanarak, mikrobiyal organizmaların tüm proteom (genomun protein karşılığı) dizilerinin de *n*-gram genom-imzalarını içerdiği gösterilmiştir (Osmanbeyoglu ve Ganapathiraju, 2011).

Metin analizlerinde kullanılan kelime tabanlı algoritmalar benzer DNA dizilerinin benzer kelimeleri paylaştığı öngörüsü ile yola çıkmışlardır. Literatürde alt kelimeler için *n*-gram dışında *k*-mer, *l*-mer, *k*-tuple gibi isimleri kullanan bu metotların temelinde benzerlik hesaplaması, alt kelime oluşumları ile yapılan matematiksel işlemlerin dizi farklılığının ölçüsü olarak ele alınmasıyla yapılmaktadır.

N-gram tabanlı yaklaşım ilk olarak Blaisdell (1986) tarafından kullanılmıştır. Ökaryotik DNA dizilerinin analizinde kullanılan metotta birinci ve ikinci dereceden Markov zinciri homojenliğinin belirlenmesi ile iki ve üç gramların dağılımlarını kullanarak benzerlik tespiti yapılmıştır. Daha sonra Wu ve ark. (2001) iki DNA dizisi üzerinden kayan pencere ile elde edilen *n*-gramların bağıl frekanslarının vektörlerinden benzerlik hesabını yapmışlardır. Luo ve ark. (2008) üç nükleotid sınıflandırmasına göre, on altı komşu çift nükleotidi dört sınıfa ayırmışlardır. Her diziyi, her sınıftaki çift nükleotidlerin nispi frekansları ile ilişkilendirerek on altı çift nükleotide göre on altı bileşenli bir vektör elde etmişlerdir. Wang ve Zheng (2008) yaptıkları çalışmada, biyolojik dizilerin kısa sözcük birleşimine dayanan yeni bir uzaklık metriği olan Ağırlıklı Dizi Entropisi (WSE) sunmuşlardır. Klasik bağıl entropinin bir revizyonu olarak, bu metrik, *n* küçük olduğunda filogenetik çıkarımdaki bağıl entropiye eşit olarak çalışmaktadır ve büyük *n* durumunda dejenerasyondan kaçınmıştır. Kelime tabanlı bu çalışmalarda DNA dizilerinin benzerliği, dizi içerisindeki her gramın konumu göz ardı edilerek hesaplanmıştır.

Yang ve Wang (2013) çeşitli uzunluklarda biyolojik diziler için uygun yeni bir hizalama gerektirmeyen dizi karşılaştırma yöntemi oluşturmuşlardır. Konumların da dahil edilmesiyle biyolojik dizilerden bire bir haritalama altında çıkarılan kısa *n* kelimesinin alt dizileri üzerinde lineer regresyon modeli değerlendirilmiştir. Buna göre, dizi karşılaştırması için bir benzerlik mesafesi önerilmiştir. Regresyon analizinde bazlar, dinükleotid ve trinükleotid olarak düşünülmüştür. Bao ve ark. (2014) DNA dizisini nükleotid bazlarının kelime frekansı, pozisyonu ve sınıflandırmasına göre üç diziye dönüştüren bir kategori-pozisyon-frekans modeli tasarlamışlardır. Xie ve ark. (2015) ise DNA dizilerini sık geçen dizi desenlerine ve entropisine dayanan, DNA dizisini aynı uzunlukta birkaç bloğa bölen bir yöntem önermişlerdir. Trinükleotid tabanlı bu yöntemler, trinuleotidlerin biyolojik bilgilerini, tekil ve dinükleotidlerden daha makul bir şekilde yansıtabilmiştir. Ayrıca, yukarıdaki üç çalışmada DNA dizilerini temsil etmek için geleneksel yöntemlere göre bazların daha karmaşık bileşimleri olarak kabul edilebilecek sık sıralı örüntüler temelli yöntemler benimsenmiştir. Bazların birleşiminden oluşan diğer geleneksel yöntemlerle karşılaştırıldığında, üç yöntem DNA dizilerinde daha uzun vadeli bağımlılık bilgisi ve yerel bilgi bulabilmiştir.

Kantorovitz ve ark. (2007) çalışmalarında *D2z skor* adı verilen, hizalamasız dizi karşılaştırması için yeni bir puanın kullanımını araştırmışlardır. Metot, iki dizideki tüm sabit uzunluklu kelimelerin frekanslarını karşılaştırmaya dayanmaktadır. Skorun önemli,

yeni bir özelliği, rastgele arka plan dağılımlarından çizilen dizi çiftleri ile karşılaştırılabilir olmasıdır. Dai ve ark. (2008) Markov modeline doğrudan *k*-kelime dağılımlarını ekleyerek, dizi karşılaştırması için *wre.k.r* ve *S2.k.r* adı verilen iki yeni istatiksel ölçüm üzerine araştırma yapmışlardır. Bu ölçümler Jensen-Shannon ıraksama hesabının genişletilmiş versiyonudur.

Qi ve ark. (2011) yaptıkları çalışmada nükleotidlerin iki uzunluklu pencerenin kaydırılması ile elde edilen grupların geçiş bilgilerini çift-yönlü bir graf üzerine aktarmışlardır. Bu geçişler için, kullandıkları bir fonksiyonla 16 boyutlu bir vektör elde ederek bu vektörler arasında benzerlik metrikleri ile analizlerini gerçekleştirmişlerdir. Zhou ve ark. (2016) ise karmaşık ağların karakterizasyonuna dayanan yeni bir matematiksel tanımlayıcı oluşturmuşlardır. Özellikle her DNA dizisinin filogenetik ilişkilerinin analizinde kullanılacak olan 60 boyutlu bir karakterizasyon vektörü oluşturmak üzere karmaşık bir ağ oluşturulmuştur. Bu ağın tasarımında dizideki nükleotidlerin 2, 3, 4, 5 ve 6 uzunluklu parçaları ve bunlar arasındaki ilişki modellenmiştir.

DNA dizilerindeki alt metin parçalarının istatistikleri de bazı çalışmalarda doğrudan benzerlik metriği olarak kullanılmıştır. Reinert ve ark. (2009) çalışmalarında karşılaştırılan iki dizi için *k*-tuple içeriğinin karşılaştırılmasına dayanan, D_2 istatistiğini kullanmışlardır. Daha önce bu tür bir çalışmaya uygun olmadığı düşünülen metodun, D_2^S ve D_2^* olarak adlandırılan iki yeni geliştirilmiş versiyonunu önermişlerdir. Kendinden standardize edilmiş bir istatistik olan D_2^S için, dizinin uzunlukları sonsuzluğa eğilimli olduğu ve münferit dizilerdeki gürültü tarafından baskın olmadığı zaman, istatistiğin normalde asimptotik olarak dağıldığı gösterilmiştir. İkinci istatistik olan D_2^*, iki dizi arasındaki ilişkiyi tespit etme gücü bakımından D_2^S'ten daha iyi performans göstermiştir, ancak D_2^* 'ın asimptotik dağılımından benzetim yapmak kolay olsa da güç hesaplamaları için kapalı bir form sağlayamamıştır.

2.3. Kimyasal Özelliklere Göre Metotlar

DNA dizilerini oluşturan nükleotidler kimyasal yapı özelliklerine göre farklı gruplamalara tabi tutulabilmektedir. İkili gruplar halinde üç farklı kategoride incelenebilir: (i) Halka yapısına göre *Pürin* (Adenin, Guanin) ve *Pirimidin* (Sitozin, Timin), (ii) Fonksiyonel gruba *Amino* (A, C) ve *Keto* (G, T), (iii) Hidrojen bağa göre *Güçlü H-bağ* (G, C) ve *Zayıf H-bağ* (A, T). Kimyasal yapılarının ortak yönü heterojen

bir altıgen döngüdür. Literatürdeki bazı çalışmalar nükleotidlerin bu şekilde gruplandırılmasından faydalanarak farklı karakterizasyon metotları geliştirmişlerdir. Shi ve Huang (2012) DNA dizilerindeki dört nükleotidi, kimyasal özelliklerine göre üç çeşit sınıflandırmaya ayırmıştır. Bir DNA birincil dizisini üç sembolik diziye dönüştürerek üç sembolik dizinin grup mutasyonlarının frekansları, DNA dizisini temsil etmek için on iki bileşenli bir vektörde gruplandırılmıştır. Girilen vektörler arasındaki Öklid mesafeleri, kodlama dizilerini karakterize etmek ve karşılaştırmak için uygulanmıştır. Ayrıca Kumar ve ark. (2016a) belirli bir nükleotid dizisini temsil eden ve Kolekar ve ark. (2010)'ın stokastik modeli kavramından uyarlama ile oluşturulan 36-B periyodik sayım değeri (PCV) sunmuştur. PCV yapısı, nükleotidlerin fizikokimyasal özellikleri ve bir gen içindeki konumsal dağılım deseni hakkındaki bilgiyi kullanmaktadır.

2.4. Bilgi Teorisi Tabanlı Metotlar

Biyolojik dizi analizi, bilgi teorisinden türetilen kavram ve metotlardan büyük ölçüde faydalanmıştır. Örneğin, ilk olarak gazların termodinamiğini incelemek için kullanılan *entropi* kavramı, daha sonra olasılıklı bir deneyle ilişkili belirsizliğin bir ölçüsü olarak tanımlanmıştır ve dizi rastgeleliğinin tahmininde uygulanmıştır. Literatürde termodinamik (Boltzmann ilkesi) ve istatistik (Fisher bilgi matrisi) anlamlarını karşılaştırarak disiplinler arasındaki entropi ve bilgi tanımları araştırılmıştır. *Karmaşıklık* kavramı bir kaynağın entropisi ile ilişkilidir. Bir dizinin fiziksel karmaşıklığı kavramı, belirli bir ortam hakkında belirli bir dizide depolanan bilgi miktarını ifade etmektedir. *Sıkıştırma,* Shannon'un entropi tanımlarıyla da ilişkilidir ve biyolojik dizilere de uygulanmıştır. Bu kavramlar arasında açık bir ilişki bulunmaktadır. Şöyle ki; düşük entropili bir dizi, prensip olarak, daha sıkıştırılabilir olacaktır ve sıkıştırılmış dizinin uzunluğu, karmaşıklığının ve sonuç olarak entropisinin tahminini vermektedir (Vinga, 2014a).

Bilgi teorisi tabanlı kavramlara dayalı ölçümler, dizileri hizalamasız karşılaştırmak için yaygın olarak uygulanmıştır (Vinga ve Almeida, 2003). Genellikle bu uygulamalar, filogenetik rekonstrüksiyon çalışmalarının temel bir yönü olan genomik dizileri sınıflandırmak ve / veya kümelemek için farklılıkları tanımlamaya çalışmıştır. Bu bağlamda, karşılıklı bilgi ve sıkıştırmaya dayalı yaklaşımlar, bilinen moleküler evrimsel olayların çoğuna uyan sağlam sonuçlar vermiştir.

Karşılıklı bilgi bağlamında entropi kavramı karşılaştırma işlemleri için anahtar özelliklerdendir. Özellikle Kullback-Leibler uyumsuzluğu (K-LD), hizalamasız teknik olarak popüler bir seçim olmuştur. Sadovsky (2003)'nin önerdiği yöntem, karşılaştırılan dizilerin küçük parçalarının frekanslarının karşılaştırılmasına dayanmaktadır. Ne metin düzenlemesi ne de karşılaştırılan nesnelerin diğer dönüşümlerini gerektirmez. Karşılaştırma, bir frekans sözlüğünün hibrit denilen özel sözlüğe karşı özgül entropisinin hesaplanmasıyla gerçekleştirilir. Wu ve ark. (2005)'nın önerdikleri simetrik K-LD versiyonu (SK-LD), karıştırılmış açık okuma çerçevesi dizilerinin sınıflandırılması için test edilmiş ve BLAST ile karşılaştırıldığında genom yeniden düzenlemelerinin varlığında daha yüksek performans gösterdiği tespit edilmiştir.

Yang ve ark. (2005) tarafından *l*-mer frekansları ve entropi kavramlarını birleştiren karma bir yaklaşım önerilmiştir. Bilgi tabanlı benzerlik endeksinin ana fikri, frekans sıralamasındaki ağırlıklı bir şehir-blok farklılığına karşılık gelen bağıl Shannon entropisi ile ağırlıklandırılmış iki dizinin *l*-mer derecelerini karşılaştırmaktır. Sıralama dağılımları kullanılarak yapılan Zipf ve artıklık (redundancy) analizi, kodlanan ve kodlanmayan bölgeler arasında farklı bir çalışmada zaten ayırt edilmiştir (Mantegna ve ark., 1994). Zipf'in yaklaşımı, dilbilimsel metinlerdeki kelime oluşumlarının histogramlarını hesaplamaya ve bunları en sık olandan daha az sıklıkta olana göre sıralamaya dayanır. Doğal dillerdeki dikkat çekici özelliklerden biri, bu fonksiyonun çift logaritmik ölçekte doğrusal bir ilişkisinin bulunduğu Zipf yasasıdır. Mantegna ve ark. (1994) tarafından elde edilen sonuçlar, Zipf'in regresyonu uygulandığında, kodlamayan bölgelerin regresyon parametreleri açısından doğal dillere daha yakın olduğunu ve blok entropilerle ölçüldüğü gibi kodlama bölgelerine göre daha fazla artıklık sergilediğini göstermektedir. SARS koronavirüsüne yapılan uygulama, genomik analiz için BT ve sıralama düzeni istatistiklerini birleştirme potansiyelini göstermektedir.

3. MATERYAL VE YÖNTEM

DNA dizi analizinde temel kavramlar, özellik oluşturma, dijital görüntü ve doku analizi, kelime tabanlı benzerlik analizi kavramları, kümeleme ve hiyararşik kümelemenin benzerlik analizindeki yeri, benzerlik analizinin çıktıları ve bu çıktıların görselleştirilmesinde kullanılan araçlar bu bölümde bahsedilen konulardır. Araştırmacıların erişimine ve kullanımına açık genetik veri tabanları hakkında kapsam, erişim yöntemleri, veri formatları, önerilen metotların testinde kullanılan veri setleri de bu bölümde aktarılan başlıklardır.

3.1. DNA Dizi Analizinde Temel Kavramlar

3.1.1. Hücre

Hücreler, tüm canlıların temel yapı taşlarıdır. İnsan vücudu trilyonlarca hücreden oluşur. Vücut için yapı sağlar, besinlerden nütrientleri alır, bu nütrientleri enerjiye dönüştürür ve özel işlevler yerine getirir. Hücreler ayrıca vücudun kalıtsal materyalini içerir ve kendilerinin kopyalarını oluşturabilir. Hücreler, her biri farklı bir işleve sahip birçok parçaya sahiptir. Organel adı verilen bu parçalardan bazıları, hücre içinde belirli görevleri yerine getiren özel yapılardır.

3.1.2. Biyolojik diziler

Teknolojik gelişmelere paralel olarak, giderek artan miktarda DNA, RNA veya protein dizilerine erişim sağlanarak daha fazla tür için genetik bilgiye ulaşılmaktadır. Post-genomik çağda karşılaşılan en büyük zorluklardan biri, halk arasında “yaşamın dili” olarak adlandırılan bu genetik dizileri deşifre etmektir. Bu ifadenin vurguladığı anlamda olduğu gibi, dilbilimsel metafor genetikte uzun süredir kullanılmaktadır. Gerçekten de 1953'te DNA'nın çift sarmal yapısının keşfi, bu biyolojik makromolekülde bulunan genetik bilginin nükleotidleri simgeleyen dört harfli bir alfabe (A, G, C, T) ile ifade edildiğini ortaya koymuştur. Bu genetik bilgi, hemen hemen aynı dört harfli alfabe üzerindeki bir diziyle de temsil edilebilen RNA tek iplikli makromoleküllere, gen adı verilen DNA dizilerinin parçalarına ihtiyaç duyulduğunda transkripsiyonla canlı bir organizmayı oluşturmak ve işletmek için kullanılır. RNA’da T, metillenmemiş U

formuyla değiştirilir (A, G, C, U). Proteini kodlayan mRNA dizileri, 20 amino asitten oluşan, hücrelerdeki üç boyutlu konformasyonlarını ve fonksiyonlarını belirleyebilecek protein dizilerine dönüştürülür (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y). Bu nedenle diziler kalıtım bilgilerinin depolanması ve hücrelerin fonksiyonel birimlerini ifadesinin merkezinde yer almaktadır.

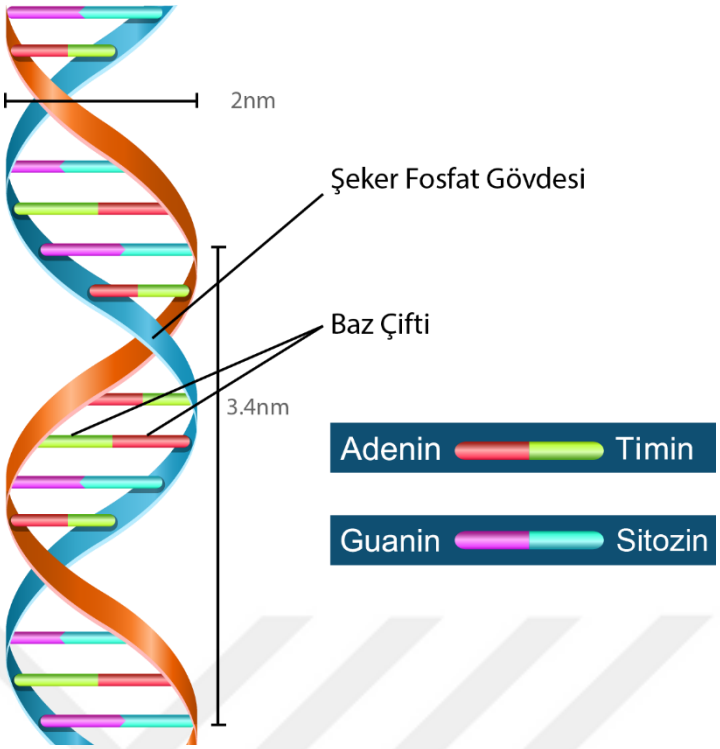
3.1.3. DNA

DNA veya deoksiribonükleik asit, insanlarda ve hemen hemen tüm diğer organizmalarda kalıtsal maddedir. Bir insanın vücudundaki neredeyse her hücre aynı DNA'ya sahiptir. Çoğu DNA, hücre çekirdeğinde bulunur (nükleer DNA olarak adlandırılır), ancak mitokondride de (mitokondriyal DNA veya mtDNA olarak adlandırılır) az miktarda DNA bulunabilir. Mitokondri, hücreler içindeki enerjiyi gıdalardan hücrelerin kullanabileceği bir forma dönüştüren yapılardır.

DNA'daki bilgiler dört kimyasal bazdan oluşan bir kod olarak saklanır: *adenin* (A), *guanin* (G), *sitozin* (C) ve *timin* (T). İnsan DNA'sı yaklaşık 3 milyar bazdan oluşur ve bu bazların yüzde 99'undan fazlası tüm insanlarda aynıdır. Bu bazların sırası veya konumu, bir organizmanın oluşturulması ve sürdürülmesi için mevcut olan, alfabedeki harflerin kelimeler ve cümleler oluşturmak için belirli bir sırada görünme şekline benzer bilgileri belirler.

DNA bazları, baz çiftleri olarak adlandırılan birimler oluşturmak için birbirleriyle, A ile T ve C ile G'yi birleştirir. Her baz ayrıca bir şeker molekülüne ve bir fosfat molekülüne bağlanır. Birlikte, bir baz, şeker ve fosfata bir nükleotid denir. Nükleotidler, çift sarmal adı verilen bir spiral oluşturan iki uzun iplik halinde düzenlenir. Çift sarmalın yapısı bir merdiven gibidir, taban çiftleri merdivenin basamaklarını ve şeker ve fosfat molekülleri merdivenin dikey yanlarını oluşturur.

DNA'nın önemli bir özelliği, çoğaltılabilmesi veya kendi kopyalarını (replikasyon) oluşturabilmesidir. Çift sarmaldaki her DNA ipliği, baz dizisini çoğaltmak için bir desen görevi görebilir. Hücreler bölündüğünde bu önemlidir, çünkü her yeni hücrenin eski hücrede bulunan DNA'nın tam bir kopyasına sahip olması gerekir. Şekil 3.1’de bir DNA yapısı gösterilmiştir.



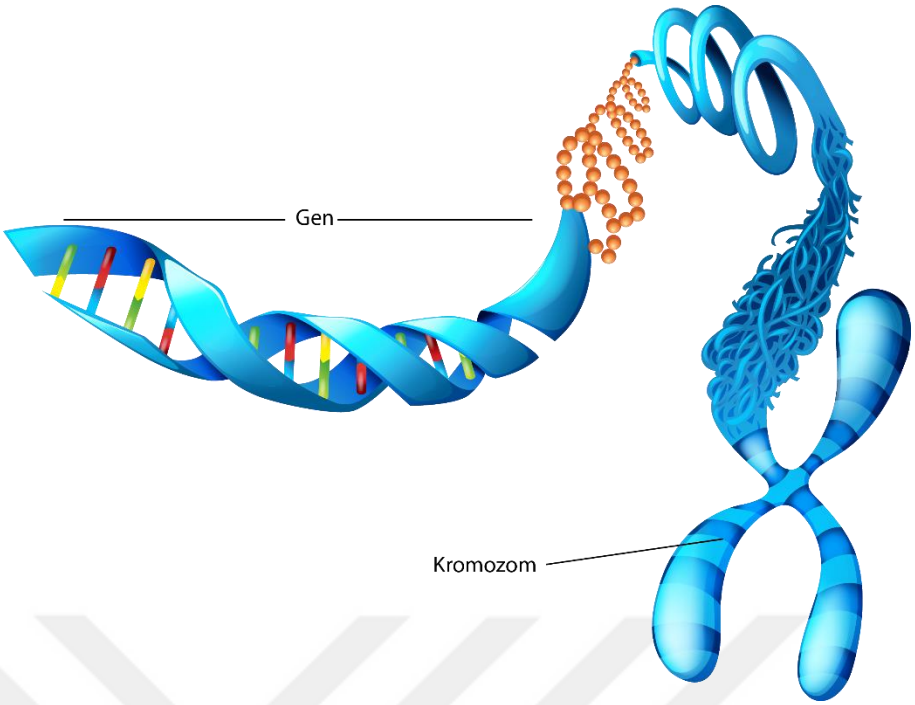
Şekil 3.1. Bir DNA'nın yapısı

3.1.4. Gen

Gen, kalıtımın temel fiziksel ve fonksiyonel birimidir. Genler DNA'dan oluşur. Bazı genler protein adı verilen molekülleri yapmak için komut görevi görmekle birlikte, bazı genler proteinleri kodlamaz. İnsanlarda, genlerin büyüklüğü birkaç yüz ile 2 milyon bazdan fazla bir aralıkta DNA bazı içerebilmektedir. İnsan Genom Projesi, insanların 20,000 ile 25,000 arasında gene sahip olduklarını tahmin etmiştir.

Her insanda, her bir ebeveyninden miras alınan her genin iki kopyası vardır. Çoğu gen tüm insanlarda aynıdır, ancak az sayıda gen (toplamın yüzde 1'inden az) insanlar arasında biraz farklılık göstermektedir. Aleller DNA genleri dizilerinde küçük farklılıklar gösteren aynı genin formlarıdır. Bu küçük farklılıklar her bireyin kendine özgü fiziksel özelliklerine katkıda bulunur.

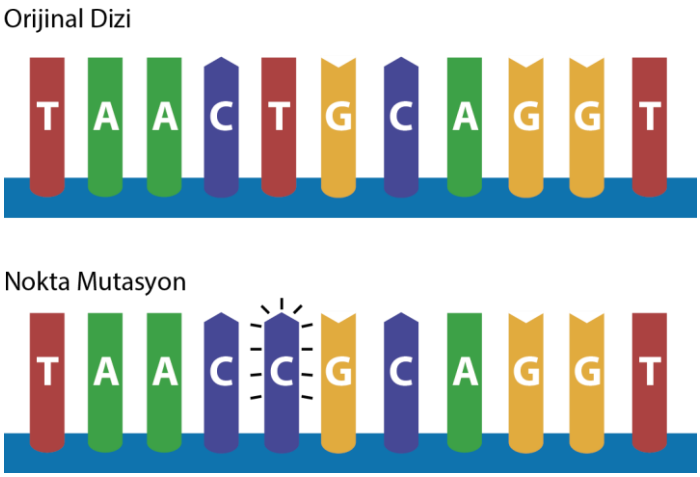
Bilim adamları genlere benzersiz isimler vererek takip ederler. Gen adları uzun olabileceğinden, genlere, gen adının kısaltılmış bir versiyonunu temsil eden kısa harf kombinasyonları (ve bazen rakamlar) olan semboller de atanır. Şekil 3.2'de bir gen görseli sunulmuştur.



Şekil 3.2. DNA'nın oluşturduğu gen ve kromozom

3.1.5. Mutasyon

Mutasyon, DNA kopyalandığında hatalar nedeniyle veya çevresel faktörlerin sonucu olarak DNA dizisinde meydana gelen değişikliklerdir. Bir canlıda ömür boyu DNA'lar, A, G, C ve T bazında değişikliklere ya da "mutasyonlara" uğrayabilir. Bu, DNA'nın kodladığı proteinlerde değişikliklere neden olur. Hatalar oluşur ve zamanında tamir edilmezse DNA replikasyonu sırasında mutasyonlar meydana gelebilir. Mutasyonlar ayrıca sigara, güneş ışığı ve radyasyon gibi çevresel faktörlere maruz kalmanın bir sonucu olarak da ortaya çıkabilir. Genellikle hücreler mutasyona neden olabilecek herhangi bir hasarı tanıyabilir ve sabit bir mutasyon haline gelmeden önce onarabilir. Mutasyonlar türler içindeki genetik çeşitliliğe katkıda bulunur. Özellikle olumlu bir etkiye sahiplerse miras alınabilir. Mutasyonlar boyut olarak değişiklik gösterebilir. Tek bir DNA baz çiftinden (nokta mutasyon), birden fazla gen içeren bir kromozomun büyük bir segmentine kadar (gen mutasyonu) herhangi bir yeri etkileyebilirler. Şekil 3.3'te bir nokta mutasyon örneği gösterilmiştir.



Şekil 3.3. DNA mutasyonu (yourgenome.org, 2020)

3.1.6. Hizalama

Biyolojik dizilerde hizalama DNA, RNA veya protein dizilerini, diziler arasındaki fonksiyonel, yapısal veya evrimsel ilişkilerin bir sonucu olabilecek benzerlik bölgelerini tanımlamak için düzenlemenin bir yoludur. Hizalanmış nükleotid veya amino asit dizi parçaları tipik olarak bir matris içindeki sıralar olarak temsil edilir. Parçalar arasına boşluklar eklenir, böylece özdeş veya benzer karakterler birbirini izleyen sütunlarda hizalanır. Dizi hizalamaları, doğal bir dilde veya finansal verilerde dizeler arasındaki mesafe maliyetinin hesaplanması gibi biyolojik olmayan diziler için de kullanılmaktadır. Şekil 3.4'te hizalama yapılmış bir DNA parçası örneği verilmiştir.

```
A C T T G T C T T A T G C
A C T _ G _ _ T T A _ _ C
```

Şekil 3.4. Hizalama İşlemi Yapılmış DNA Parçası

Bir hizalamadaki iki dizi ortak bir atayı paylaşıyorsa, uyumsuzluklar nokta mutasyonları ve boşluklar şeklinde indeller (ekleme veya silme mutasyonları) olarak yorumlanabilir. Proteinlerin dizi hizalamalarında, dizide belirli bir pozisyonu işgal eden amino asitler arasındaki benzerlik derecesi, belirli bir bölgenin veya dizi motifinin soylar arasında nasıl korunduğunun kaba bir ölçüsü olarak yorumlanabilir. Dizinin belirli bir bölgesinde substitüsyon (yer değişikliği) olmaması veya sadece çok konservatif substitüsyonların (yani yan zincirleri benzer biyokimyasal özelliklere sahip amino asitlerin substitüsyonu) varlığı, bu bölgenin yapısal veya fonksiyonel öneme sahip

olduğunu düşündürmektedir (Ng ve Henikoff, 2001). DNA ve RNA nükleotid bazları amino asitlerden daha çok birbirlerine benzese de baz çiftlerinin korunması benzer bir fonksiyonel veya yapısal rolü gösterebilir.

Çok kısa veya çok benzer diziler el ile hizalanabilir. Ancak pratikte insan çabasıyla hizalanamayan uzun, çok değişken veya çok sayıda dizinin hizalanması söz konusudur. İnsan bilgisi, yüksek kaliteli dizi hizalamaları üretmek için algoritmaların oluşturulmasında ve zaman zaman nihai sonuçların, algoritmik olarak temsil edilmesi zor olan desenleri (özellikle nükleotid dizileri durumunda) yansıtacak şekilde ayarlanmasında uygulanır. Dizi hizalamaya yönelik hesaplama yaklaşımları genellikle iki kategoriye ayrılır: *global hizalamalar* ve *lokal hizalamalar*. Genel hizalamanın hesaplanması, hizalamayı tüm sorgu dizilerinin tüm uzunluğunu kapsayacak şekilde yapılmaya "zorlayan" bir global optimizasyon biçimidir. Buna karşın lokal hizalamalar, genellikle farklılığı yüksek diziler içindeki benzer bölgeleri tanımlar. Dizi hizalama problemine çeşitli hesaplama algoritmaları uygulanmıştır. Bunlar dinamik programlama gibi yavaş ama biçimsel olarak doğru yöntemleri içerir. Bunlar arasında, en iyi eşleşmeleri bulmayı garanti etmeyen, büyük ölçekli veri tabanı araması için tasarlanmış verimli, sezgisel algoritmalar veya olasılık yöntemleri de bulunur (Earl ve ark., 2014).

3.1.6.1. İkili hizalama

Dizi analizi yöntemlerinin neredeyse tamamında ilk adım, ikili hizalamadır. Çoğu dizi hizalama yöntemi benzerlik kriterlerini optimize etmeye çalışır. Bu benzerliği değerlendirmenin iki şekli vardır: lokal ve global. Lokal yöntemler, bir A dizisinin alt dizilerinin başka bir B dizisinde mevcut olup olmadığını belirlemeye çalışır. Bu yöntemlerin veri tabanı arama ve erişiminde büyük faydaları vardır (ör. BLAST, Altschul et al., 1990). Belirli bir benzerlik derecesine sahip dizilerin tespitinde faydalı olmasına rağmen, homolog (türdeş) olsun ya da olmasın, filogenetik analizde karşılaştırılan dizilerin ortolog[1] olduğu varsayılmaktadır. Global yöntemler, dizilerin tam uzunlukları üzerinde karşılaştırmalar yapar; diğer bir deyişle, A dizisinin her bir elemanı, B dizisindeki her bir elamanla karşılaştırılır. Global karşılaştırma, filogenetik analiz için başlıca hizalama yöntemidir (Phillips ve ark., 2000).

[1] Ortolog genler, farklı türlerde bulunan ancak aynı atadan geldikleri için birbirleriyle benzer olan genlerdir.

Benzerlik maksimizasyonunun temel noktası, iki dizi arasındaki minimum düzenleme mesafesinin hesaplanmasıdır. Düzenleme mesafesi, A dizisini B dizisine dönüştürmek için gereken işlemlerin sayısıdır (substitüsyonlar, eklemeler veya silmeler). Her işleme bir maliyet (veya ceza) verilmelidir. Toplam optimal maliyet, iki dizinin benzerliğini yansıtan bir ölçü olacaktır. İkili dizi hizalamanın temel yöntemi ilk olarak Needleman ve Wunsch (1970) tarafından tanımlanmıştır. Needleman – Wunsch (N-W) yöntemi başlangıçta proteinler için tasarlanmıştır, ancak herhangi bir ikili düzenleme mesafesi problemine uygulanabilmektedir. Prosedür, iki dizi arasındaki benzerlik ölçüsünü maksimize etmeye çalışır. Smith ve Waterman (1981), bir mesafe metriğini en aza indiren Seller metriğinin (Sellers, 1974) N-W algoritmasına eşdeğer olduğunu göstermiştir. Bu algoritmalar, daha küçük alt problemleri özyinelemeli olarak çözüp nihai bir küresel sonuca birleştirerek daha büyük bir sorunun çözülmesine izin veren dinamik programlamaya (Eddy, 2004) bir örnektir.

3.1.6.2. Çoklu hizalama

Çoklu dizi hizalaması (ÇDH), üç veya daha fazla biyolojik dizinin (protein, DNA veya RNA) dizi hizalamasıdır. Çoğu durumda, sorgu dizilerinden oluşan girdi kümesinin, bir bağlantıyı paylaştıkları ve ortak bir atadan türetildikleri evrimsel bir ilişkiye sahip oldukları varsayılır. Elde edilen ÇDH'den dizi homolojisi çıkarılabilir ve dizilerin paylaşılan evrim kökenlerini değerlendirmek için filogenetik analiz yapılabilir. Tek bir hizalama sütununda farklı karakterler olarak görülen nokta mutasyonları, hizalamadaki bir veya daha fazla dizideki tireler ekleme veya silme mutasyonları (indeller veya boşluklar) gibi olaylarını gösterir. Çoklu dizi hizalaması genellikle protein alanlarının, ikincil ve üçüncül yapıların ve hatta münferit amino asitlerin veya nükleotidlerin dizi korunumunu değerlendirmek için kullanılır. Çoklu dizi hizalaması aynı zamanda böyle bir dizi setini hizalama işlemini de ifade eder. Biyolojik olarak ilgili uzunluktaki üç veya daha fazla dizi zor olabileceğinden ve elle hizalamak yüksek zaman ve işlem maliyetine sahip olduğundan, hizalamaları üretmek ve analiz etmek için hesaplamalı metotlar içeren algoritmalar kullanılır. ÇDH'ler, hesapsal açıdan maliyetli oldukları için ikili hizalamadan daha karmaşık yöntemlere ihtiyaç duymaktadır. Birçok çoklu dizi hizalama aracı, global optimizasyon yerine sezgisel yöntemler kullanır, çünkü birkaç orta uzunlukta diziden daha fazlası arasındaki optimum hizalamanın belirlenmesi, hesaplama açısından maliyetlidir (Lakshmi ve ark., 2016).

Çoklu dizi hizalamada kesin, aşamalı ve yinelemeli hizalama yaklaşımları olmak üzere üç yaklaşım kategorisi sıklıkla kullanılmaktadır. Kesin algoritmalar genellikle optimum seviyeye çok yakın sonuçlar sağlar (Lipman ve ark., 1989). Aynı anda birden fazla diziyi hizalamaya çalışır ve bu nedenle dinamik programlamaya bağlı olması gerekir. Yukarıda belirtildiği gibi hizalamasındaki kesin yöntemin dezavantajı nedeniyle, çoğu çoklu dizi hizalama metodu diğer iki kategori kapsamındadır.

3.1.6.2.1. Aşamalı hizalama

Aşamalı hizalama yaklaşımını ilk olarak Hogeweg ve Hesper (1984) literatüre kazandırmıştır. Bu yaklaşım, karmaşık ÇDH probleminin alt problemlere ayrıldığı bir sezgisel tarama yaklaşımıdır. Bu, dolaylı olarak ikili hizalama kullanarak doğrudan MSA sorununu çözer. En iyi ikili hizalama, ilk önce dikkate alındığı tüm dizileri aşamalı olarak birleştirir. Aşamalı hizalama, her yaprağın hizalanacak bir diziyi temsil ettiği ÇDH problemini çözmek için bir *kılavuz ağaç* kullanır. Kılavuz ağaç, karşılaştırması yapılacak diziler arasında tüm ikili kombinasyonlara göre benzerlik hesaplaması yaparak bir ağaç oluşturur ve çoklu hizalamada dizilerin hangi sırayla hizalama listesine ekleneceğini belirler. Ziyaret edilen her iç düğüm, karşılık gelen alt ağacındaki dizilerin bir ÇDH'si ile ilişkilidir. Son olarak, dikkate alınan tüm dizilerin ÇDH'si kök düğümü ile ilişkilidir. En bilinen aşamalı çoklu dizi hizalama metotları arasında ClustalW (Thompson ve ark., 1994), T-Coffee (Notredame ve ark., 2000), KAlign (Lassmann ve Sonnhammer, 2005) sayılabilmektedir. Bu yaklaşımın açgözlü doğası, boşlukların değiştirilmesine izin veremez ve bu nedenle, sonraki aşamada hizalama değiştirilemez. Bu açgözlülük için yerel minimuma takılma olasılıkları vardır (Thompson ve ark., 2005). Diğer bir dezavantaj, aşamalı ÇDH'nin ilk hizalamadan etkilenmesidir. Sonuç olarak, bu aşamada yapılan herhangi bir hata nihai ÇDH sonuçlarına yayılır. Öte yandan, yinelemeli hizalama yöntemleri, dizileri veya dizi gruplarını yeniden düzenleyerek hizalamayı yinelemeli olarak değiştirir ve böylece aşamalı yöntemin dezavantajının üstesinden gelir (Chowdhury ve Garai, 2017).

3.1.6.2.2. Yinelemeli hizalama

Yinelemeli hizalama yaklaşımı genellikle ardışık yöntemlerle, yapılan hizalamada değişiklikler yaparak işlem sonrası müdahaleler gerçekleştirir. Kılavuz ağacın yapısını

değiştirir. Kılavuz ağaç tahminini güncelleyen bu yaklaşımı benimseyen başlıca algoritmalar olarak MAFFT (Katoh ve ark., 2002), MUSCLE (Edgar, 2004), PRIME (Yamada ve ark., 2006) ve PRRP (Gotoh, 1996) sayılabilir. Aşamalı hizalama ile elde edilen bir ÇDH kullanarak yeni mesafe matrislerini hesaplarlar. Bu da ikinci bir aşamalı hizalamaya yol açan yeni bir kılavuz ağacı oluşturur. Bazı yinelemeli yöntemler, hizalanmış dizileri art arda iki gruba bölerek ve daha sonra hizalama işlemi yakınlaşana kadar bu grupları yeniden düzenleyerek gerçekleştirir (Chowdhury ve Garai, 2017).

3.2. Özellik Oluşturma ve Özellik Vektörü

Makine öğrenimi ve örüntü tanımada bir özellik, gözlemlenebilecek bir fenomenin bireysel olarak ölçülebilir özelliğidir. Veri toplamak ve işlemek maliyetli ve zaman alıcı bir işlem olabilmektedir. Bu nedenle, bilgilendirici, ayırt edici ve bağımsız özelliklerin seçilmesi, örüntü tanıma, sınıflandırma ve regresyonda etkili algoritmalar için çok önemli bir adımdır. İçinde yaşadığımız bilgi çağında, veri kümesi boyutları hem örnek sayısı hem de özellik sayısı bakımında giderek büyümektedir. On binlerce veya daha fazla özelliğe sahip veri kümelerini görmek oldukça yaygın hale gelmiştir. Ayrıca, algoritma geliştirici uzmanlar çoğu kez özellik oluşturma (feature generation) adı verilen bir işlem kullanmaktadır. Özellik oluşturma, istatistiksel veya yapay zekâ sistemlerinde yapılan analizde potansiyel kullanım için bir veya daha fazla özellikten yeni özellikler oluşturma işlemidir. Genellikle bu süreç modele yeni bilgiler ekler ve daha doğru hale getirir. Özellik oluşturma, bir özellik etkileşimi olduğunda model doğruluğunu artırabilir. Özellik etkileşimini kapsayan yeni özellikler ekleyerek, tahmin modeli için yeni bilgilere erişilebilir hale getirmektedir (Cohen, 2019).

Özellik çıkarımında kullanılan yaygın veri türlerinden dizi verileri insanların günlük yaşamlarında her yerde bulunur ve genellikle veri analizinde kullanılır. Biyotıp, müzik, edebiyat ve sağlık hizmetleri gibi çeşitli uygulama türlerinde dizi verilerinden bahsedebiliriz. Büyük bilgisayar şirketlerinin, hangi olay kombinasyonlarının bilgisayar çökmelerine yol açabileceğini belirlemek için yazılan bilgisayar günlüğü girişlerini incelemesi, siber güvenlik ve internette gezinme işlemlerinin optimizasyonu, müzik verileri analizi için müzik parçalarının sembolik gösterimi ile hangi tür müzik parçalarının sıklıkla mutluluğa veya üzüntüye neden olabileceğini belirlemek, biyolojik dizi analizi için önemli bölgelerin belirlenmesi, dizi veri setlerinde yapılabilecek uygulamalara örnekler olarak sayılabilir.

Dizi analizi için çoğu zaman diziler, özellik vektörleri olarak temsil edilmelidir. Diğer veri tipleriyle karşılaştırıldığında, dizi verilerinin ayırt edici özellikleri şunları içermelidir (Dong ve Liu, 2018):

- Dizi yapısındaki verilerin uzunlukları çeşitlilik gösterebilir. Çok kısa veriler olabileceği gibi çok uzun yapıda verilerden de bahsedilebilir. Örneğin DNA/protein dizisinde on binlerle belirtilen uzunluklarla çalışılabilir.
- Bir dizideki elemanların sırası sabittir. Başka bir deyişle, bir dizideki iki farklı öğenin konumlarını değiştirmek, farklı bir dizi üretir. Örneğin, bir cümledeki iki farklı kelimenin konumunu değiştirirsek, tüm cümle değişebilmektedir. Bu da cümlenin anlamını önemli ölçüde değiştirebilir.
- Elemanların dizilerdeki konumu çok önemli olabilir. Örneğin, DNA'daki bir gen başlangıç bölgesinin promotör bölgesi, ilgili genin transkripsiyon başlangıç bölgesine belirli mesafedeki yakınlıkta bir pozisyonda bulunur.

Dizi verilerinin benzersiz özellikleri ve önemi nedeniyle, dizi verileri madenciliği çok sayıda araştırmacının dikkatini çekmiştir ve diğer veri türlerinden oldukça farklı bir sonuç ortaya çıkarmıştır. Desenler biyolojik veri analizinde sıklıkla "motif" olarak adlandırılmıştır.

3.3. Dijital Görüntülerde Doku Analizi

3.3.1. Dijital görüntü

Dijital görüntü, her biri x ekseni ve y ekseni üzerindeki (x, y) ile ifade edilmiş uzamsal koordinatlarının girdi olarak verildiği iki boyutlu işlevlerin bir çıktısı olan gri seviyesi veya yoğunluğu için sonlu ve ayrık miktarlardaki, piksel olarak adlandırılan resim öğelerinden oluşur. Yükseklik ve genişlik değerlerini piksellerin oluşturduğu matris formatındaki vektörlerle ifade edilir.

3.3.1.1. Gri-seviye görüntü

Gri-seviyeli veya gri tonlamalı görüntü, her pikselin değerinin yalnızca bir miktar ışığı temsil eden tek bir örnek olduğu bir görüntüdür. Sadece yoğunluk bilgisi taşır. Bir tür siyah beyaz veya gri monokrom olan gri tonlamalı görüntüler yalnızca gri tonlarından

oluşur. Kontrast, en zayıf yoğunlukta siyahtan en güçlüde beyaza değişir. n bit sayısı olmak üzere, görüntüdeki piksellerin yoğunluğunu ifade edecek sayı aralığı $0 - 2^n$'dir.

3.3.1.2. Renkli görüntü

Renkli görüntü, her piksel için renk bilgisi içeren dijital bir görüntüdür. Görsel olarak kabul edilebilir sonuçlar için, her piksel için bazı renk uzaylarında koordinat olarak yorumlanan örnekler (renk kanalı) sağlamak gereklidir. Üç kanallı RGB (Red-Green-Blue / Kırmızı-Yeşil-Mavi) renk uzayı bilgisayar ekranlarında yaygın olarak kullanılır. Renkli bir görüntünün piksel başına üç değeri (veya kanalı) vardır ve ışığın yoğunluğunu ve renkliliğini ölçer. Dijital görüntü verilerinde saklanan gerçek bilgiler, her spektral banttaki parlaklık bilgisidir.

3.3.2. Doku

Doku, görüntü bölgesinin bir özelliğidir. Görsel alandaki bölgeler dokudaki parlaklık, renk veya diğer niteliklerin farklılıkları ile karakterize edilebilir. Bir görüntüde bulunan bilgi deseni veya yapının düzenlemesi olan doku, birçok görüntü türünün önemli bir özelliğidir. Genel bir tanımlamayla doku, görüntünün temel parçalarının şekli, boyutu, düzeni, yoğunluğu ve oranı ile tanımlanan bir nesnenin yüzey karakteristiğini ve görünümünü ifade etmektedir (Shapiro ve Stockman, 2001). Görsel sistemdeki ilk işlemler, sonraki hesaplama aşamalarında işlem yükünü hafifletmek için görüntünün bölgelere geçici bir segmentasyonunu gerçekleştirmek için doku bilgilerini kullanabilir. Tek bir dokulu görüntü bölgesinin analizi, o bölge için kategorik etiketlerin algılanmasına yol açabilir ("Bu ahşaptır" veya "Bu yüzey kaygandır" gibi). Dokunun görünümü, gözlemcinin iki dokulu bölgenin aynı veya farklı şeylerden yapılmış gibi görünmesini belirlemesini sağlar. İki bitişik görüntü bölgesi farklı yüzey dokusuna sahipse bu, araya giren doku sınırının saptanmasını sağlayabilir. Bu tür doku tanımlı sınırlar daha sonra şekli zeminden ayırmak ve 2 boyutlu şekil tanımlaması için kullanılabilir. Bu alanda birçok araştırmanın odağı, dokuyu karakterize etmek için kullanılan mekanizmaları ve temsili şemaları karakterize etmek ve böylece yukarıdaki algısal yeteneklerin her birinden aynı temel mekanizmaların sorumlu olup olmadığını tespit etmektir (Landy ve Graham, 2004).

3.3.2.1. Doku analizi

Doku analizi, tıbbi görüntüleme, uzaktan algılama, endüstriyel kontrol, içerik tabanlı görüntü erişimi gibi birçok alanda önemli bir rol oynamıştır ve görevleri temel olarak sınıflandırma, segmentasyon ve sentezdir. Bilgisayarla görmede doku araştırmalarının temel hedefi dokuyu anlamak, modellemek ve işlemektir. Dokuyu analiz etme yaklaşımları çok çeşitlidir ve esas olarak dokusal özelliklerin çıkarılması için kullanılan yönteme göre birbirinden ayrılırlar. Bu çerçevede tanımlanabilecek dört kategori; istatistiksel metotlar, yapısal metotlar, model tabanlı metotlar ve dönüşüm tabanlı metotlardır (Bharati ve ark., 2004).

3.3.2.2. Doku analizinde istatistiksel metotlar

İstatistiksel yöntemler, görüntünün her bir noktasında yerel özellikleri hesaplayarak ve yerel özelliklerin dağılımlarından bir dizi istatistik türeterek gri değerlerin uzamsal dağılımını analiz eder (Srinivasan ve Shobha, December 2008). Bu teknik bilgisayarlı görmede ilk yöntemlerden biridir. Yerel özelliği tanımlayan piksel sayısına bağlı olarak, istatistiksel yöntemler birinci dereceden (bir piksel), ikinci dereceden (piksel çifti) ve daha yüksek dereceli (üç veya daha fazla piksel) istatistiklere göre sınıflandırılabilir. Bu sınıflar arasındaki fark, birinci dereceden istatistiklerin, görüntü pikselleri arasındaki uzamsal etkileşimi bırakarak tek tek piksel değerlerinin özelliklerini tahmin etmesidir. İkinci ve daha yüksek dereceli istatistiklerde ise, birbirine göre belirli yerlerde meydana gelen iki veya daha fazla piksel değerinin özelliklerini tahmin eder (Materka ve Strzelecki, 1998).

3.3.2.2.1. Birinci derece istatistiksel metotlar

Birinci dereceden doku analizi (veya histogram analizi) gri-seviyeli bir görüntüde piksel yoğunluğu değerlerini çıkarır (Alobaidli ve ark., 2014). Histogramı hesaplayan bu değerler üzerinden görüntünün birinci dereceden istatistik özelliklerini tanımlamak için farklı parametreler kullanılır. (Levine, 1985; Pratt, 1991). Buna göre $x = 0, 1, \dots, N - 1$ ve $y = 0, 1, \dots, M - 1$ olmak üzere; görüntünün iki uzay değişkeni x ve y 'nin bir $f(x, y)$ fonksiyonu olduğunu varsayalım. $f(x, y)$ fonksiyonu, G görüntüdeki maksimum yoğunluk seviyesi olmak üzere, ayrık $i = 0, 1, \dots, G - 1$ değerlerini alabilir. Yoğunluk

seviyesi histogramı, tüm görüntüdeki yoğunluğa sahip piksel sayısını (her yoğunluk seviyesi için) gösteren bir fonksiyondur. Bu fonksiyon Denklem 3.1'de verilmiştir:

$$h(i) = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \delta(f(x,y), i) \tag{3.1}$$

Burada $\delta(j,i)$ Kronecker delta fonksiyonudur ve Denklem 3.2' de tanımlanmıştır:

$$\delta(j,i) = \begin{cases} 1, & eğer\ j = i \\ 0, & eğer\ j \neq i \end{cases} \tag{3.2}$$

Yoğunluk seviyelerinin histogramı, görüntüdeki istatistiksel bilgilerin kısa ve basit bir özetidir. Gri-seviye histogramın hesaplanması tekil pikselleri içerir. Bu nedenle, histogram, görüntü veya bir parçası hakkında birinci dereceden istatistiksel bilgileri içerir. $h(i)$ değerlerinin görüntüdeki toplam piksel sayısına bölünmesi, yoğunluk seviyelerinin meydana gelme olasılığını (Denklem 3.3) yaklaşık olarak verir (Materka ve Strzelecki, 1998).

$$p(i) = \frac{h(i)}{NM}, \qquad i = 0, 1, \ldots, G-1 \tag{3.3}$$

Histogramı hesaplayan değerler üzerinden görüntünün birinci dereceden istatistik özelliklerini tanımlamak için parametreler olarak kullanılan metriklere ait denklemler Denklem 3.4 – 3.9' da (Materka ve Strzelecki, 1998) verilmiştir:

$$\mu = \sum_{i=0}^{G-1} ip(i) \tag{3.4}$$

$$\sigma^2 = \sum_{i=0}^{G-1} (i-\mu)^2 p(i) \tag{3.5}$$

$$\mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i-\mu)^3 p(i) \tag{3.6}$$

$$\mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i-\mu)^4 p(i) - 3 \quad (3.7)$$

$$E = \sum_{i=0}^{G-1} [p(i)]^2 \quad (3.8)$$

$$H = -\sum_{i=0}^{G-1} p(i) log_2[p(i)] \quad (3.9)$$

Denklem 3.4'te verilen *ortalama* değeri, görüntünün veya dokunun ortalama yoğunluk seviyesini ifade eder. Denklem 3.5'te verilen *Varyans*, ayrıca ortalamadaki yoğunluğun varyasyonunu da tanımlar. Histogram, ortalamanın etrafında simetrik ise *çarpıklık* değeri (Denklem 3.6) sıfırdır, aksi takdirde ortalamanın altında veya üstünde olmasına bağlı olarak pozitif veya negatif olabilir. Buna göre, μ_3 simetrinin bir göstergesidir. *Basıklık* (Denklem 3.7), histogramın düzlüğünün bir ölçüsüdür. Denklem 3.7 içine yerleştirilen "-3" bileşeni, Gauss şeklindeki bir histogram için μ_4'ü sıfıra normalleştirir. *Enerji* (Denklem 3.8), histogram homojenliğinin bir ölçüsüdür. Son olarak, bir görüntünün *Entropisi* (Denklem 3.9) rastgelelik tahminidir ve dokusunu ölçmek için sıklıkla kullanılır. Entropi, tanımlanmış yapısal bilgilerle doğrudan ilişkili olan histogram piklerinin keskinliğinin bir ölçümü olarak düşünülebilir (Vazquez-Fernandez ve ark., 2010)

3.3.2.2.2. İkinci derece istatistiksel metotlar

Birinci dereceden istatistiklerle oluşturulan özellikler, görüntünün gri düzeyindeki dağılımı ile ilgili bilgi sağlar. Ancak, görüntüdeki çeşitli gri düzeylerin bağıl konumları hakkında herhangi bir bilgi vermezler. Bu özellikler, tüm düşük değerli gri düzeylerin birlikte konumlandırılıp konumlandırılmadığını veya yüksek değerli gri düzeylerle değiştirilip değiştirilmediğini ölçmez. Bazı gri seviye konfigürasyonların bir oluşumu, bağıl $P_{\theta,d}(I_1, I_2)$ frekansları matrisi ile tarif edilebilir. Gri düzeylerde I_1, I_2 olan iki pikselin θ yönündeki d mesafesiyle ayrılmış pencerede ne sıklıkta göründüğünü açıklar. Bilgiler, piksellerin çiftler halinde değerlendirildiği ikinci dereceden görüntü istatistiklerini ölçen *eş oluşum* matrisinden elde edilebilir. Eş oluşum matrisi iki parametrenin bir fonksiyonudur: (d) piksel sayısında ölçülen bağıl mesafe ve bunların

bağıl yönü θ. θ yönü yatay, köşegen, dikey ve ters köşegen yönlerini ifade eden sırasıyla $0^o, 45^o, 90^o$ ve 135^o'lik açılarla nicelendirilmiştir (Aggarwal ve Agrawal, 2012).

3.4. Kelime Tabanlı Dizi Benzerlik Analizi

Metin analizlerinde kullanılan kelime tabanlı algoritmalar benzer DNA dizilerinin benzer kelimeleri paylaştığı öngörüsü ile yola çıkmışlardır. Literatürde alt kelimeler için *n*-gram, *k*-mer, *l*-mer, *k*-tuple gibi isimleri kullanan bu metotların temelindeki benzerlik hesaplaması, alt kelime oluşumları ile yapılan matematiksel işlemlerin, dizi farklılığının ölçüsü olarak ele alınmasıyla yapılmaktadır. Bu yöntemlerin arkasındaki mantık basittir: Benzer diziler benzer kelimeleri paylaşır ve kelimelerin oluşumları ile yapılan matematiksel işlemler dizi farklılığının iyi bir ölçüsünü verir. Yöntem aynı zamanda ilk önce dinükleotid bileşimi için başlatılan ve daha uzun kelimelere genişletilen genomik imzalar fikri ile sıkı bir şekilde birleştirilmiştir. Bu işlem üç temel aşamaya bölünebilir (Zielezinski ve ark., 2017):

i. İlk olarak karşılaştırılan diziler için belirli uzunluklardaki benzersiz kelimelerin koleksiyonları oluşturulur.
ii. İkinci adımda her dizi bir sayı dizisine (vektör) dönüştürülür.
iii. Son adımda ise diziyi temsil eden vektörler arasında bir uzaklık fonksiyonu uygulanır ve diziler arasındaki farklılıklar ölçülür.

3.4.1. N-gram

Kelime tabanlı yöntemler arasında n-gram-tabanlı yaklaşım ilk olarak Blaisdell tarafından önerilmiştir (Blaisdell, 1986). N ve n pozitif tamsayı olmak üzere, bir $\mathcal{A}$ alfabesi üzerinden oluşturulmuş $S = (s_1, s_2, \dots, s_{N+(n-1)})$ simgeler dizisi verildiğinde S dizisinin bir *n*-gramı, ardışık simgelerin *n*-uzunluklu herhangi bir alt dizisidir. S'nin i. *n*-gramı $(s_i, s_{i+1}, \dots, s_{i+(n-1)})$ dizisidir. $\mathcal{A}$ alfabesi üzerinden $(|\mathcal{A}|)^n$ farklı *n*-gram elde edilebilir ($|\mathcal{A}|, \mathcal{A}$ alfabesindeki harf sayısıdır (Tomović ve ark., 2006)). *n*-gramda temel birim bir sesbirim, harf, hece, kelime veya baz çifti olabilir. *n*-gram uygulamaları olarak metin sıkıştırma, yazım hatası tespiti ve düzeltme, optik karakter tanıma, bilgi alma, dil tanıma, eksik sesbirim tahmini, otomatik metin sınıflandırma gibi çalışmalar sayılabilir.

Bir n-gram modeli, n-gramların istatistiksel özelliklerini kullanarak dizileri, özellikle doğal dilleri modeller. N-gramların metin içerisindeki dağılımları ve olasılıkları kullanılarak gerek metin karakterizasyonu ile benzerlikler gerekse de tahmin modelleri tasarlanabilmektedir.

3.4.2. Üst-*k* *n*-gram

N-gram kullanılan doğal dil işleme (NLP) çalışmalarında, dize tahminlerini işlemek için n-gram metin sinopsisleri kullanılır. Bir *n*-gram sinopsisi, belirli n-gramları "atlayarak" kelimeleri özetler. Bu nedenle bir sinopsis, verilerde meydana gelen tüm olası *n*-gramların bir alt kümesini oluşturur. Basit bir strateji olarak, verilerden rastgele n-gram örnekleri seçmek kullanılabilir. Diğer bir yaklaşım ise, bir üst-*k* *n*-gram sinopsisi oluşturmaktır (Wagner ve ark., 2014).

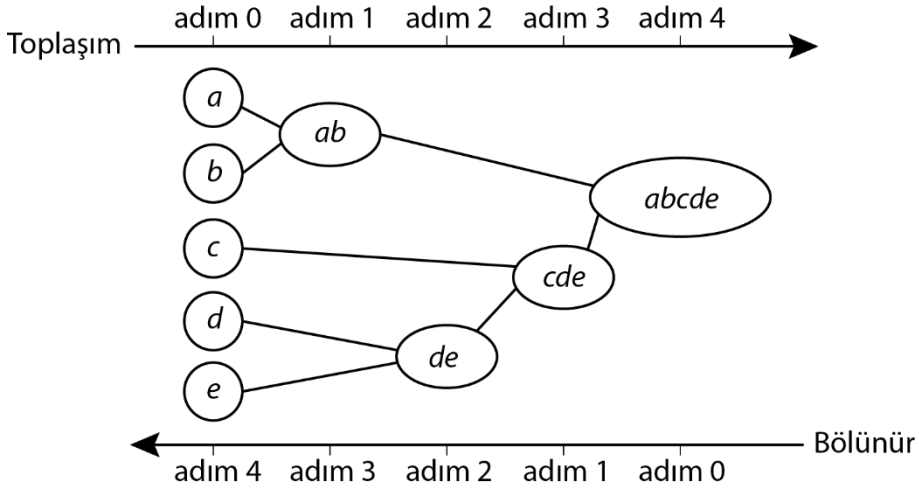
Oluşturulmuş doküman kitaplığı *S*'nin üst-*k* *n*-gram sinopsisi, *S*'deki en yüksek frekansa sahip *n*-gram kümesinden oluşan bir {*n*-gram, sayı} haritasıdır. Üst-*k* *n*-gram sinopsisi $\emptyset_{\text{üst}-k}$'yi oluşturmak için gerekli işlemler çok basittir.

- Tüm benzersiz *n*-gramlar ve sayıları *S* kitaplığından belirlenir;
- *n*-gramlar, azalan sırayla sıralanır;
- En yüksek frekansa sahip *n*-gramlar, *S* doküman kitaplığının büyüklüğüne göre hesaplanan *k* miktara (yüzde veya adet) erişilene kadar $\emptyset_{\text{üst}-k}$'ye eklenir ve tahmin veya benzerlik modeli bu kısıtlanmış *n*-gram listesinden oluşturulur (Wang ve ark., 2011).

3.5. Hiyerarşik Kümeleme

Kümeleme analizi veya kümeleme, istatistiksel veri analizi ve keşifsel veri madenciliğinde kullanılan önemli bir prosedürdür. Bir grup veri noktasını aynı gruptaki üyelerin, farklı gruplardaki üyelere daha çok benzeyecek şekilde gruplama sürecini ifade eder. Küme analizi makine öğrenimi, örüntü tanıma, bilgi alma, görüntü işleme ve biyoinformatik gibi birçok bilgi keşif alanında yaygın olarak kullanılmaktadır. Hesaplamalı biyoloji ve biyoinformatik bağlamında, evrim biyolojisinde homolog genetik dizileri gen ailelerinde gruplandırmak; transkriptomide, ilgili ifade desenleri olan grup genlerine veya ekolojide heterojen ortamlardaki organizma topluluklarını

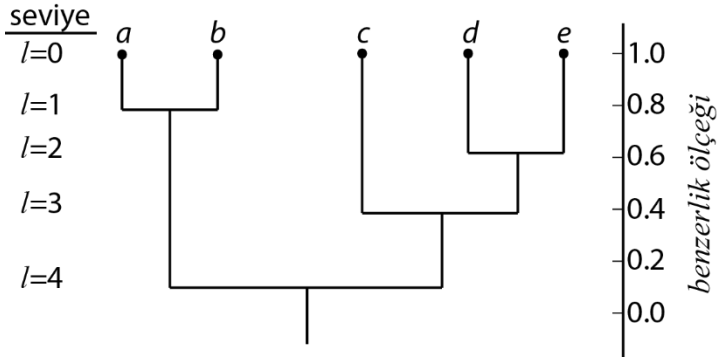
tanımlamak için başarıyla uygulanmıştır. Kümeleme analizi algoritmaları temel olarak iki açıdan farklılık gösterir: küme kavramı ve algoritmanın etkinliği. Popüler küme tanımları, üyeler arasında yüksek düzeyde benzerliklere sahip grubu, veri alanındaki yakın paketlenmiş bölgeleri veya belirli istatistiksel dağılımları içermektedir. Bu nedenle, kümelenme çok amaçlı bir optimizasyon problemi olarak formüle edilebilir (Taboada ve Coit, 2007). Kümeleme algoritmalarının seçimi ve ilgili parametre ayarları (örneğin mesafe fonksiyonu, beklenen küme sayısı, yoğunluk eşiği) giriş veri kümesinin özelliklerine ve sonuçların amaçlanan kullanımına bağlıdır. Kümeleme teknikleri genellikle iki gruba ayrılır: ürettikleri küme yapısına bağlı olarak hiyerarşik kümeleme grubu ve bölümlemeli kümeleme grubu. Her yaklaşımın kendi artıları ve eksileri vardır ve farklı yaklaşımlar verilere farklı bakış açıları sağlar. Hiyerarşik kümeleme algoritması, evrimsel biyoloji çalışmalarında genetik veri kümelerinin analizi için özellikle yararlıdır. Çünkü bu veri kümelerinde ilgili organizmalardan çıkarılan genetik diziler arasında doğal hiyerarşik ilişkiler vardır (Nguyen ve Kwoh, 2015). Toplaşım (agglomerative) kümeleme ve bölünür (divisive) kümeleme şeklinde yaklaşımları olan hiyerarşik kümelemede sırasıyla, her elemanın tek küme olarak ele alınması ve mesafe değerlerine göre birleştirilerek tek küme oluşturması ve tüm elemanları tek küme olarak ele alarak daha sonra yine uzaklıklarına göre ayırarak bireysel kümelere dönüştürmesi söz konusudur. Şekil 3.5'te toplaşım ve bölünür hiyerarşik algoritmaların $\{a, b, c, d, e\}$ den oluşan 5 nesne için bir uygulaması gösterilmiştir.



Şekil 3.5. $\{a, b, c, d, e\}$ veri nesneleri için toplaşım ve bölünür hiyerarşik kümeleme

Hiyerarşik kümeleme yöntemi sonucunda elde edilen küme ilişkilerinin gösterildiği ağaç yapısı dendrogram olarak adlandırılır ve evrimsel genetik biyolojide

kullanılan filogenetik ağaç ile aynı gösterim şekline sahiptir. Nesnelerin nasıl adım adım gruplandığını gösterir. Şekil 3.6'da, Şekil 3.5'te verilen beş nesne için bir dendrogram gösterilmiştir.



Şekil 3.6. $\{a, b, c, d, e\}$ veri nesneleri için oluşturulmuş dendrogram

Uzaklık ölçümleri yapılarak birbirine uzaklıklarına göre oluşturulan benzerlik matrisinden dendrogramın elde edilmesi konusunda değişen yaklaşımlara göre çeşitli hiyerarşik kümele yöntemleri bulunmaktadır.

3.5.1. Benzerlik hesaplama

Bir veri setindeki elemanlar kümelenirken benzerlik kavramından faydalanılır. Veri setindeki her bir elemanın diğer bir elemanla olan benzerliğinin veya her bir elemanın veri setindeki diğer elemanlardan olan uzaklığının tespiti şeklinde ifade edilebilir. Söz gelimi 10 elemandan oluşan bir veri setindeki elemanlar başlangıçta 10 farklı kümeye ayrılmışsa, bu 10 farklı kümenin farklı özelliklere sahip olduğunun belirlenmesi gerekmektedir. Bu da kümeler arasında bir uzaklık ölçümü ile gerçekleştirilebilir. Birbirinden çok farklı olmayan elemanlar aynı küme içerisinde ifade edilebilir. Bu işlem, tüm elemanların taranarak veya veri setinin taranması sırasında benzerlik ve mesafe ölçümlerinin yapıldığı sırada yapılabilir. Literatürde benzerlik hesaplamasında kullanılan çok sayıda uzaklık ölçüm metriği bulunmaktadır (Dahal, 2015).

3.5.1.1. Öklid mesafesi

Matematikte Öklid mesafesi veya Öklid metriği olarak adlandırılan, Öklid uzayındaki iki nokta arasındaki "sıradan" düz çizgi mesafesidir. Bu mesafe ile Öklid uzayı

metrik bir alan haline gelir. İlişkili norma Öklid normu denir. p ve q noktaları arasındaki Öklid mesafesi, onları birleştiren çizgi parçasının uzunluğudur ($\overline{pq}$). Kartezyen koordinatlarda $p = (p_1, p_2, \dots, p_n$ ve $q = (q_1, q_2, \dots, q_n)$ Öklid n-uzayında iki nokta ise, p'den q'ya veya q'dan p'ye Öklid mesafesi (d) Pisagor formülü ile verilir (Denklem 3.10):

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{3.10}$$

3.5.2. Hiyerarşik kümeleme algoritmaları

Kümeler arasındaki mesafe için yaygın olarak kullanılan algoritmalar Denklem 3.11 – 3.13 ile verilmiştir. Burada $|p - p'|$, iki nesne veya nokta olan p ve p' arasındaki mesafedir. n_i ise C_i kümesindeki nesne sayısıdır.

Minimum mesafe : $d_{min}(C_i, C_j) = \ min_{p \in C_i, p' \in C_j} |p - p'|$ (3.11)

Maksimum mesafe : $d_{max}(C_i, C_j) = \ max_{p \in C_i, p' \in C_j} |p - p'|$ (3.12)

Ortalama mesafe : $d_{avg}(C_i, C_j) = \ \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$ (3.13)

Bir algoritma, kümeler arasındaki mesafeyi ölçmek için minimum mesafe olan $d_{min}(C_i, C_j)$ kullandığında, buna *en yakın komşu kümeleme algoritması* da denir. Ayrıca, en yakın kümeler arasındaki mesafe keyfi bir eşiği aştığında kümeleme işlemi sonlandırılırsa buna *tek bağlantı algoritması* denir.

Bir algoritma, kümeler arasındaki mesafeyi ölçmek için $d_{max}(C_i, C_j)$ maksimum mesafesini kullandığında, buna *en uzak komşu kümeleme algoritması* da denir. En yakın kümeler arasındaki maksimum mesafe keyfi bir eşiği aştığında kümeleme işlemi sonlandırılırsa buna *tam bağlantı algoritması* denir. İki küme arasındaki mesafe, iki kümedeki en uzak düğümler tarafından belirlenir. En uzak komşu algoritması, her bir yinelemede kümelerin çapındaki artışı mümkün olduğunca en aza indirme eğilimindedir.

Gerçek kümeler oldukça kompakt ve yaklaşık olarak eşit büyüklükte ise, yöntem yüksek kaliteli kümeler üretecektir. Aksi takdirde, üretilen kümeler anlamsız olabilmektedir.

Yukarıdaki minimum ve maksimum ölçümler, kümeler arasındaki mesafenin ölçülmesinde iki uç noktayı temsil etmektedir. Aykırı değerlere veya gürültülü verilere aşırı duyarlı olma eğilimindedirler. *Ortalama mesafe*nin kullanımı, minimum ve maksimum mesafeler arasında bir uzlaşıdır ve aşırı duyarlılık sorununun üstesinden gelmektedir.

3.5.3. Normalizasyon

Veri biliminde veri önişleme işlemleri yapılacak analizler için önem arz etmektedir. Önişleme aşamasında yapılacak veri dönüşümleri yapılacak analiz için uygun biçimlere dönüştürülür veya birleştirilir. Veri dönüşümü düzleme (smoothing), birleştirme (aggregation), genelleme (generalization), öznitelik ekleme (attribute construction) ve normalizasyon şeklinde yapılabilir.

Düzleme bir veri temizleme biçimidir. *Birleştirme* ve *genelleme* veri azaltma biçimleri olarak kullanılır. *Öznitelik ekleme*de ise verilen özniteliklerden yeni öznitelik oluşturulur ve yüksek boyutlu verilerdeki yapının doğruluğunu ve anlaşılırlığını geliştirmeye yardımcı olmak için veri setine eklenir.

Bir öznitelik, değerleri 0.0 ile 1.0 gibi küçük bir aralık içine düşecek şekilde ölçeklendirerek normalleştirilir. ***Normalleştirme***, sinir ağlarını içeren sınıflandırma algoritmaları veya en yakın komşu sınıflandırma ve kümeleme gibi mesafe ölçümleri için özellikle yararlıdır. Örneğin sınıflandırma madenciliği için sinir ağı geri yayılım algoritması kullanıyorsa, eğitim gruplarında ölçülen her özellik için giriş değerlerinin normalleştirilmesi öğrenme aşamasının hızlanmasına yardımcı olacaktır. Mesafeye dayalı yöntemler için normalleştirme, başlangıçta büyük aralıklara (ör. gelir) sahip özniteliklerin, başlangıçta daha küçük aralıklara (örneğin, ikili öznitelikler) sahip niteliklerin etkisinden kaçınmasına yardımcı olur. Veri normalleştirmede kullanılan yöntemlerden en yaygın olanı "*min-maks normalizasyonu"*dur.

Min-maks normalizasyonu orijinal veriler üzerinde lineer bir dönüşüm gerçekleştirir. min_A ve $maks_A$ değerlerinin bir *A* özelliğine ait minimum ve maksimum değerler olduğunu varsayalım. Min-maks normalizasyonu *A*'nın bir değeri olan v'yi$[yeni_maks_A - yeni_min_A]$ aralığında v' değerine dönüştürür. Min-maks normalizsyonu Denklem 3.14'te verilmiştir.

$$v' = \frac{v - min_A}{maks_A - min_A}(yeni\_maks_A - yeni\_min_A) + yeni\_min_A \quad (3.14)$$

3.6. Veri seti

DNA dizi benzerlik analizinde hizalamasız karşılaştırma yöntemlerinin test edilmesinde literatürde yer alan çeşitli veri setleri bulunmaktadır. Bu veri setleri farklı uzunluklarda DNA dizilerinden oluşmaktadır. Önerilen metotların test edilmesinde yalnızca bir veri setinin kullanılması, hizalamasız karşılaştırmanın doğası gereği uzunluktan bağımsız olma yönünün test edilmesinde yetersiz kalmaktadır. Bu bakımdan farklı uzunluklardaki veri setlerinin kullanılması literatüre girmiş metotların test edilmesinde daha faydalı görülmektedir.

3.6.1. NCBI genetik veri tabanı

Bilgisayarlı bilgi işleme yöntemlerinin biyomedikal araştırmaların yürütülmesindeki öneminin anlaşılmasıyla 4 Kasım 1988'de Amerika Birleşik Devletleri'nde Ulusal Sağlık Enstitüleri (NIH)'ndeki Ulusal Tıp Kütüphanesi (NLM)'nin bir bölümü olarak Ulusal Biyoteknoloji Bilgi Merkezi (NCBI) kurulmuştur. Kuruluşun amacı hesaplamalı moleküler biyoloji alanında intramural bir araştırma programı oluşturmak ve biyomedikal veri tabanları oluşturarak bunların bilimsel çalışmalarla sürdürülmesini sağlamaktır. Moleküler biyoloji ve genetik alanındaki araştırmacıların ücretsiz olarak erişebildiği bu veri tabanın yanı sıra kuruluş, moleküler biyoloji, biyokimya ve genetik alanında mevcut bilgileri depolamak ve analiz etmek için otomatik sistemler oluşturmakla yükümlüdür. Bilim insanları tarafından bu tür veri tabanlarının ve yazılımların kullanımını kolaylaştırmak, ulusal ve uluslararası biyoteknoloji bilgilerini toplama çabalarını koordine etmek ve biyolojik açıdan önemli moleküllerin yapısını ve işlevini analiz etmek için bilgisayar tabanlı bilgi işlemenin ileri yöntemleri üzerine araştırma yapmak gibi alanlarda da çalışma yürütmektedir (NCBI, 2020).

3.6.2. FASTA formatı

FASTA formatı, baz çiftlerinin veya amino asitlerin tek harfli kodlar kullanılarak temsil edildiği nükleotid dizilerini veya peptit dizilerini temsil etmek için metin tabanlı

bir formattır (Choudhuri, 2014). FASTA biçimindeki bir dizi, tek satırlık bir açıklama ve ardından dizi verisi satırlarıyla başlar. Açıklama satırı, ilk sütundaki dizi verilerinden büyük (">") sembolü ile ayırt edilir. Tüm metin satırlarının 80 karakterden daha kısa olması önerilir. Şekil 3.7'de FASTA formatında örnek bir dizi verilmiştir.

```
>J01859.1 Escherichia coli 16S ribosomal RNA, complete sequence
AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGT
AACAGGAAGAAGCTTGCTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATG
GAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCG
GGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACG
ATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGG
CAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTT
CGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTACCCG
CAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAAC
TGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGT
AGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCG
TGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCC
TTGAGGCGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACT
CAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCT
TACCTGGTCTTGACATCCACGGAAGTTTTCAGAGATGAGAATGTGCCTTCGGGAACCGTGAGACAGGTGC
TGCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCCT
TTGTTGCCAGCGGTCCGGCCGGGAACTCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGA
CGTCAAGTCATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGGCGCATACAAAGAGAAGCGA
CCTCGCGAGAGCAAGCGGACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATG
AAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCG
CCCGTCACACCATGGGAGTGGGTTGCAAAAGAAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACTT
TGTGATTCATGACTGGGGTGAAGTCGTAACAAGGTAACCGTAGGGGAACCTGCGGTTGGATCACCTCCTT
```

Şekil 3.7. FASTA formatında verilmiş örnek bir DNA dizisi (NCBI Genbank)

3.6.3. NADH dehidrogenaz alt birim 4 genleri

Önerilen metotların test edilmesinde kullanılan veri setlerinden ilki 4 farklı primat grubundan oluşan 12 türe ait NADH dehidrogenaz alt birim 4 genleridir. Veri seti, Eski-Dünya maymunlarından 4 tür, Yeni-Dünya maymunlarından 1 tür, prosimiyenlerden 2 tür ve hominoidlerden 5 türden oluşmaktadır. Tablo 3.1'de tür detayları ve erişim kodları (accession number) verilen veri seti 893 ile 896 bp arasında değişen uzunluklardadır. Daha önceki çalışmalarda ilk olarak Hayasaka ve ark. (1988) tarafından kullanılmış olan veri seti daha sonra Zhang ve Chen (2007); Zhang (2008); Qi ve ark. (2011); Chen ve ark. (2018); Delibaş ve Arslan (2020) ve Delibaş ve ark. (2020) tarafından kullanılmıştır.

Tablo 1.1. 12 Türe Ait NADH dehidrogenaz alt birim 4 genlerine ait bilgiler (NCBI)

	Türler	Erişim Kodu	Uzunluk (bp)
1	*Macaca fascicularis*	M22653	896
2	*Macaca fuscata*	M22651	896
3	*Macaca mulatta*	M22650	896
4	*Macaca sylvanus*	M22654	896
5	*Saimiri sciureus*	M22655	893
6	*Chimpanzee*	V00672	896
7	*Lemur catta*	M22657	895
8	*Gorilla*	V00658	896
9	*Hylobates*	V00659	896
10	*Sumatran Orangutan*	V00675	895
11	*Tarsius syrichta*	M22656	895
12	*Human*	L00016	896

3.6.4. 13 bakteri türüne ait 16S ribozomal DNA dizisi

Önerilen metodun farklı uzunluklardaki başarısının test edilmesi adına laboratuvar ortamlarında sıklıkla karşılaşılan bakteriyel analizlerde kullanılan DNA uzunluğuna uygun bir veri seti oluşturulmuştur. 16S ribozomal DNA dizileri veri setine eklenen bakteriler rastgele seçilmiştir. Veri setinde kullanılan bakteriler üç farklı grup olacak şekilde oluşturulmuş ve bu 3 farklı gruba tek bakteri içeren bir adet dış grup eklenmiştir. Bu şekilde birbirine çok yakın grup içi ayrıştırmayı yaparken, farklı gruplardakilerle iyi ve doğru ayrıştırmayı test etmesi hedeflenmiştir. Tüm diziler NCBI genetik veri tabanından alınmış ve detay bilgileri Tablo 3.2'de verilmiştir. Dizi uzunlukları 1,509 ile 1,541 bp arasında değişmektedir. Veri seti Delibaş ve Arslan (2020) ve Delibaş ve ark. (2020) tarafından kullanılmıştır.

Tablo 1.2. 13 bakteri türüne Ait 16S ribozomal DNA dizi bilgileri

	Türler	Erişim Kodu	Uzunluk (bp)
1	*Bacillus maritimus*	KP317497	1,515
2	*Bacillus wakoensis*	NR_040849	1,524
3	*Bacillus australimaris*	NR_148787	1,513
4	*Bacillus xiamenensis*	NR_148244	1,513
5	*Escherichia coli*	J01859	1,541
6	*Streptococcus himalayensis*	NR_156072	1,509
7	*Streptococcus halotolerans*	NR_152063	1,520
8	*Streptococcus tangierensis*	NR_134818	1,520
9	*Streptococcus cameli*	NR_134817	1,518
10	*Thermus amyloliquefaciens*	NR_136784	1,514
11	*Thermus tengchongensis*	NR_132306	1,523
12	*Thermus thermophilus*	NR_037066	1,515
13	*Thermus filiformis*	NR_117152	1,514

3.6.5. 18 eteneli memeliye ait tüm mitokondriyal DNA

18 eteneli memeliye ait tüm mitokondriyal genomlar bol miktarda genetik bilgi içermektedir. Son yıllarda literatürdeki analizlerde de karşılaşılan veri setindeki tüm diziler NCBI veri tabanından elde edilmiştir ve uzunlukları 16,295 bp ile 17,019 bp arasındadır (Yang ve Wang, 2013; Hou ve ark., 2014; Jin ve ark., 2016; Delibaş ve Arslan, 2020; Delibaş ve ark., 2020). Veri setindeki türlere ait detaylı bilgiler Tablo 3.3'te verilmiştir.

Tablo 1.3. 18 eteneli memeliye ait tüm mitokondriyal DNA bilgileri

	Türler	Erişim Kodu	Uzunluk (bp)
1	*Human*	V00662	16,569
2	*Pygmy chimpanzee*	D38116	16,563
3	*Common chimpanzee*	D38113	16,554
4	*Gorilla*	D38114	16,364
5	*Orangutan*	D38115	16,389
6	*Gibbon*	X99256	16,472
7	*Baboon*	Y18001	16,521
8	*Horse*	X79547	16,660
9	*White rhinoceros*	Y07726	16,832
10	*Harbor seal*	X63726	16,826
11	*Gray seal*	X72004	16,797
12	*Cat*	U20753	17,009
13	*Fin whale*	X61145	16,397
14	*Blue whale*	X72204	16,402
15	*Cow*	V00654	16,338
16	*Rat*	X14848	16,300
17	*Mouse*	V00711	16,295
18	*Platypus*	X83427	17,019

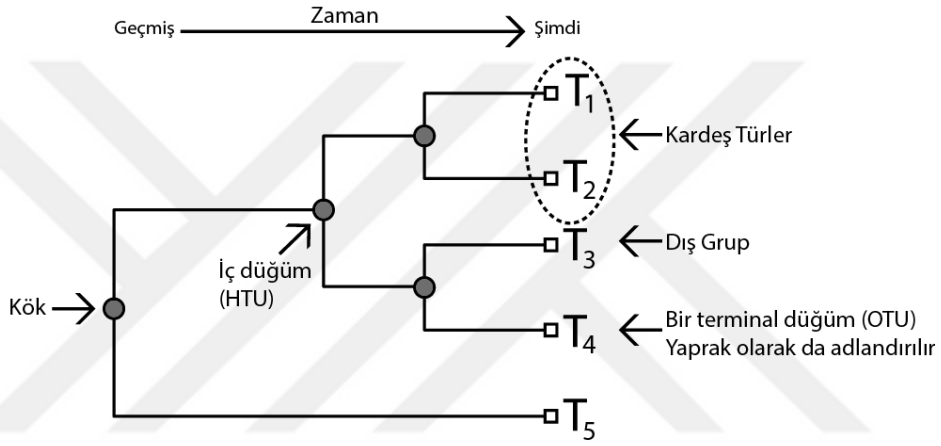
3.7. Çıktılar

3.7.1. Filogenetik ağaç

Filogeni türlerin evrimsel tarihini ifade eder. *Filogenetik,* filogenilerin, yani türlerin evrimsel ilişkilerinin incelenmesidir. Filogenetik analiz, evrimsel ilişkileri tahmin etmenin bir yoludur. Moleküler filogenetik analizde, türlerin evrimsel ilişkisini değerlendirmek için ortak bir gen veya proteinin dizisi kullanılabilir. Filogenetik analizden elde edilen evrimsel ilişki genellikle dallanma, ağaç benzeri diyagram yani filogenetik ağaç olarak tasvir edilir (Choudhuri, 2014).

Bir filogenetik ağaç veya evrim ağacı, çeşitli taksonlar arasındaki evrimsel ilişkilerin diyagramatik bir temsilidir. Düğüm ve dallardan oluşan bir dallanma diyagramıdır. Bir ağacın dallanma düzenine ağacın topolojisi denir. Düğümler, türler (veya daha yüksek taksonlar), popülasyonlar, genler veya proteinler gibi taksonomik

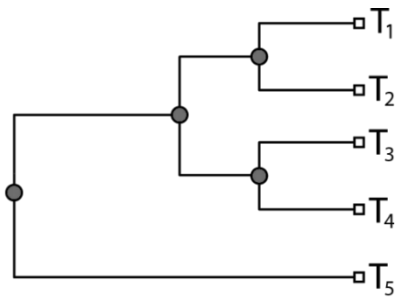
birimleri temsil eder. Dallara kenar denir ve taksonomik birimler arasındaki evrimsel ilişkilerin zaman tahminini temsil eder. Bir dal yalnızca iki düğümü bağlayabilir. Bir filogenetik ağaçta, terminal düğümleri operasyonel taksonomik birimleri (OTU) veya yaprakları temsil eder. OTU' lar türler, popülasyonlar veya gen veya protein dizileri gibi gerçek nesneler karşılaştırılırken, iç düğümler varsayımsal taksonomik birimleri (HTU'lar) temsil eder. Bir HTU, çıkarılmış bir birimdir ve bu noktadan kaynaklanan düğümlerin son ortak atalarını (LCA) temsil eder. Aynı düğümden ayrılan torunları (takson) kardeş gruplar oluşturur (Choudhuri, 2014). Şekil 3.8'de örnek bir filogenetik ağaç gösterilmiştir.



Şekil 3.8. Filogenetik ağaç

3.7.2. Newick formatı

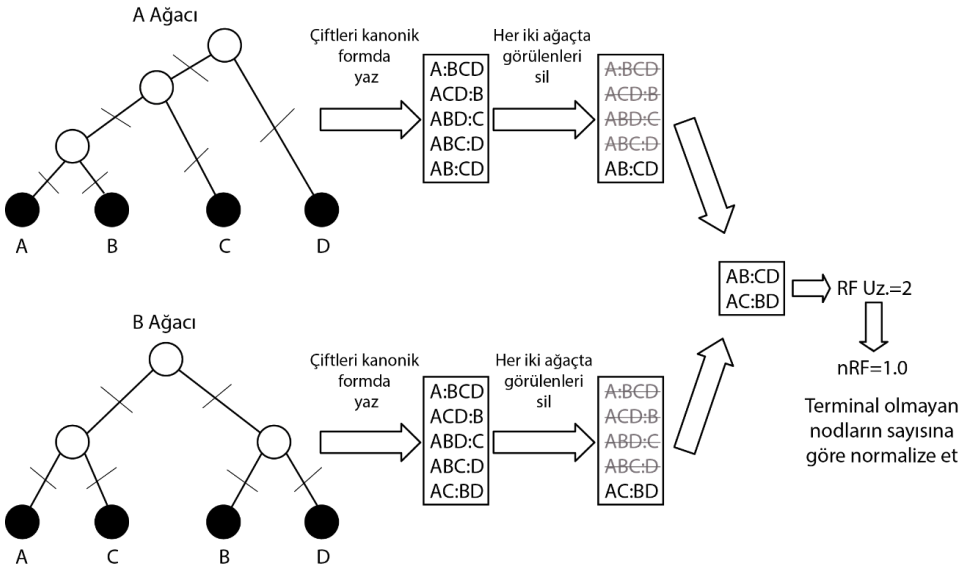
Newick formatı 24 Haziran 1986'da Evrim Araştırmaları Derneği'nin bir toplantısı ile duyurulmuştur. Bu formatın hedef kitlesi filogenetik ağaçlar üzerinde çalışan araştırmacılardır. Etiketli düğümlerle oluşturulmuş köklü ağaçların gösterimi ve türlerle ebeveynler arası evrimsel uzunlukları ifade etmek için kullanılır. Standart format noktalı virgülle bitirilir. Köke giden çizgi için Filogeni bağlamında bir anlamı olmadığından uzunluk belirtilmez. Bir ağaç, düğüm adlarının iç içe parantezlerle gruplandırılması ile ifade edilir. Düğüm adlarının yanına ":" işareti konularak düğümden ayrılma noktasına uzaklığı verilebilir (Durbin ve ark., 1998). Örneğin Şekil 3.9'da verilen filogenetik ağaç için newick formatında gösterim; $((T_1, T_2), (T_3, T_4), T_5)$ şeklindedir.



Şekil 3.9. Newick formatında gösterim için örnek filogenetik ağaç

3.7.3. Robinson-Foulds uzaklığı

Filogenetik ağaçlar arasındaki topolojik korelasyonu belirlemek için, 0-1aralığına normalleştirilmiş Robinson-Foulds (nRF) (Robinson ve Foulds, 1981) değerlerini kullanarak uyumluluk ölçümü yapılabilmektedir. Bu tür karşılaştırmalarda sıklıkla kullanılan metriklerden olan Robinson-Foulds (RF) metodu, iç dalların herhangi bir ağacı iki terminal düğüm grubuna ayırdığını kabul eder. RF her iki ağaç için tüm çift bölmeleri numaralandırır, mükerrer çift bölümleri dışarı atar son olarak kalan benzersiz çift bölümleri sayar. RF uzaklığı sadece tekrarlamayan çiftlerin sayısıdır. Üçgen eşitsizliğini sağladığı için RF, gerçek bir matematiksel metrik olarak düşünülebilir. Elde edilen 0 puanı, incelenen ağaçların uyumlu olduğunu, 1 puanı ise ağaçlar arasında hiçbir uyumun olmadığını göstermektedir. Düşük nRF skorları da iki ağaç arasında yüksek düzeyde uyum olduğunu göstermektedir. Metodun hesaplama basamakları Şekil 3.10'da verilmiştir (Sheneman ve Foster, 2006).



Şekil 3.10. nRF uzaklığının hesaplanması

3.7.4. MEGA7 moleküler evrimsel genetik analiz yazılımı

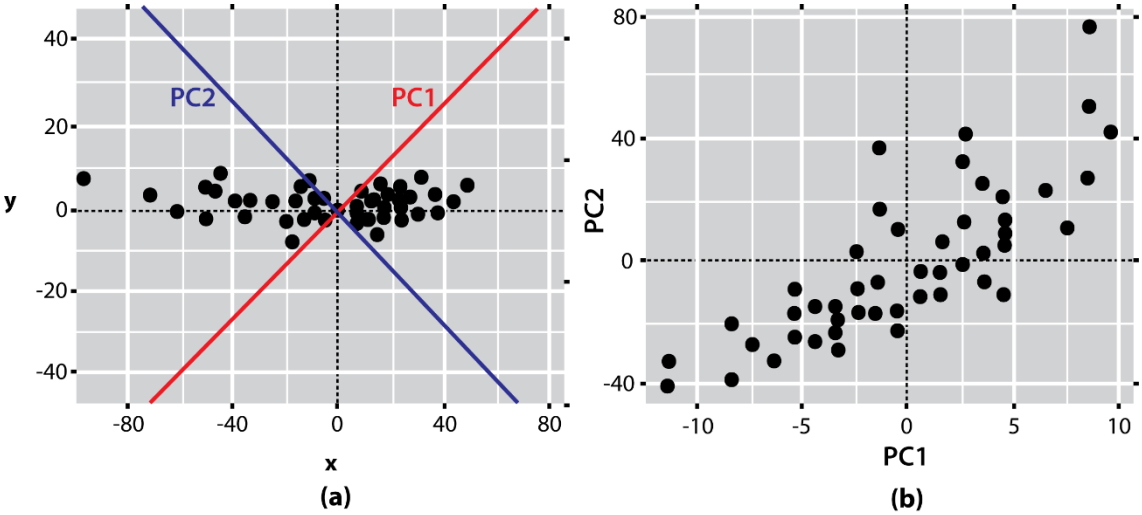
Genom dizilemesi, çok çeşitli organizmalardan büyük miktarlarda DNA dizi verisi üretmektedir. Sonuç olarak, gen dizisi veri tabanları hızla büyümektedir. Bu verilerin verimli analizini yapmak için, çeşitli algoritmalar ve yararlı istatistiksel yöntemler içeren kullanımı kolay bilgisayar programlarına ihtiyaç hasıl olmaktadır. MEGA (Molecular Evolutionary Genetics Analysis) yazılımı da bu ihtiyaca karşılık üretilmiş bir programdır. MEGA yazılımının amacı DNA ve protein dizilerini evrimsel bir bakış açısıyla keşfetmek ve analiz etmek için araçlar sağlamaktır. Akademik alt yapıyla donatılmış yazılım ilk olarak Kumar ve ark. (1994) tarafından duyurulmuştur. Literatüre girdiği yıldan itibaren çeşitli sürümleri ile literatürdeki ve teknolojideki gelişmelere paralel olarak versiyonlamalar yapan yazılımın mevcut en gelişmiş versiyonu olan MEGA7, kaynak kodunun önemli bir yenilemesi ve performans açısından geliştirilmiş bir sürümüdür (Kumar ve ark., 2016b).

3.7.5. Temel bileşen analizi

Temel bileşen analizi (PCA), bir veri kümesini basitleştirmek için kullanılan istatiksel bir tekniktir. Hotelling (1933) tarafından geliştirilen metodun amacı, içerilen ilgili bilginin mümkün olduğunca çoğunu korurken çok değişkenli verilerin boyutsallığını azaltmaktır. Karşılık gelen hedef verilere atıfta bulunulmaksızın tamamen girdi verisine dayandığı için gözetimsiz bir öğrenmedir. Maksimize edilecek kriter varyanstır. PCA'nın başlangıç noktası, orijinal veri kümesinden türetilen korelasyon katsayılarının matrisidir. Yöntemin arkasındaki mantık, korelasyonların sürekli bir ölçekte ölçülen değişkenlerden elde edilmesini gerektirir. Uygulamada bu, tam veri seti için kovaryans matrisinin hesaplanmasıyla elde edilir. Daha sonra, kovaryans matrisinin özvektörleri ve özdeğerleri hesaplanır ve azalan özdeğere göre sıralanır. Yöntem bazen korelasyon katsayıları metrik değişkenlerden hesaplanmadığında kullanılır. Böyle bir analiz geçerli değildir, ancak buna rağmen yöntem, verilerin yapısının kaba, yararlı bir özetini verebilir.

Aşağıdaki Şekil 3.11(a)'da, veriler X-Y koordinat sisteminde temsil edilmektedir. Boyut küçültme, verilerin değiştiği temel bileşenler olarak adlandırılan ana yönlerin tanımlanmasıyla elde edilir. PCA, en büyük varyanslı yönlerin en "önemli" (yani, en temel) olduğunu varsayar. Şekil 3.11(b)'de, PC1 ekseni, örneklerin en büyük varyasyonu gösterdiği ilk ana yöndür. PC2 ekseni en önemli ikinci yöndür ve PC1 eksenine diktir. İki

boyutlu verilerimizin boyutsallığı, her bir örnek ilk temel bileşene yansıtılarak tek bir boyuta indirilebilir.



Şekil 3.11. X-Y koordinat düzlemine yerleştirilmiş verilerin PCA ile temsili

Teknik olarak, her bir ana bileşen tarafından tutulan varyans miktarı, özdeğer (eigenvalue) olarak ölçülür. PCA yönteminin, özellikle veri kümesindeki değişkenler yüksek derecede korelasyonlu olduğunda faydalı olduğu belirtilmelidir. Korelasyon, verilerde fazlalık olduğunu gösterir. Bu fazlalık nedeniyle, PCA, orijinal değişkenleri orijinal değişkenlerdeki varyansın çoğunu açıklayan daha az sayıda yeni değişkene (temel bileşenler) azaltmak için kullanılabilir (Kassambara, 2017).

4. ÖNERİLEN METOTLAR

DNA dizilerinin benzerlik analizi yapılırken ele alınan diziler farklı uzunluklarda olabilmektedir. Literatürde önerilen metotların testi için kullanılan veri setleri yaklaşık olarak 100 baz ile 20 bin baz aralığında değişiklik göstermektedir. Dizi uzunluklarındaki bu farklılıklar önerilen analiz metotları için büyük zorluklar oluşturmaktadır. Bu yöntemlerin bazıları hem kısa hem de uzun dizilerin analizi konusunda başarısız olabilmektedir. DNA dizilerinin benzerlikleri biyolojik işlevlerini ve evrimsel bilgilerini öğrenmek için kullanılmaktadır. Bu bağlamda benzerlikler biyolojik mekanizmalar ile temsil edilmelidir. Bununla birlikte biyolojik özelliklerin ve kimyasal özelliklerin makul bir şekilde nasıl dikkate alınacağı dikkatle ele alınması gereken bir problemdir. Ayrıca, yüksek hesaplama karmaşıklığı ve uzun DNA dizileri için büyük işlem belleği ihtiyaçları DNA dizi analiz yöntemlerinin ciddi sorunları arasındadır (Jin ve ark., 2017). Daha doğru ve kullanışlı benzerlik analizi yöntemleri, bilinmeyen DNA dizilerinin fonksiyon bilgilerini ve evrim bilgilerini bulmamıza daha etkili bir şekilde yardımcı olabilecektir (Wang ve ark., 2010).

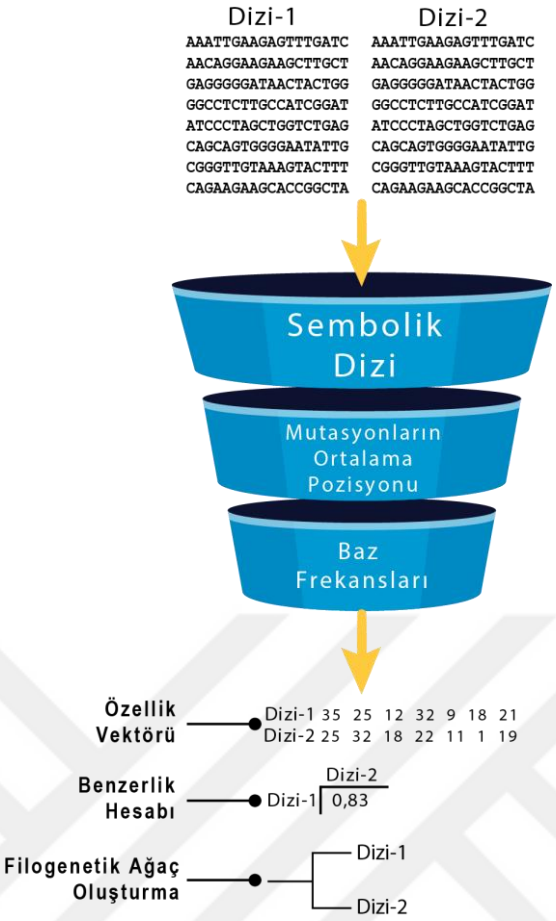
Bu tez çalışmasında yukarıda bahsedilen sorunlara çözüm üretebilmek adına, bilgisayar mühendisliği disiplininde kullanılan özellik çıkarımı yöntemlerinin DNA dizilerinde benzerlik analizine uyarlanmasıyla elde edilecek metotlar önerilmiştir. Bu maksatla "Grup Mutasyonlarının Ortalama Konumları ve Nükleotid Bazlarının Frekanslarına Dayalı Benzerlik Analizi", "Birinci derece istatistik tabanlı görüntü doku analizi kullanarak benzerlik analizi" ve "Üst-*k* *n*-gram Eşleşmesi Tabanlı Benzerlik Analizi" yöntemleri geliştirilmiştir. Önerilen bu metotlarda işlem adımları ana hatlarıyla DNA dizilerinin sayısallaştırılarak çeşitli boyutlarda vektörlere dönüştürülmesi, bu vektörler arasında bir benzerlik metriği kullanarak ilişkilerin hesaplanması ve bu benzerlik ilişkilerine göre sonuçların değerlendirilmesi şeklinde teşekkül etmiştir. Yöntemlerin performansının test edilmesi adına kullanılan veriler, literatürde sıklıkla tercih edilen veri kümelerinden alınmıştır.

4.1. Grup Mutasyonlarının Ortalama Konumları ve Nükleotid Bazların Frekanslarına Dayalı Benzerlik Analizi

DNA dizileri nükleotid denilen basit moleküllerden oluşmaktadır. Nükleotidler, nükleositlerin fosfat esterleridir ve DNA'nın bileşenleridir. Tüm nükleotitleri oluşturan

üç bileşen, bir azot hetero-halkalı baz, bir pentoz şekeri ve bir fosfat grubudur. Temel bazlar tek halkalı pirimidinler (Y) ve iki halkalı pürinler (R)' dir. Prürinler adenin (A) ve guanin (G), pirimidinler ise sitozin (C) ve timin (T)'dir. Bu nükleotidler DNA dizisini temsil eden yukarıdaki dört harften (A, G, C ve T) oluşan bir metin dizisi olarak ifade edilebilir.

Bilgisayar bilimi perspektifinden bakıldığında bir DNA dizisi *string* veri tipinde bir metin parçasıdır. Bu nedenle DNA dizileri arasındaki benzerlik analizi için metin benzerliğinde kullanılabilecek metotlar rahatlıkla uygulanabilmektedir. Metinler arasındaki benzerlik ölçümü yapılacağı zaman metinler uzunluklarından bağımsız olarak farklı türden vektörlere dönüştürülmektedir. Son olarak da bu vektörler arasında benzerlik hesaplaması yapılmaktadır. Bir özellik vektörü oluşturulurken dikkate alınması gereken konu vektörü oluşturan özelliklerin, benzerlik hesaplamasına tabi tutulan öğeleri en iyi şekilde temsil kabiliyetine sahip özelliklerin seçilip kullanılmasıdır. Bu değerlerin ayrıştırıcılık kabiliyeti benzerlik hesaplamasında belirleyici unsur olacaktır. Bu çalışmada DNA dizilerinden 7 boyutlu bir özellik vektörü elde edilmiştir. Elde edilen bu özellik vektörleri nükleotidlerin kimyasal özellikleriyle ilişkili grup mutasyonlarının ortalama pozisyonu ve nükleotidlerin salt frekanslarından oluşmaktadır. Bu iki özelliğin birleşimiyle benzerlik analizinde kullanılabilecek düzeyde ayrıştırıcı öznitelik tespiti yapılmaya çalışılmıştır. Metodun işlem süreci Şekil 4.1'de sunulmuştur.



Şekil 4.1. Grup mutasyonlarının ortalama konumları ve nükleotid bazlarının frekanslarına dayalı benzerlik analizi süreci

Oluşturulacak özellik vektöründeki birinci grup değerler nükleotidlerin mutasyon bilgilerinden oluşmaktadır. Pürin ve pirimidinler sırasıyla R ve Y harfleriyle sembolize edilmiştir. ***n*** uzunluklu DNA dizisi $S = s_1 s_2 s_3 \dots s_n$ şeklinde ifade edilsin. Bu diziden $\emptyset_{RY}$ şeklinde isimlendirilen bir simgesel haritalama dizisi elde edilir:

$$\emptyset_{RY}(S) = \emptyset_{RY}(s_1)\emptyset_{RY}(s_2)\emptyset_{RY}(s_3) \dots \emptyset_{RY}(s_n),$$

burada $\emptyset_{RY}(s_i) = \begin{cases} R, & eğer\ s_i \in R \\ Y, & eğer\ s_i \in Y \end{cases}$, $i = 1,2,3, \dots, n$'dir.

Bu haritalama uygulandığında örnek olarak alınabilecek $S = AAGCTTATAGGCCCT$ dizisi aşağıdaki gibi dönüştürülecektir:

$$\emptyset_{RY}(S) = RRRYYYRYRRRYYYY$$

Önerilen metotta aynı gruptan olan nükleotidlerin peş peşe geldiği korunmuş bölgeler dikkate alınmaktadır. Vektörün oluşturan bu değerler $(\mu_{RR}, \mu_{YY}, \mu_D)$ aynı nükleotid grubu olarak devam edilen bölgeler (R'den R'ye veya Y'den Y'ye) ve aynı nükleotidin kopyası şeklinde devam eden bölgelerin (duplication) ortalama konumları

hesaplanmıştır. Ortalama uzaklıkları ifade eden μ değerleri Denklem 4.1, 4.2 ve 4.3' te tanımlanmıştır:

$$\mu_{RR} = \sum_{i=1}^{n_{RR}} \frac{d_i}{n_{RR}} \tag{4.1}$$

burada d_i ilk nükleotidden RR_i'nin pozisyonu arasındaki uzaklığı, n_{RR} R'den R'ye geçişlerin toplam sayısını ifade etmektedir.

$$\mu_{YY} = \sum_{i=1}^{n_{YY}} \frac{d_i}{n_{YY}} \tag{4.2}$$

burada d_i ilk nükleotidden YY_i'nin pozisyonu arasındaki uzaklığı, n_{YY} Y'den Y'ye geçişlerin toplam sayısını ifade etmektedir.

$$\mu_D = \sum_{i=1}^{n_D} \frac{d_i}{n_D} \tag{4.3}$$

burada d_i ilk nükleotidden nükleotid duplikasyonunu ifade eden D_i'nin pozisyonu arasındaki uzaklığı, n_D duplikasyon olan noktaların toplam sayısını ifade etmektedir.

İki DNA dizisi benzer ise ortalama uzaklık değerleri de benzerlik gösterebilir. Ancak bu değerler tek başına DNA dizileri hakkında bilgi verme konusunda yetersiz olacağından ve farklı pozisyonlardaki mutasyonların ortalama mesafeleri aynı olabileceğinden vektör daha fazla güçlendirmeye muhtaçtır.

Özellik vektörünün ikinci grup parametreleri DNA dizisini oluşturan A, G, C ve T içeriği hakkındaki bilgiden oluşturulmuştur. Bu 4 sayısal değer A, G, C ve T nükleotidlerinin frekanslarıdır ve sırasıyla n_A, n_G, n_C ve n_T şeklinde sembolize edilen değerlerdir. Bu 4 tam sayı değeri basit ancak dizi hakkında içerdiği önemli ve ayırt edici bilgiyle vektörün karakterizasyonuna katkı sağlamaktadır. Bu şeklide elde edilen iki parametre grubunun kombinesiyle oluşan sayısal vektör DNA dizileri arasında benzerlik hesaplaması yapabilmek için kullanılmıştır. Elde edilen 7-boyutlu vektör Denklem 4.4'te ifade edilmiştir:

$$V = [\mu_{RR}, \mu_{YY}, \mu_D, n_A, n_G, n_C, n_T] \quad (4.4)$$

Yukarıda verilen S DNA dizisi ve $\emptyset_{RY}(S)$ sembolik dizisine metodun uygulaması yapıldığında 7-boyutlu V vektörü aşağıdaki hesaplamalarla elde edilir:

$$\mu_{RR} = \frac{22}{4} = 5.5$$
$$\mu_{YY} = \frac{48}{5} = 9.6$$
$$\mu_D = \frac{41}{5} = 8.2$$
$$n_A = 4, n_G = 3, n_C = 4, n_T = 4$$
$$V = [5.5, 9.6, 8.2, 4, 3, 4, 4]$$

4.1.1. Benzerlik hesaplaması

Önceki bölümde 7-boyutlu lineer uzayda bir özellik vektörü elde edilmiştir. Bu vektör DNA dizileri arasındaki benzerliğin hesaplanmasında kullanılmıştır. DNA dizileri arasındaki benzerliğin hesaplanmasında uzaklık hesaplaması analizin önemli bir adımıdır. Dizi karşılaştırmaları yapılırken en yaygın kullanılan uzaklık ölçüsü Öklid uzaklığıdır (Jin ve ark., 2017). Elde edilen karakterizasyon arasındaki benzerlik, uç noktalar arasında Denklem 3.10'da verilen Öklid uzaklığı ile hesaplanmıştır:

4.2. Birinci Derece İstatistik Tabanlı Görüntü Doku Analizi Kullanarak Benzerlik Analizi

Benzerlik hesaplamalarında literatürde birçok yöntem kullanılmış, bu yöntemlerin bazıları doğrudan DNA benzerliği hesaplamaları için tasarlanmış bazıları ise farklı alanlardaki uygulamalarının, farklı başarı seviyelerinde, DNA benzerliğine uyarlanması şeklinde ortaya konmuştur. Bu bağlamda, DNA dizi benzerlik analizi için önerilen ikinci sistemde, bilgisayar biliminde geniş uygulama alanı bulan görüntü işlemede doku analizi uygulamasına dayanan bir yöntem önerilmiştir. Bir görüntüde bulunan bilgi deseni veya yapının düzenlemesi olan *doku*, birçok görüntü türünün önemli bir özelliğidir. Genel olarak doku, görüntünün temel parçalarının şekli, boyutu, düzeni, yoğunluğu ve oranı ile tanımlanan bir nesnenin yüzey karakteristiğini ve görünümünü ifade etmektedir.

Dokunun içerdiği bilgiler nedeniyle, doku özelliği çıkarma, çeşitli görüntü işleme uygulamalarında, uzaktan algılamada ve içerik tabanlı görüntü erişiminde önemli bir işlevdir. Doku özellikleri istatistiksel, yapısal, model tabanlı ve dönüştürülmüş bilgiler kullanılarak çeşitli yöntemlerle elde edilebilir (Pham, 2010). Görüntü dokularının özellik çıkarımı için kullanılan istatistiksel yöntemler, görüntünün gri seviyeleri arasındaki dağılımları ve ilişkileri dolaylı olarak yöneten deterministik olmayan özelliklere göre dokuyu temsil eder. Bu teknik, makine görmesinin ilk yöntemlerinden biri olarak literatüre katılmıştır (Tuceryan ve Jain, 1998). İstatistiksel yöntemler, görüntüdeki her bir noktada yerel karakteristikler hesaplanarak ve yerel karakteristiklerin dağılımından bir dizi istatistik elde edilerek gri seviye değerlerin uzamsal dağılımını analiz etmek için kullanılabilir. Birinci dereceden bir istatistik olan histogram tabanlı özellikler, orijinal görüntü özelliklerinden hesaplanır ve komşuluk ilişkilerini dikkate almaz. Doku analizine yönelik histogram tabanlı yaklaşım, bir görüntünün tamamı veya görüntünün histogram olarak gösterilen bir kısmı için yoğunluk değerlerinin konsantrasyonlarına dayanır. Yaygın özellikler ortalama, varyans, enerji, entropi, çarpıklık ve basıklıktır (Srinivasan ve Shobha, December 2008).

Belirli bir görüntünün histogramı kolayca hesaplanabilir. Bir histogramın şekli görüntü hakkında önemli bilgiler içerir. Örneğin, dar dağılmış bir histogram düşük kontrastlı bir görüntüyü gösterir. Bir bimodal histogram genellikle görüntünün farklı yoğunluktaki bir arka plana karşı dar yoğunluklu bir nesne içerdiğini gösterir (Srinivasan ve Shobha, December 2008). Doku, histogramda bulunan bilgilerle karakterize edilebilir. Bu şekilde, histogram tabanlı özellikler görüntülerin sınıflandırılmasında, görüntünün alınması ve görüntü segmentasyonunda sıklıkla kullanılır (Mapayi ve ark., 2015). Buna ek olarak, doku analizi, göğüs lezyonlarının saptanması (Gomez ve ark., 2012), damar segmentasyon tekniklerinde (Fraz ve ark., 2012) ve PET / CT görüntülerinin evrelenmesi (Win ve ark., 2013) gibi birçok biyomedikal uygulamada kullanılmıştır.

Yukarıda belirtilen uygulamalarda görüldüğü üzere doku analizi çeşitli uygulamalardaki benzerlik tabanlı sınıflandırmada yaygın olarak kullanılmıştır. Önerilen metotta da görüntü dokuları için histogram tabanlı özellik çıkarma yöntemi DNA dizi analizine uygulanmıştır. Bir DNA dizisinin özelliklerini tanımlamak ve hesaplamak için yukarıda bahsedilen momentlere dayanan doku analizi teorisi uygulanmaktadır. Hizalamasız DNA dizi analiz yöntemlerinde olduğu gibi, her dizi birbirinden bağımsız olarak vektörel sayısallaştırmaya tabi tutulur. Bu sayısallaştırmanın temelinde DNA dizilerinin birer gri-seviye dijital görüntüye dönüştürülmesi ve bu görüntüler üzerinde

gerçekleştirilecek doku analizi yatmaktadır. Doku analizinin birinci derece istatistik özellikleri, başka bir deyişle histogram özellikleri özellik vektörünü teşkil etmektedir. Histogram özelliklerinden *çarpıklık, basıklık, entropi* ve *enerji*, 4-boyutlu özellik vektörünün metrikleridir (Delibaş ve Arslan, 2020).

4.2.1. DNA dizisini bir sayısal görüntüye çevirme

Bir resmi oluşturan piksellere benzer olarak DNA dizileri de nükleotidleri ifade eden A, G, C ve T karakterlerinden oluşan alt birimlerden oluşmaktadır. Gri-seviye bir görüntü için pikseller, bu pikselin gri-seviye değerine karşılık gelen sayılarla temsil edilir. Önerilen metotta da DNA dizisini bir görüntüye dönüştürmek için nükleotidlere karşılık gelen harfler için sayısal eşdeğerler atanmıştır. Histogram tabanlı özellikler, yukarıda belirtildiği gibi, orijinal görüntü özelliklerinden hesaplanır ve komşuluk ilişkilerini dikkate almaz. Bu nedenle, karakterizasyonu daha da güçlendirmek için nükleotidler ikili gruplar halinde kullanılmıştır. Böylece, komşu olarak temel geçişlerle ilgili veriler orijinal görüntü özelliklerine gömülmektedir. Değerler harflere atanırken, 4 harfin ikili kombinasyonları olan 16 durum için sayısal karşılıklar belirlenmiştir. Değer atanmış ikili nükleotid grupları, bir baz kaydırılarak elde edilmiştir. Atanan değerler 1-255 aralığında sırasıyla ve eşit aralıklı olarak dağılacak şekilde belirlenmiştir. Bu dinükleotidler aşağıdaki gibidir:

$$\alpha = \{AA, AG, AC, AT, GA, GG, GC, GT, CA, CG, CG, CT, TA, TG, TC, TT\}$$

DNA dizilerine, bir erişim kodu ile genetik veri tabanlarından erişilebilir. Bu veri tabanlarından bir DNA dizisi FASTA formatında alındığında, bir satırda 70 baz olacak şekilde düzenlenir. Görüntüye dönüştürme, DNA dizisinin bu formattaki sunumuna göre oluşturulur. Bu şekilde, görüntü bir satırda 70 piksel genişliğinde ve dizinin büyüklüğüne göre değişen bir yükseklikle elde edilir. Bu şekilde, gri-seviyeli bir görüntü elde edilir ve analiz edilecek tüm diziler dönüştürülür.

4.2.2. Histogram analizine dayalı özellik çıkarımı

Birinci dereceden doku analizi (veya histogram analizi) gri-seviyeli bir görüntüde piksel yoğunluğu değerlerini çıkarır(Alobaidli ve ark., 2014). Histogramı hesaplayan bu

değerler üzerinden görüntünün birinci dereceden istatistik özelliklerini tanımlamak için farklı parametreler kullanılır. Merkezi momentler olarak da adlandırılan bu metrikler aynı zamanda DNA dizisini karakterize edecek değerlerdir (Levine, 1985; Pratt, 1991). Metriklere ait denklemler Denklem 3.4 – 3.9' da verilmiştir. Özellik vektörü, histogramın şekli hakkında bilgi içeren bu metriklerle tasarlanmıştır. Bu şekilde, görüntüye dönüştürülen DNA dizileri, bu görüntünün özelliklerini karakterize eden aşağıdaki metriklerle 4 boyutlu vektörler elde edilmiştir. Elde edilen V vektörü Denklem 4.5'te verilmiştir:

$$V = [\mu_3, \mu_4, E, H] \tag{4.5}$$

4.2.3. Benzerlik hesaplaması

4-boyutlu lineer uzayda bir karakterizasyon vektörü olan V elde edildikten sonra, bu vektörlerle DNA dizilerinden dönüştürülen dokuların histogramları arasında bir karşılaştırma yapılmıştır. Bu vektörler arasındaki benzerlikler, uç noktalar arasına Öklid uzaklığı (Denklem 3.10) uygulanarak hesaplanmıştır.

4.3. Üst-*k* *n*-gram Eşleşmesi Tabanlı Benzerlik Analizi

Kelime tabanlı benzerlik ölçümü kullanan yöntemler yaygın kullanılan hizalamasız DNA dizisi benzerlik analizi yöntemleri arasındadır. Kelime tabanlı yöntemler arasında *n*-gram-tabanlı yaklaşım ilk olarak Blaisdell tarafından önerilmiştir (Blaisdell, 1986). N-gram bazlı yöntemler, genetik dizilerin filogenetik analizi için, evrimsel modellere ihtiyaç duymadan hızlı ve kolay bir şekilde kullanılabilmektedir (Osmanbeyoglu ve Ganapathiraju, 2011; Ganapathiraju ve ark., 2012). Dizi karşılaştırması ve filogenetik analiz için çok sayıda *n*-gram bazlı yöntem önerilmiş ve kullanılmıştır. Bu yöntemlere ek olarak, optimal *n* aralığını bulmak için çeşitli yaklaşımlar da önerilmiştir. Bazı yöntemler, DNA dizileri arasındaki mesafeyi Jensen – Shannon ıraksaması ve Kullback – Leibler ıraksaması gibi benzerlik fonksiyonlarıyla ölçerek doğrudan *n*-gram frekansını kullanmışlardır. Sıklıkla kullanılmasına ve başarılı sonuçlar elde edebiliyor olmasına rağmen kelime tabanlı metotların en büyük dezavantajı hesapsal maliyetin yüksek olmasıdır.

N-gram tabanlı yöntemlerde çeşitli metin işlemlerinin hesaplama maliyetini azaltmak için, *n*-gramın bazı bölümleri yok sayılır. Bunu yaparken, üst-*k* *n*-gram olarak adlandırılan ve *n*-gram frekanslarına göre en yüksek puanı alan bölge belirlenir ve bu bölgenin benzerlik hesaplamasında baskın etkisi kullanılır (Cheng ve ark., 2005; Wagner ve ark., 2014; Alhanahnah ve ark., 2018). Önerilen bu yöntemde de özellik vektörleri elde etmek için DNA dizilerinin ikili karşılaştırmasında üst-*k* *n*-gram eşleşme sayısı kullanılmıştır. Yöntemin arka planında, önceki çalışmalardan esinlenilerek, doğrudan vektörel frekanslardan oluşturulmuş özellik vektörleri arasında benzerlik hesaplaması yapılarak hibrit bir yaklaşım ortaya çıkmıştır. Önerilen yöntem, üst-*k* *n*-gram'ı kullanarak DNA dizisi benzerliğini ortaya çıkarmak için en etkili baz gruplarını belirlemeyi ve kullanmayı amaçlamaktadır. Bu da hesaplama maliyetlerini önemli ölçüde azaltmaktadır. Analiz daha sonra benzerlik metrikleri kullanılarak gerçekleştirilir.

4.3.1. Üst-*k* *n*-gram tabanlı özellik çıkarımı

Bir DNA dizisi 4 harfli bir alfabe içermektedir: A, G, C ve T. Bir NLP uygulaması olan *n*-gram, bu 4 harften oluşan bir metin olarak düşündüğümüz DNA dizilerine uygulanmıştır. DNA'yı tanımlayacak *n*-gramları belirlemek için bir üst-*k* stratejisi kullanılmıştır. Buradaki önemli husus tekrarlayan bölgelerin bir DNA'nın karakterizasyonunda ayırt edici özellikler olduğudur. Bu sebeple sık tekrarlanan, yüksek frekanslı metin parçaları DNA'nın karakterizasyonu için değerlidir. Bu yaklaşımla, dizideki tüm *n*-gramlar yerine, sık sık tekrarlanan yüksek frekanslı *n*-gramların kullanılması, tüm *n*-gramları kullanma yükünü büyük ölçüde hafifletecek ve DNA dizisinin karakterizasyonunda özünü kullanmamızı sağlayacaktır. S, DNA dizisindeki bir *n*-gram kümesiyse, $S = (s_1, s_2, s_3, \dots, s_N)$; $\emptyset_{topk}$, *S* listesindeki *n*-gramların frekanslarına göre azalan sıralı listedeki ilk *k*-yüzdedir. Karşılaştırılacak DNA dizilerinin üst-*k* *n*-gram kümeleri $\emptyset_{topk(i)}$ise, bu dizilerin özellik vektörleri aşağıdaki gibi hesaplanır:

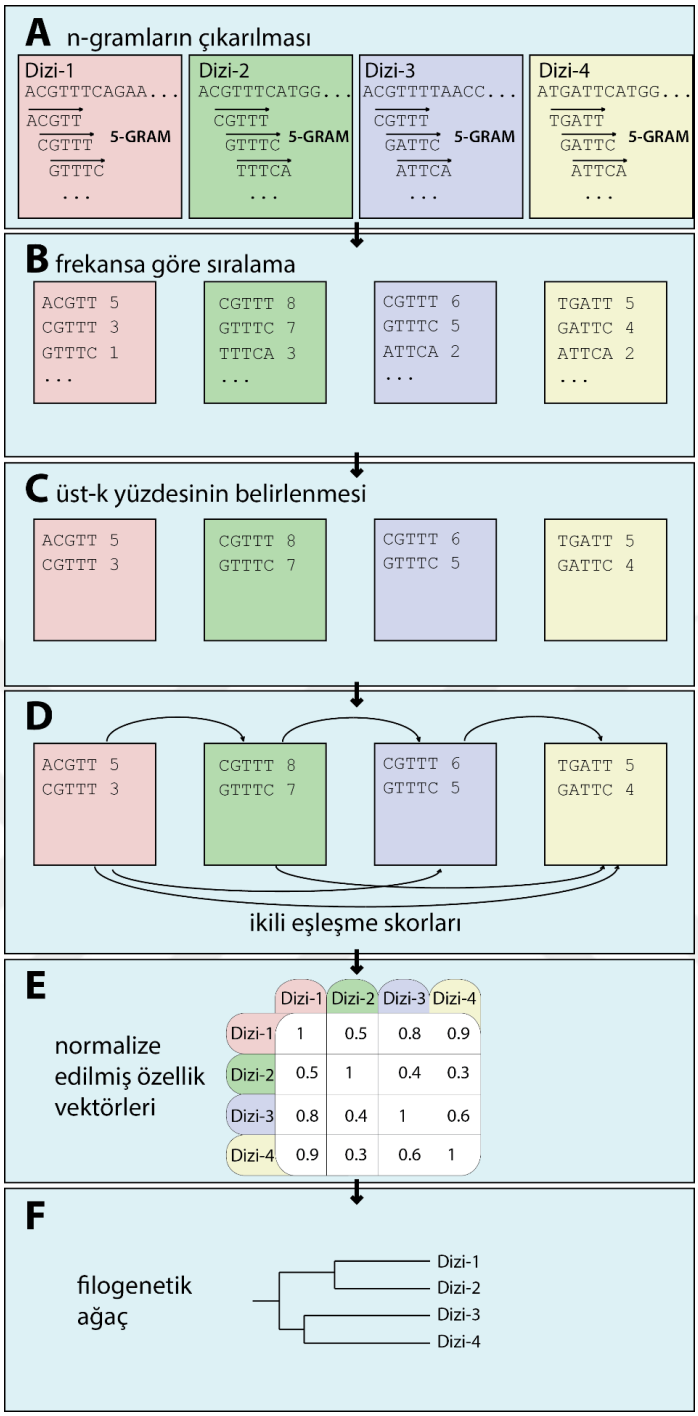
$$Eğer\ \ \emptyset_{topk(i,j)} \in \emptyset_{topk(l)}, bir\_artır(m_{i,l})$$

Burada $i, l = (1, 2, \dots, d)$, d karşılaştırılacak DNA dizisi sayısı; $j = (1, 2, \dots, t), t = N * k$-yüzde; m eşleşme skorudur. $d = 5$ için oluşacak $V_{(i)}$ özellik vektörleri Tablo 4.1'deki gibi oluşacaktır.

Tablo 4.1. Eşleşme skorları ile elde edilen DNA dizilerinin özellik vektörleri

Diziler	**1**	**2**	**3**	**4**	**5**
1	1	$m_{1,2}$	$m_{1,3}$	$m_{1,4}$	$m_{1,5}$
2	$m_{2,1}$	1	$m_{2,3}$	$m_{2,4}$	$m_{2,5}$
3	$m_{3,1}$	$m_{3,2}$	1	$m_{3,4}$	$m_{3,5}$
4	$m_{4,1}$	$m_{4,2}$	$m_{4,3}$	1	$m_{4,5}$
5	$m_{5,1}$	$m_{5,2}$	$m_{5,3}$	$m_{5,4}$	1

DNA dizileri arasındaki benzerliği, *n*-gram frekanslarını doğrudan benzerlik fonksiyonları kullanarak ölçen metotların aksine önerilen metotta bu frekanslar özellik vektörlerinin boyutlarını oluşturan değerler olarak kullanılmıştır. Daha sonra özellik vektörlerinin normalize edilmiş değerleri arasında benzerlik metrikleri kullanarak karşılaştırma yapılmıştır. Önerilen metot sürecini özetleyen görsel Şekil 4.2'de verilmiştir (Delibaş ve ark., 2020).



Şekil 4.2. Üst-k n-gram eşleşmesi tabanlı benzerlik analizi uygulama süreci

Örnek olarak, *AGCTCTACCG* ve *AAAGCTTACTG* dizileri karşılaştırıldığında azalan sırayla frekanslarına göre sıralandığında *S* listeleri aşağıdaki gibi oluşur:

$$S_1 = \{CT: 2, AG: 1; GC: 1;\ TC: 1;\ TA: 1;\ AC: 1; CC: 1; CG: 1\}$$

$$S_2 = \{AA: 2, CT: 2, AG: 1, GC: 1, TT: 1, TA: 1, AC: 1, TG: 1\}$$

Üst-*25 2*-gram için, *CT* ve *AG* S_1 listesinin yüzde 25'lik kısmına denk gelen ve $\emptyset_{top25(1)}$ listesini oluşturan elemanlardır. Bu elemanların $\emptyset_{top25(2)}$ listesinde olup olmadıkları sırayla incelenir. Buna göre *CT* eşleşen tek elemandır ve V_1 vektöründeki ilgili yere $m_{1,2} = 1$ olarak yerleştirilir.

4.3.2. Özellik vektörlerinin normalizasyonu

Eşleşme skorları ile elde edilen d-boyutlu özellik vektörleri $V_{(i)}$'nin normalleştirilmesi gerekmektedir. 0-1 normalizasyon için min-maks normalizasyon metodu kullanılmıştır. Min-maks normalizasyon denklemi Denklem 3.14'te verilmiştir.

4.3.3. Benzerlik hesaplaması

Önerilen metodun uygulanmasıyla lineer uzayda *d*-boyutlu bir özellik vektörü elde edilmiştir. Bu vektör DNA dizileri arasındaki benzerliğin hesaplanmasında kullanılmıştır. DNA dizileri arasındaki benzerliğin hesaplanmasında uzaklık hesaplaması analizin temelidir ve önemli bir adımdır. Elde edilen karakterizasyon arasındaki benzerlik, uç noktalar arasında Denklem 3.10'da verilen Öklid uzaklığı ile hesaplanır.

4.3.4. Filogenetik ağaçların karşılaştırılması

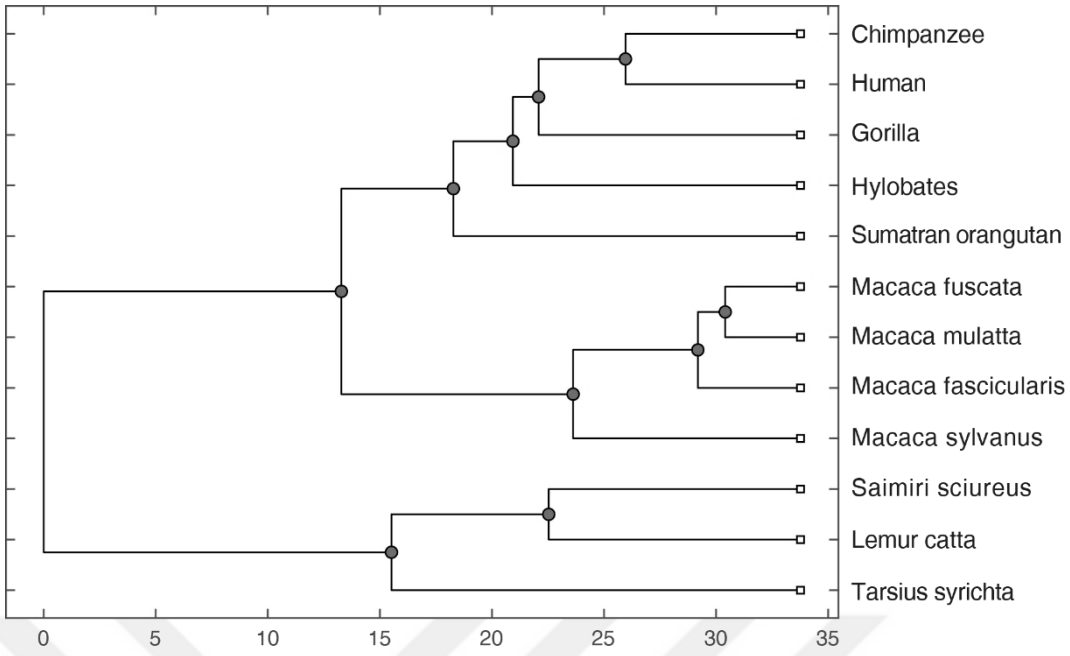
Filogenetik ağaçlar arasındaki topolojik korelasyonu belirlemek için, 0-1aralığına normalleştirilmiş Robinson-Foulds (nRF) (Robinson ve Foulds, 1981) değerlerini kullanarak uyumluluk ölçülmüştür. 0 puanı, incelenen ağaçların uyumlu olduğunu, 1 puanı ise ağaçlar arasında hiçbir uyumun olmadığını göstermektedir. Düşük nRF skorları da iki ağaç arasında yüksek düzeyde uyum olduğunu göstermektedir. Filogenetik ağaçların karşılaştırılması *Newick* formatında ifadeler kullanılarak yapılır. nRF skoru, $Doğruluk = 1 - nRF * 100$ denklemiyle bir yüzdelik dilime dönüştürülmüştür. Böylece, elde ettiğimiz yüzdelik değer, referansta bulunan kenarların hedef ağaçta ne kadar karşılık bulduğunu göstermektedir. Metodumuzun MEGA ile elde edilen referans ağaca göre başarısı "*doğruluk oranı*" olarak ifade edilmiştir.

5. DENEYSEL SONUÇLAR

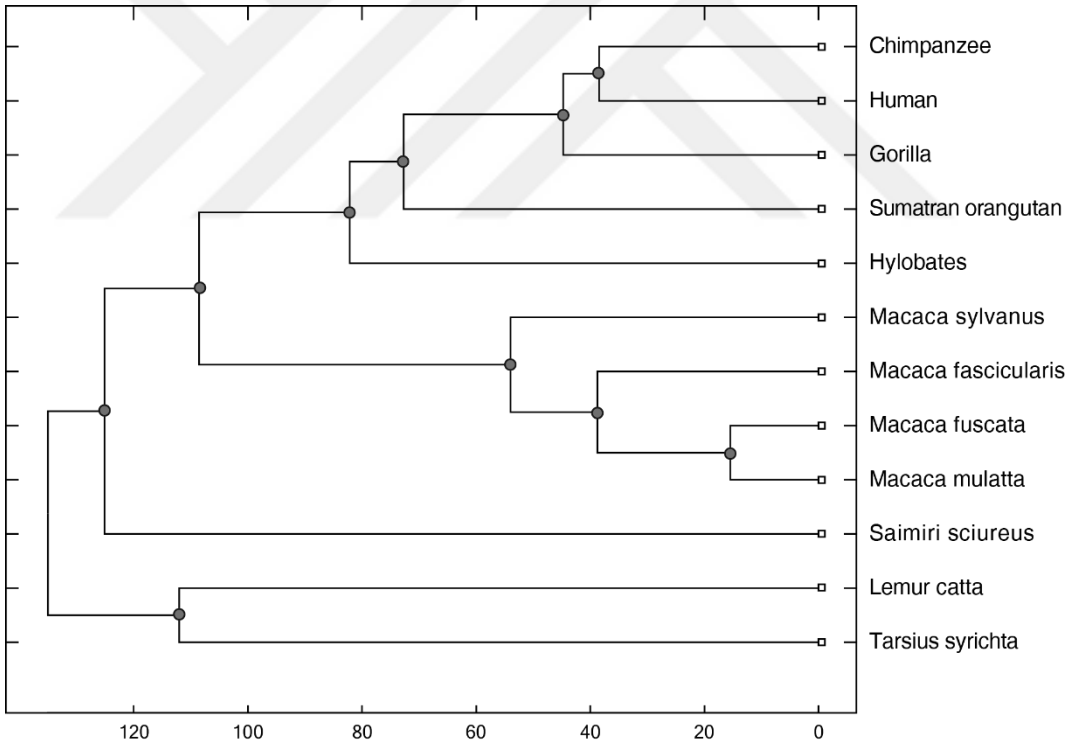
Hizalamasız DNA dizi benzerlik analizi için önerilen metotların sonuçları bu bölümde verilmektedir. Önerilen metotların tamamı Intel Core i7 1.8 GHz İşlemci, 8 GB RAM ile çalışan bir bilgisayarda, MATLAB R2018b uygulama geliştirme aracı üzerinde kodlanmıştır. Yüksek seviyeli programlama dillerinden olan MATLAB yazılımına ait "*MATLAB Statistics and Machine Learning Toolbox*" ve "*MATLAB Bioinformatics Toolbox*" araç kutusu fonksiyonları kullanılmıştır. Yazılan program kodları ile elde edilen özellik vektörlerinin benzerlik ölçümleri bu fonksiyonlarla gerçekleştirilmiştir. Kümelemede kullanılan özellik vektörlerini analiz etmek ve filogenetik ağaçları oluşturmak için "*pdist*" ve "*seqlinkage*" fonksiyonları kullanılmıştır. Fonksiyonların parametreleri olarak; pdist için "*Euclidean*" (varsayılan parametre) ve seqlinkage için "*single*" ve "*average*" kullanılmıştır. Daha sonra filogenetik ağaçların çizimi dendrogramlar ile gerçekleştirilmiştir. Elde edilen ağaçların karşılaştırılması için referans ağaçlar *MEGA7 (Molecular Evolutionary Genetics Analysis)* (Kumar ve ark., 2016b) yazılımı ile oluşturulmuştur.

5.1. Grup Mutasyonlarının Ortalama Konumları ve Nükleotid Bazların Frekanslarına Dayalı Benzerlik Analizi Sonuçları

Önerilen metot Tablo 3.1'de verilen 12 türe ait NADH dehidrogenaz alt birim 4 genlerine uygulanmış ve özellik vektörleri elde edilmiştir. Vektörler arasında daha sonra Öklid uzaklığı kullanılarak benzerlik hesaplaması yapılmıştır. Bu hesaplamadan elde edilen filogenetik ağaç Şekil 5.1'de verilmiştir. Aynı DNA dizilerinin analizi MEGA7 yazılımında hizalama tabanlı ClustalW metodu ile yapılmış, referans filogenetik ağaç oluşturmak için de UPGMA metodu kullanılmıştır. Referans filogenetik ağaç Şekil 5.2'de verilmiştir.



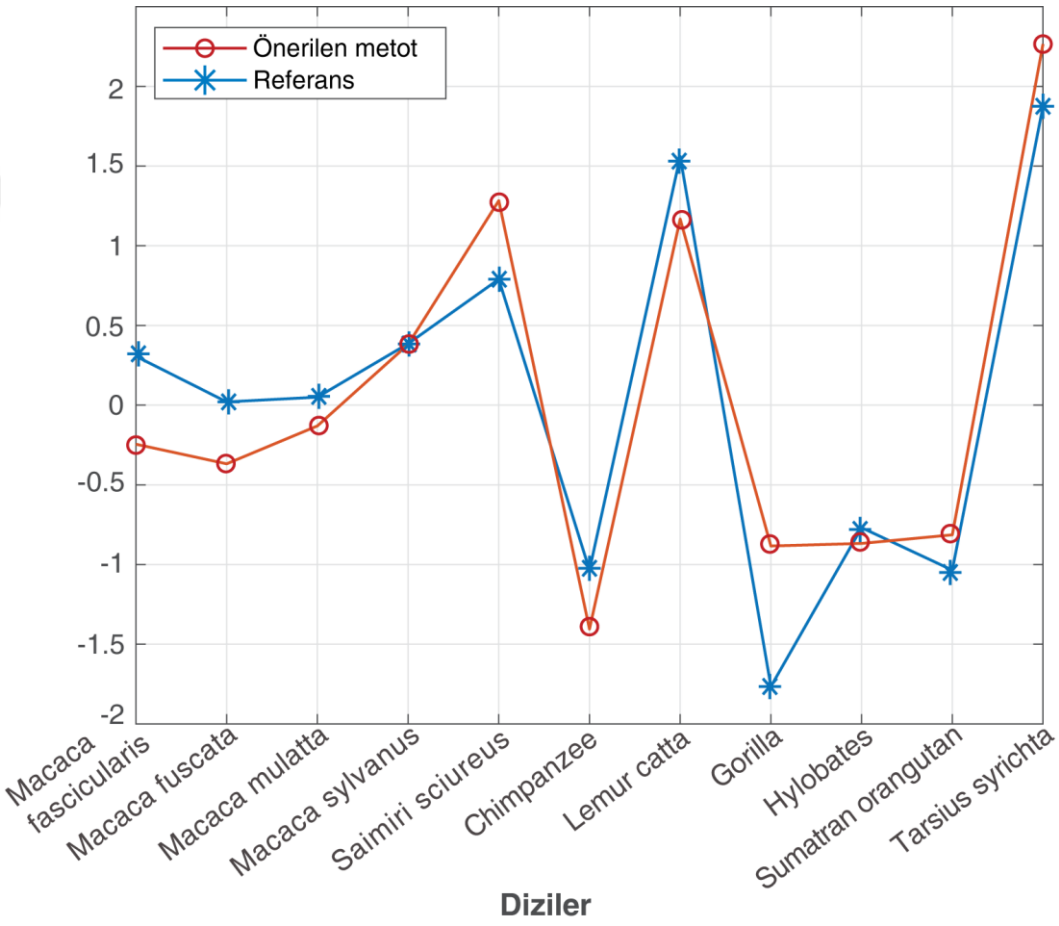
Şekil 5.1. Bölüm 4.1'de önerilen metot ile oluşturulan filogenetik ağaç



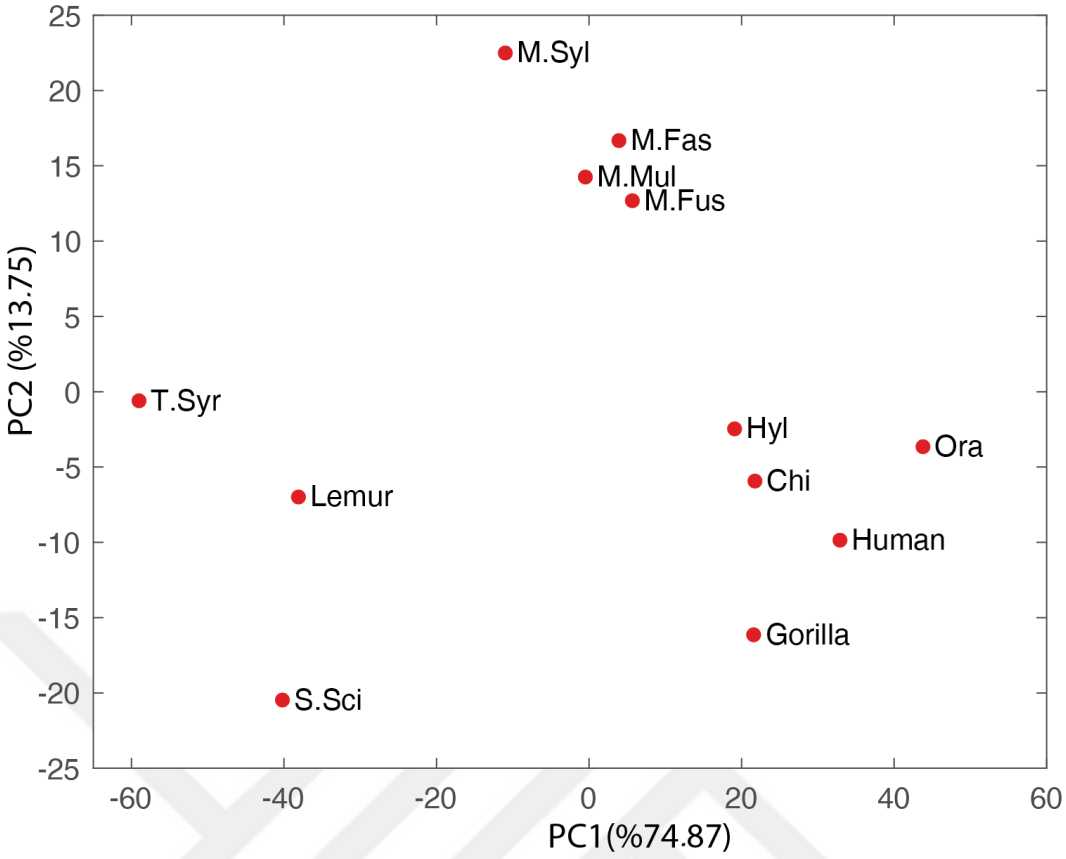
Şekil 5.2. MEGA7 ile ClustalW hizalama ve UPGMA metodu ile elde edilen filogenetik ağaç

Önerilen metotla ve MEGA7 ile üretilen filogenetik ağaçların niteliksel uyumluğu görülmektedir. Bu durumun daha netleştirilmesi ve görselleştirilmesi adına, Şekil 5.3'teki grafik ile rastgele seçilen *Human* türü ve diğer türler arasındaki uzaklıklar, önerilen metodun ve referans analizin sonuçları 0'a merkezlenerek gösterilmiştir. Bu grafikteki

çizgisel uyum, önerilen metodun sonucuyla referans ağacın sonucunun uyumuna işaret etmektedir. 12 özellik vektörüne ait PCA analizi yapılmış, vektörlerin iki ana bileşeni PC1 ve PC2'den oluşan 2-boyutlu özellik uzayına projeksiyonu Şekil 5.4'te sunulmuştur. Bu projeksiyonla, henüz uzaklık hesaplamasına tabi tutulmamış vektörlerin uzayda doğru bir şekilde konumlandığı ve yukarıdaki sonuçlarla uyumlu olduğunu genel olarak görülebilmektedir. İki bileşen, bu veri setinin 7-boyutlu vektörün toplam ataletinin %88.62'sini içermektedir.

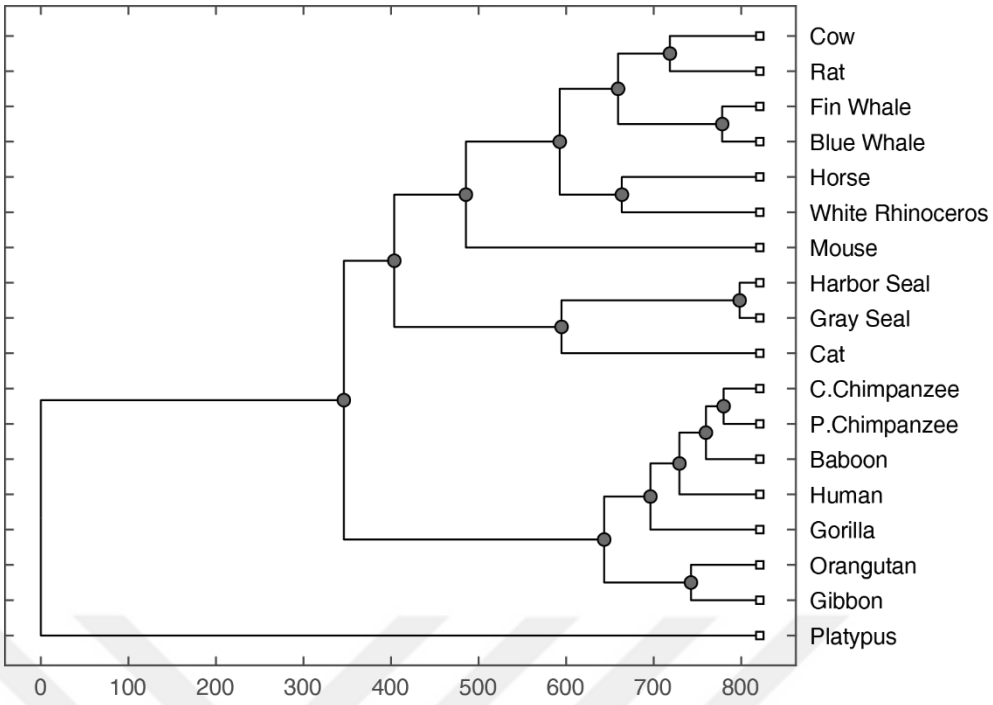


Şekil 5.3. *Human* ve 11 tür arasındaki uzaklık eğrisi

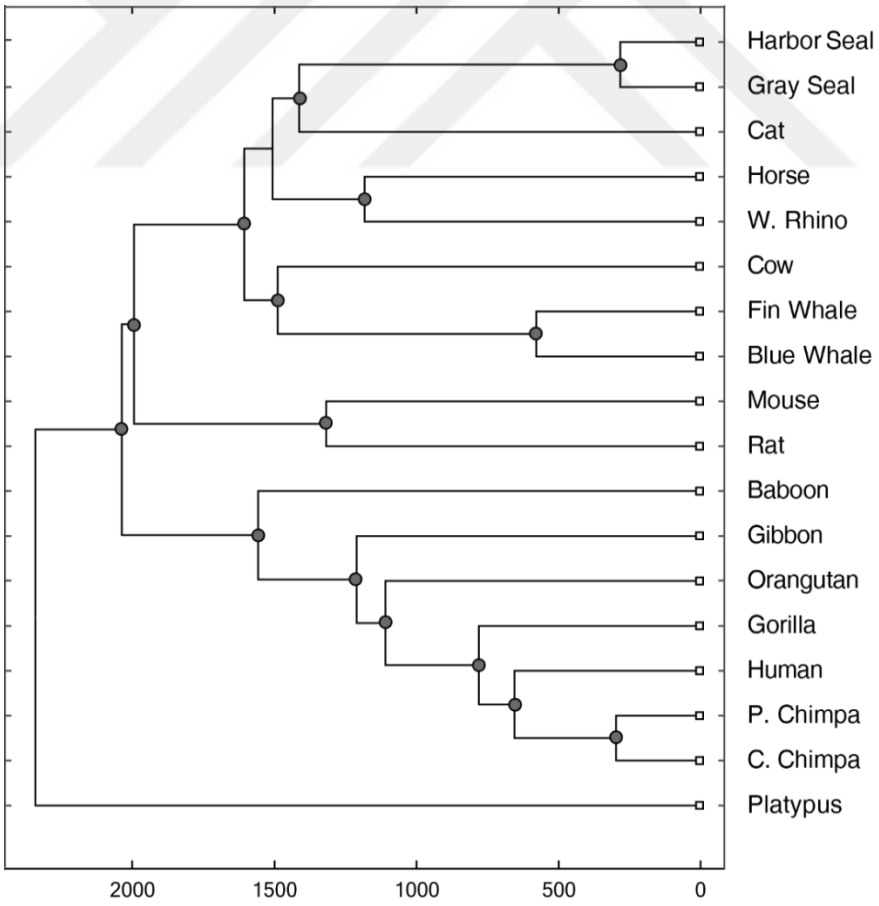


Şekil 5.4. 12 türün 7 boyutlu vektörlerinin iki temel bileşen PC1 ve PC2'den oluşan 2-B uzaya projeksiyonu

Tablo 3.3' de verilen 18 eteneli memeliye ait tüm mitokondriyal DNA veri setine önerilen metodun uygulanması ile özellik vektörleri elde edilmiştir. Vektörler arasında daha sonra Öklid uzaklığı kullanılarak benzerlik hesaplaması yapılmıştır. Bu hesaplamadan elde edilen filogenetik ağaç Şekil 5.5'de verilmiştir. Aynı DNA dizilerinin analizi MEGA7 yazılımında hizalama tabanlı ClustalW metodu ile yapılmış referans filogenetik ağaç oluşturmak için de UPGMA metodu kullanılmıştır. Referans filogenetik ağaç Şekil 5.6'da verilmiştir.

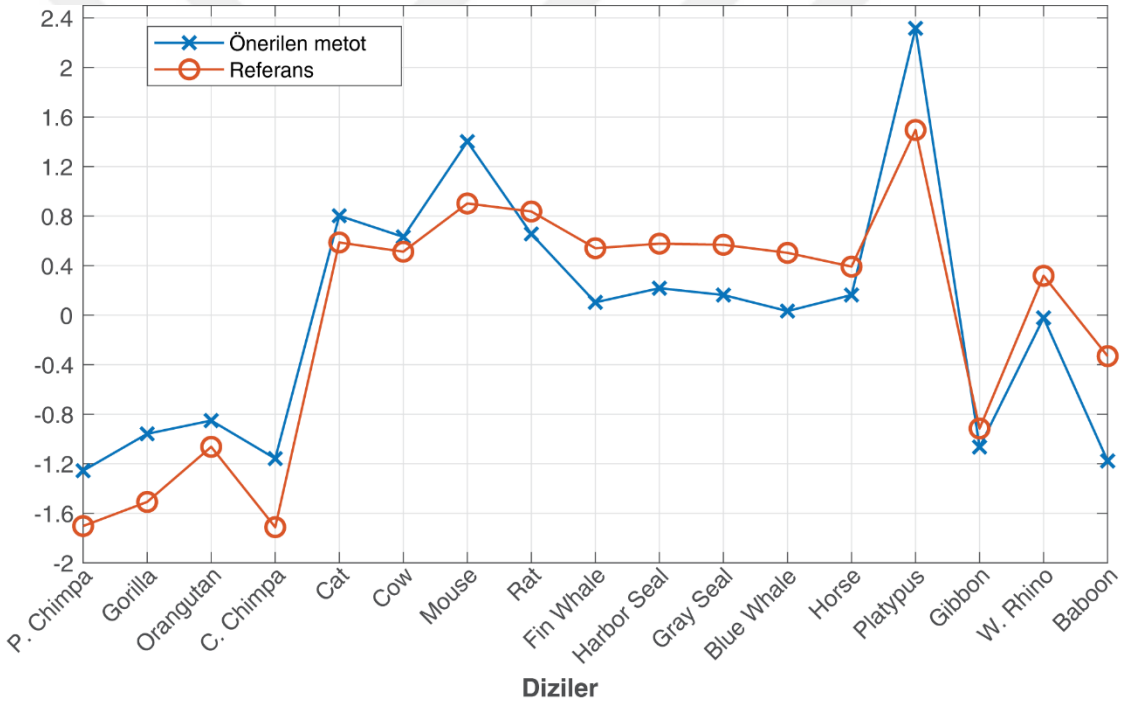


Şekil 5.5. Bölüm 4.1'de önerilen metot ile oluşturulan filogenetik ağaç

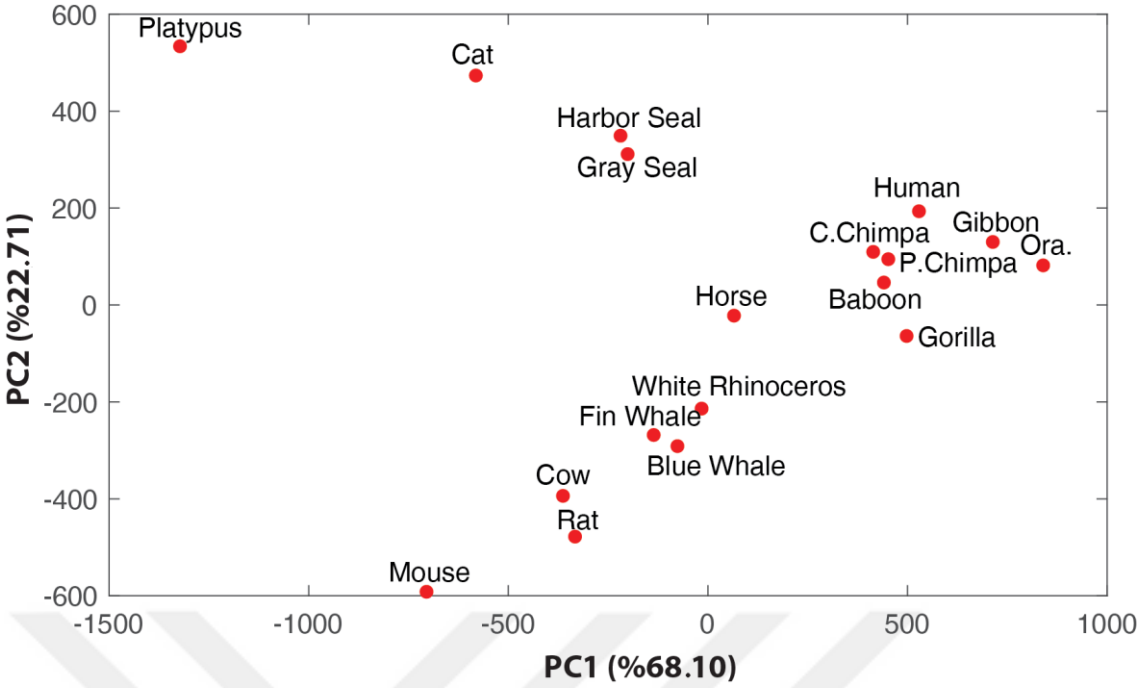


Şekil 5.6. MEGA7 ile ClustalW hizalama ve UPGMA metodu ile elde edilen filogenetik ağaç

Önerilen metotla ve MEGA7 ile üretilen filogenetik ağaçların niteliksel uyumluluğu görülmektedir. Bu durumun daha netleştirilmesi ve görselleştirilmesi adına, Şekil 5.7'deki grafik ile rastgele seçilen *Human* türü ve diğer türler arasındaki uzaklıklar, önerilen metodun ve referans analizin sonuçları 0'a merkezlenerek gösterilmiştir. Bu grafikteki çizgisel uyum, önerilen metodun sonucuyla referans ağacın sonucunun uyumuna işaret etmektedir. 18 özellik vektörüne ait PCA analizi yapılmış, vektörlerin iki ana bileşeni PC1 ve PC2'den oluşan 2-boyutlu özellik uzayına projeksiyonu Şekil 5.8'de sunulmuştur. Bu projeksiyonla, henüz uzaklık hesaplamasına tabi tutulmamış vektörlerin uzayda doğru bir şekilde konumlandığı ve yukarıdaki sonuçlarla uyumlu olduğunu genel olarak görülebilmektedir. İki bileşen, bu veri setinin 7-boyutlu vektörün toplam ataletinin %90,81'sini içermektedir.



Şekil 5.7. *Human* ve 17 tür arasındaki uzaklık eğrisi

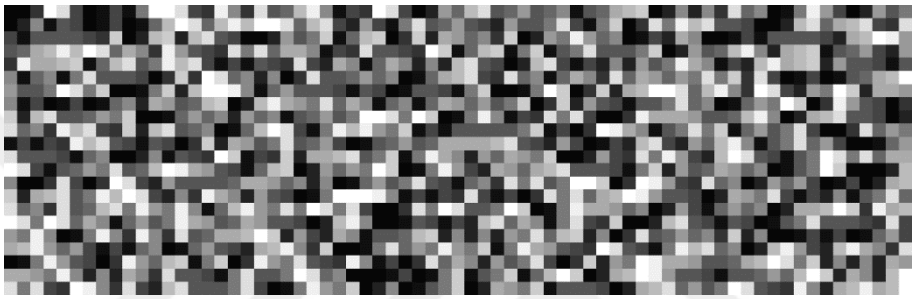


Şekil 5.8. 18 türün 7 Boyutlu vektörlerinin iki temel bileşen PC1 ve PC2'den oluşan 2-B uzaya projeksiyonu

DNA dizi benzerlik analizi metodu seçerken algoritmanın düşük maliyetli olması dikkate alınmalıdır. DNA dizilerinin farklı uzunluklarda olmaları hesaplama maliyetini etkileyen önemli faktörlerdendir. Hesaplama maliyetinin düşürülmesi hizalamasız benzerlik analizi metotlarının birincil hedefidir. Önerdiğimiz bu metodun avantajlarından biri düşük hesaplama maliyetine sahip olmasıdır. Yöntemde 7-boyutlu vektörün oluşturulması için DNA dizisi yalnızca bir kez taranmaktadır. *n* uzunluklu bir DNA dizisinin taranması için gerekli zaman karmaşıklığı *O(n)*, m adet DNA dizisinin karşılaştırma için gerekli vektörlerinin oluşturulması ise *O(mn)* karmaşıklığında olacaktır. Hizalama tabanlı metotlarda; aşamalı hizalama için algoritmalara yapılan çeşitli modifikasyonlarla indirgenmiş $O(m.log_2 m.n^2)$, yinelemeli hizalama için $O(m^2n^2)$ karmaşıklıktan bahsedilmektedir (Baichoo ve Ouzounis, 2017). Hizalamasız metotlarda ise en az doğrusal karmaşıklıkla çözüme ulaşıldığı göz önünde bulundurulduğunda önerilen metodun performansı kabul edilir sonuçlar üretmiştir (Schwende ve Pham, 2013).

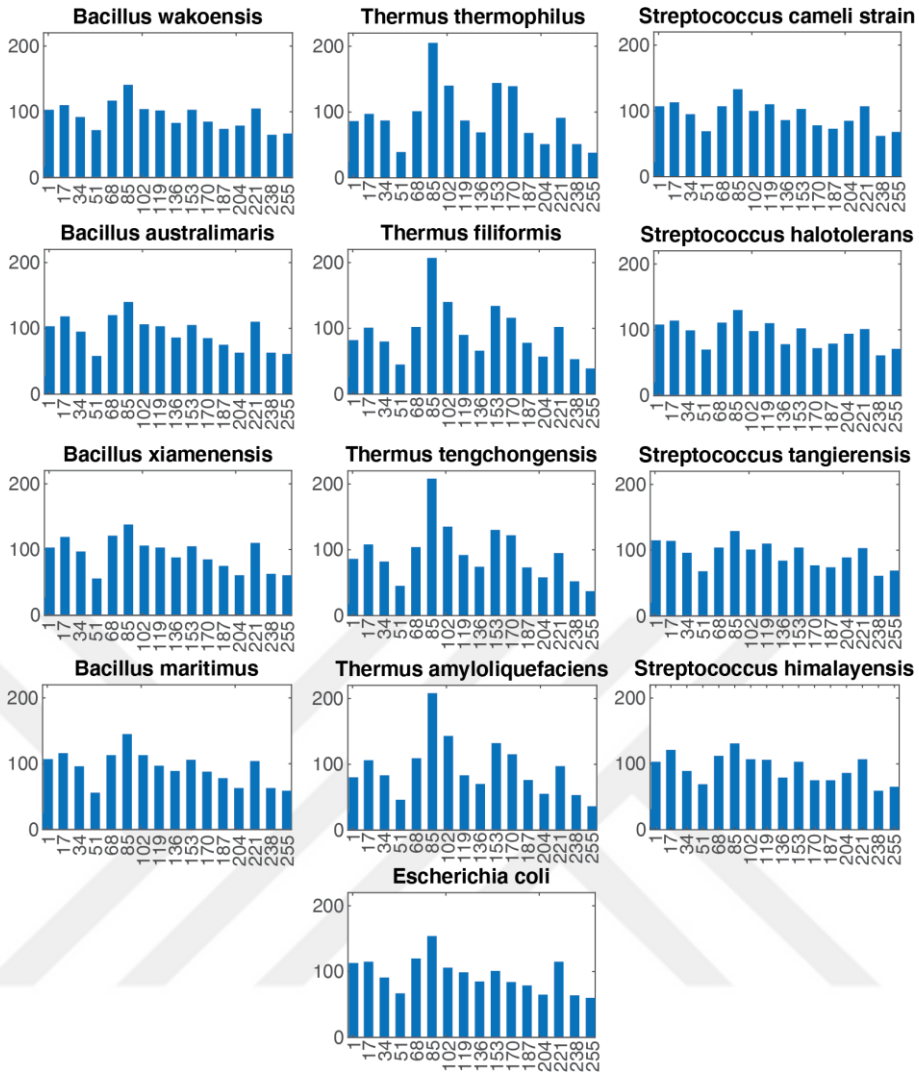
5.2. Birinci Derece İstatistik Tabanlı Görüntü Doku Analizi Kullanarak Benzerlik Analizi Sonuçları

Önerilen metodun birinci aşamasında analizi yapılacak DNA dizileri birer gri-seviye görüntüye dönüştürülmektedir. Tablo 3.2’ de verilen 13 bakteriye ait 16S ribozomal DNA dizisine metodun uygulanması ile her bir DNA dizisi için bir görüntü elde edilmiştir. Elde edilen görüntülere örnek olarak Şekil 3.7’de verilen FASTA formatındaki DNA dizisinin gri-seviye görüntüye dönüştürülmüş hali Şekil 5.9’da verilmiştir.



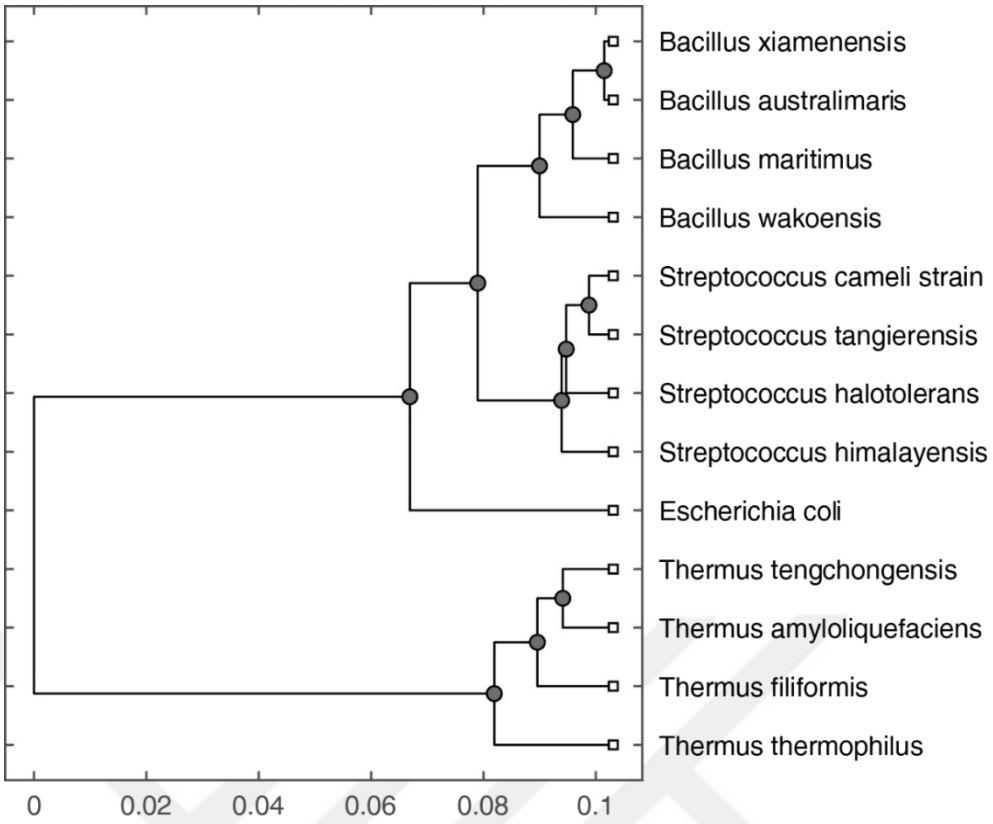
Şekil 5.9. Şekil 3.7'deki DNA dizisinden dönüştürülmüş gri-seviye görüntü

Elde edilen görüntüye uygulanan doku analizi ile histogram hesaplaması yapılmıştır. Tablo 3.2’de verilen DNA dizilerinden oluşturulmuş görüntülerin histogramları Şekil 5.10’da verilmiştir. DNA dizileri arasındaki grupların topolojik yapısı histogramların benzerliğinden kabaca görülebilmektedir.

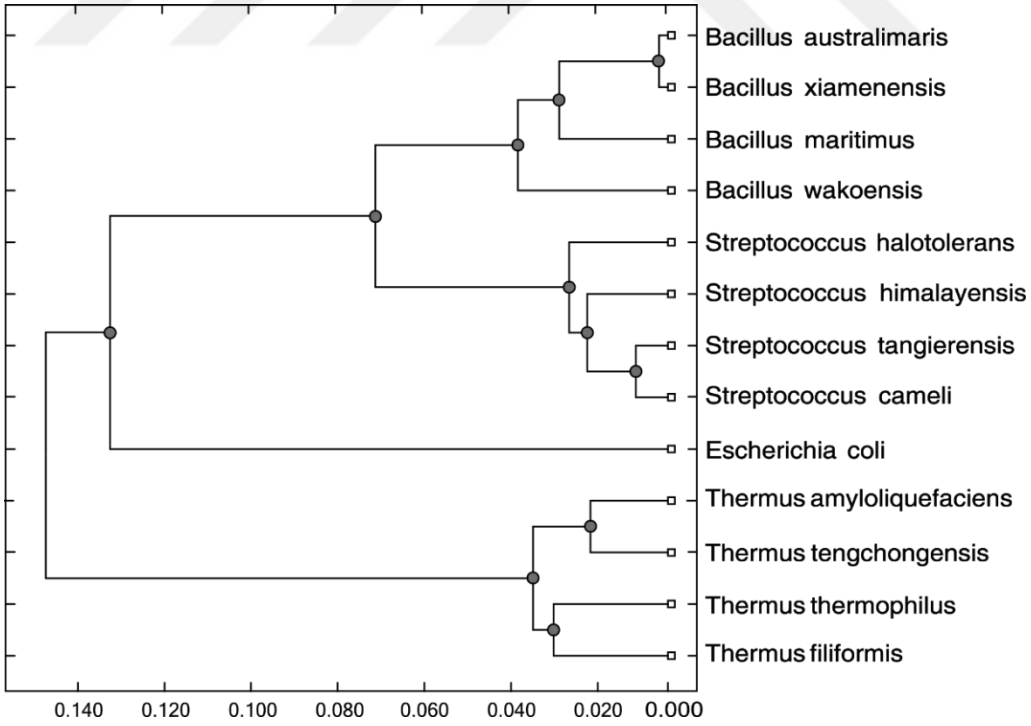


Şekil 5.10. Tablo 3.2'deki DNA dizilerinden dönüştürülmüş görüntülere ait histogramlar

Elde edilen histogramlar kullanılarak doku metrikleri hesaplanmış ve özellik vektörleri oluşturulmuştur. Vektörler arasında daha sonra Öklid uzaklığı kullanılarak benzerlik hesaplaması yapılmıştır. Bu hesaplamadan elde edilen filogenetik ağaç Şekil 5.11'de verilmiştir. Aynı DNA dizilerinin analizi MEGA7 yazılımında hizalama tabanlı ClustalW metodu ile yapılmış, referans filogenetik ağaç oluşturmak için de UPGMA metodu kullanılmıştır. Referans filogenetik ağaç Şekil 5.12' de verilmiştir.



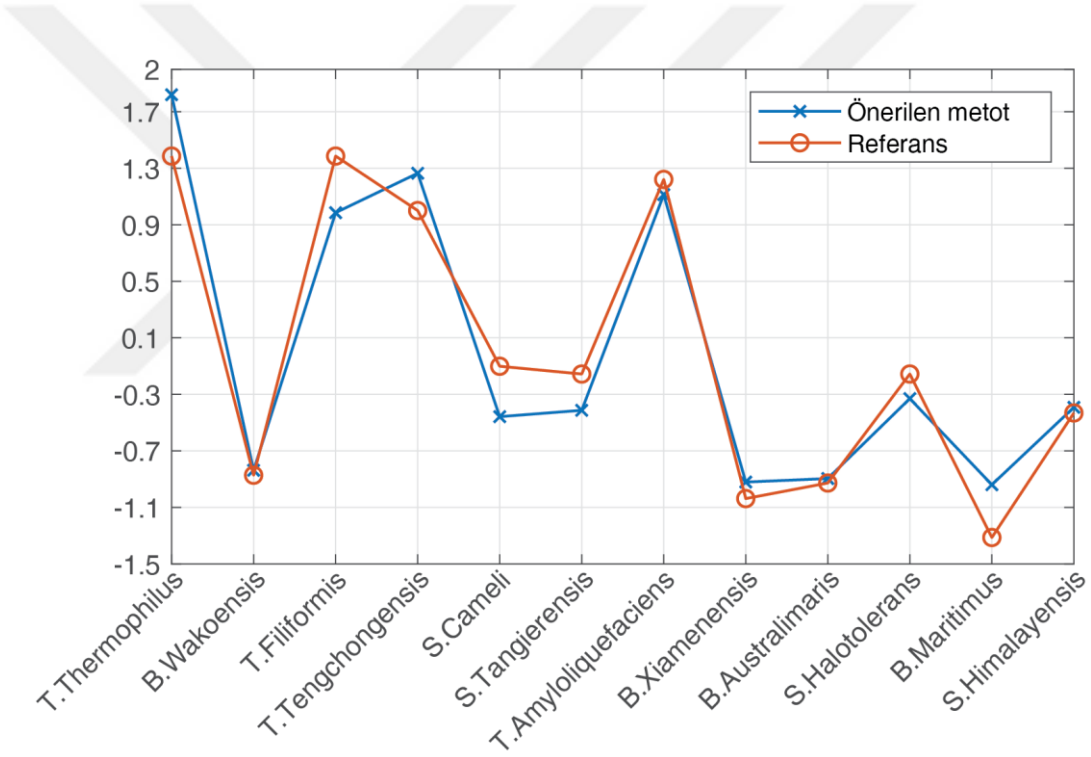
Şekil 5.11. Bölüm 4.2'de önerilen metot ile elde edilen filogenetik ağaç



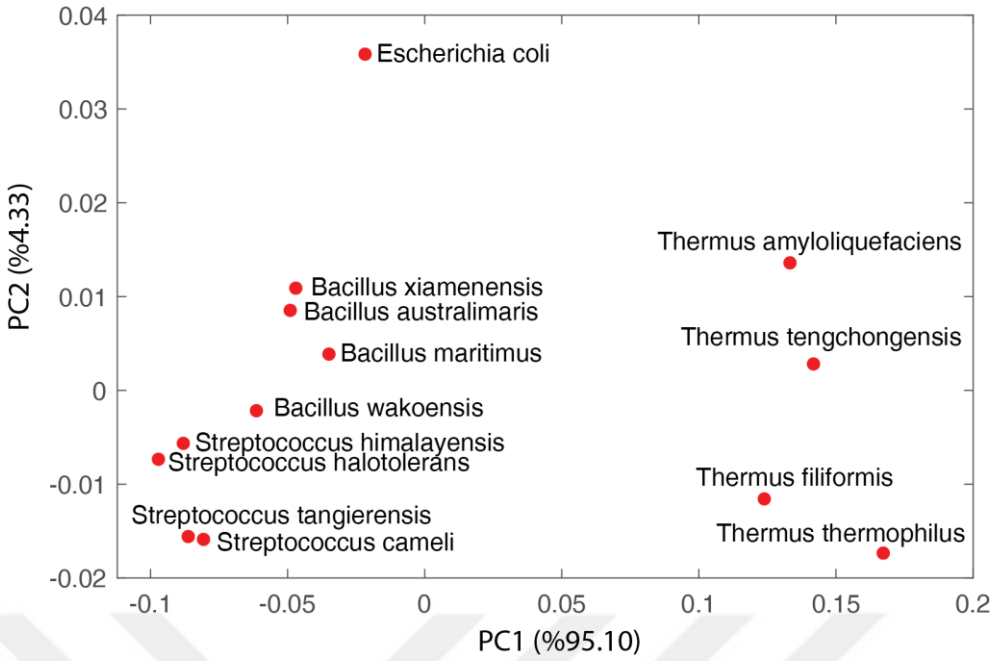
Şekil 5.12. MEGA7 ile ClustalW hizalama ve UPGMA metodu ile elde edilen filogenetik ağaç

Doğruluk göz önüne alındığında, önerilen yöntemle üretilen filogenetik ağaç, referans filogenetik ağaçla topolojik olarak tutarlıdır. Ayrıca, önerilen yöntemin benzerlik

matrisi ve MEGA7 tarafından üretilen referans ağacının niteliksel olarak uyumu hakkındaki genel örtüşmeyi de görebiliriz. Bunu ortaya çıkarmak ve görselleştirmek için, Şekil 5.13'teki grafik ile rastgele seçilen *Escherichia coli* türü ve diğer türler arasındaki uzaklıklar, önerilen metodun ve referans analizin sonuçları 0'a merkezlenerek gösterilmiştir. Bu grafikteki çizgisel uyum, önerilen metodun sonucuyla referans ağacın sonucunun uyumuna işaret etmektedir. Ayrıca 13 özellik vektörüne ait PCA analizi yapılmış, vektörlerin iki ana bileşeni PC1 ve PC2'den oluşan 2-boyutlu özellik uzayına projeksiyonu Şekil 5.14'te sunulmuştur. Bu projeksiyonla, henüz uzaklık hesaplamasına tabi tutulmamış vektörlerin uzayda doğru bir şekilde konumlandığı ve yukarıdaki sonuçlarla uyumlu olduğu genel olarak görülebilmektedir. İki bileşen, bu veri setinin 4-boyutlu vektörün toplam ataletinin %99.43'ünü içermektedir.

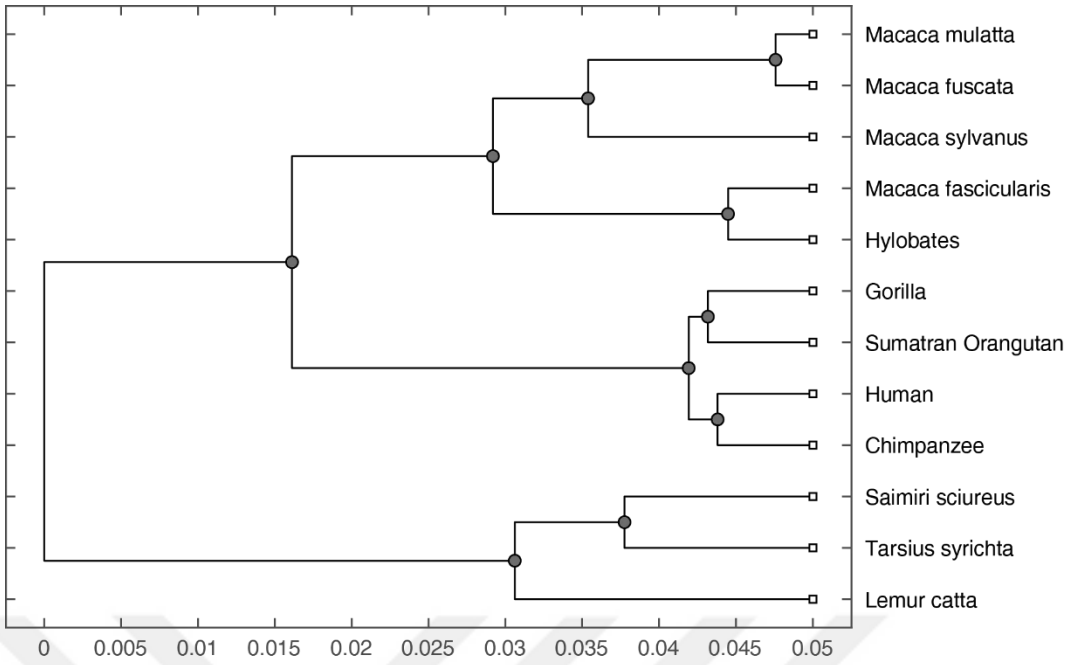


Şekil 5.13. *Escherichia coli* ve 12 bakteri arasındaki uzaklık eğrisi

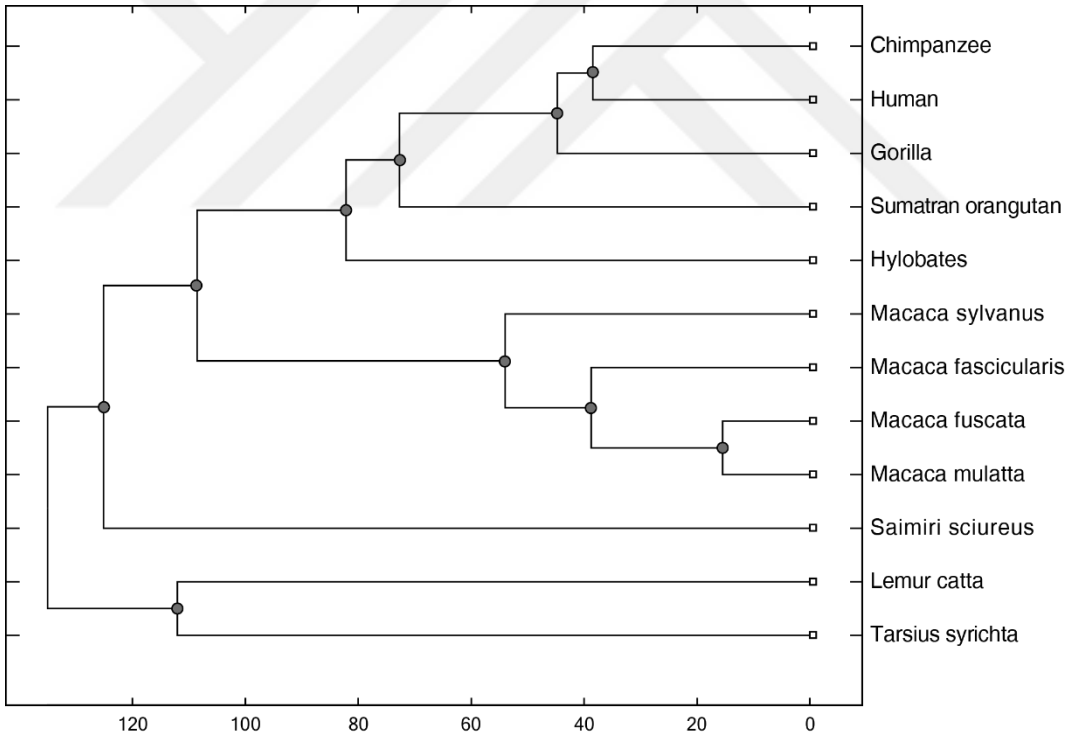


Şekil 5.14.13 bakterinin 4 boyutlu vektörlerinin iki temel bileşen PC1 ve PC2'den oluşan 2-B uzaya projeksiyonu

Tablo 3.1' de verilen 12 türe ait NADH dehidrogenaz alt birim 4 genleri veri setine önerilen metodun uygulanması ile gri-seviye görüntüler oluşturulmuş ve bu görüntülerin histogramları çıkarılmıştır. Histogram metriklerinden elde edilen vektörler arasında daha sonra Öklid uzaklığı kullanılarak benzerlik hesaplaması yapılmıştır. Bu hesaplamadan elde edilen filogenetik ağaç Şekil 5.15'te verilmiştir. Aynı DNA dizilerinin analizi MEGA7 yazılımında hizalama tabanlı ClustalW metodu ile yapılmış referans filogenetik ağaç oluşturmak için de UPGMA metodu kullanılmıştır. Referans filogenetik ağaç Şekil 5.16'da verilmiştir.



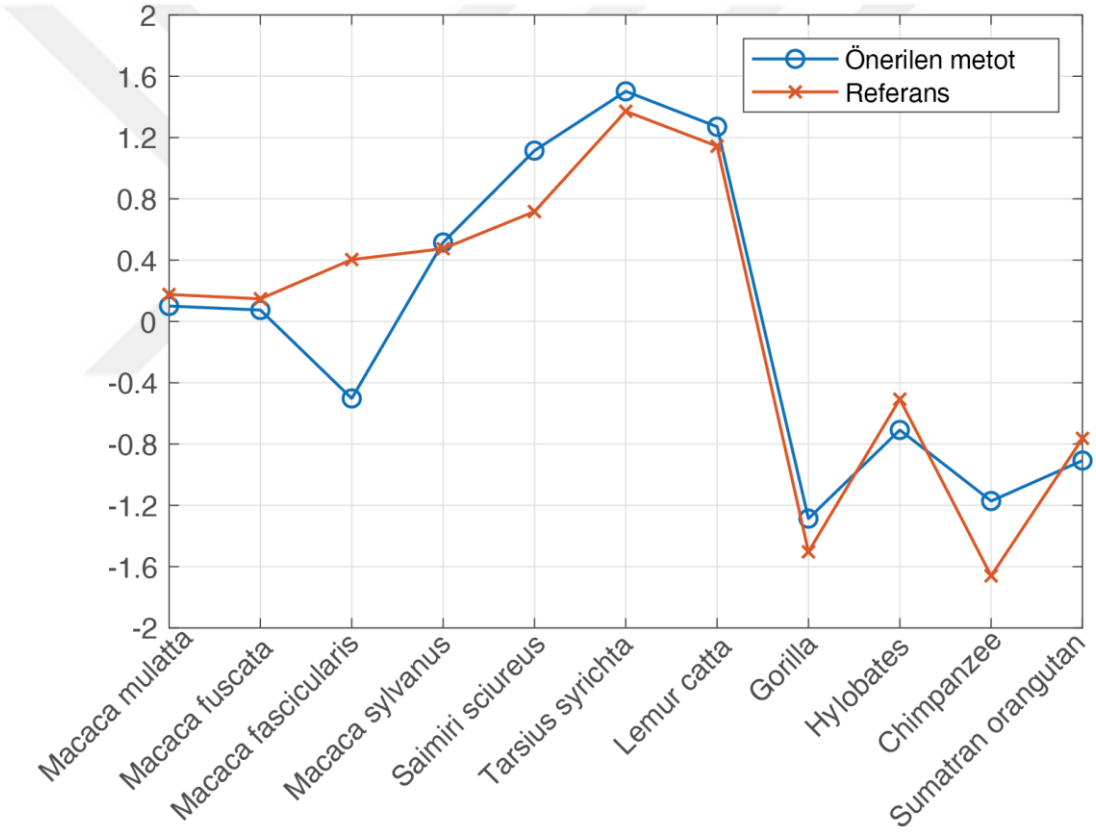
Şekil 5.15. Bölüm 4.2'de önerilen metotla elde edilen filogenetik ağaç



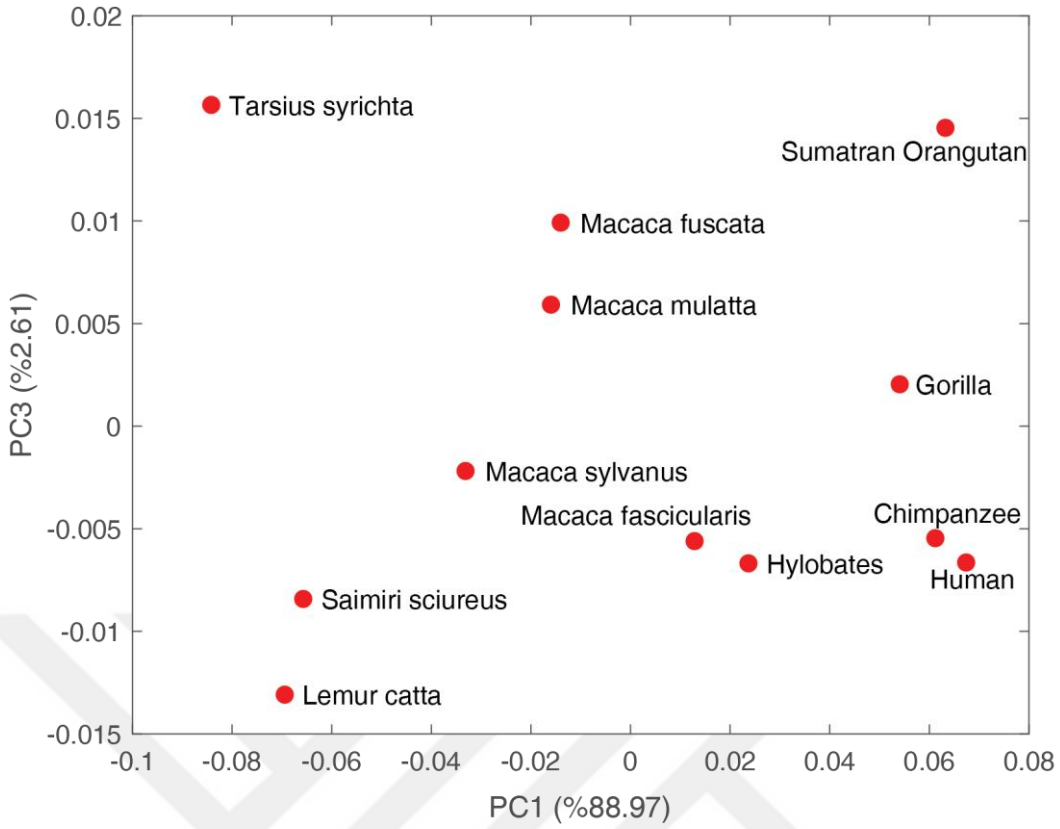
Şekil 5.16. MEGA7 ile ClustalW hizalama ve UPGMA metodu ile elde edilen filogenetik ağaç

Doğrulukla ilgili değerlendirme yapıldığında, önerilen yöntemle üretilen filogenetik ağacın, referans filogenetik ağaçla topolojik olarak tutarlı sonuç ürettiği gözlemlenmiştir. Ayrıca, önerilen yöntemin benzerlik matrisi ve MEGA7 tarafından üretilen referans ağacın niteliksel olarak uyumu hakkındaki genel örtüşmeyi de

görebiliriz. Bunu ortaya çıkarmak ve görselleştirmek için, Şekil 5.17'deki grafik ile rastgele seçilen *Human* türü ve diğer türler arasındaki uzaklıklar, önerilen metodun ve referans analizin sonuçları 0'a merkezlenerek gösterilmiştir. Bu grafikteki çizgisel uyum, önerilen metodun sonucuyla referans ağacın sonucunun uyumuna işaret etmektedir. Ayrıca 12 özellik vektörüne ait PCA analizi yapılmış, vektörlerin iki ana bileşeni PC1 ve PC3'ten oluşan 2-boyutlu özellik uzayına projeksiyonu Şekil 5.18'de sunulmuştur. Bu projeksiyonla, henüz uzaklık hesaplamasına tabi tutulmamış vektörlerin uzayda doğru bir şekilde konumlandığı ve yukarıdaki sonuçlarla uyumlu olduğunu genel olarak görülebilmektedir. İki bileşen, bu veri setinin 4-boyutlu vektörün toplam ataletinin %91.58'ünü içermektedir.

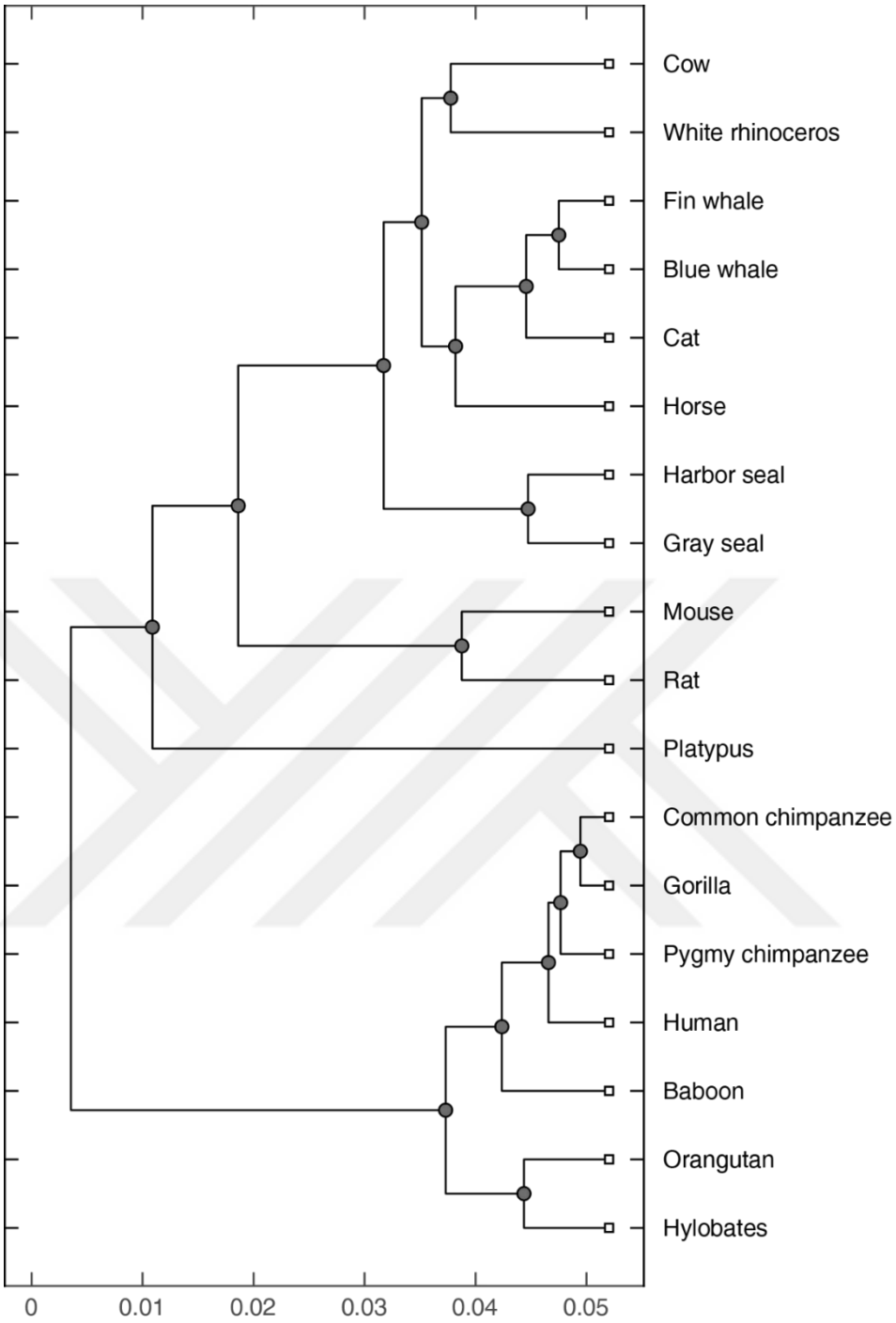


Şekil 5.17. *Human* ve 11 tür arasındaki uzaklık eğrisi

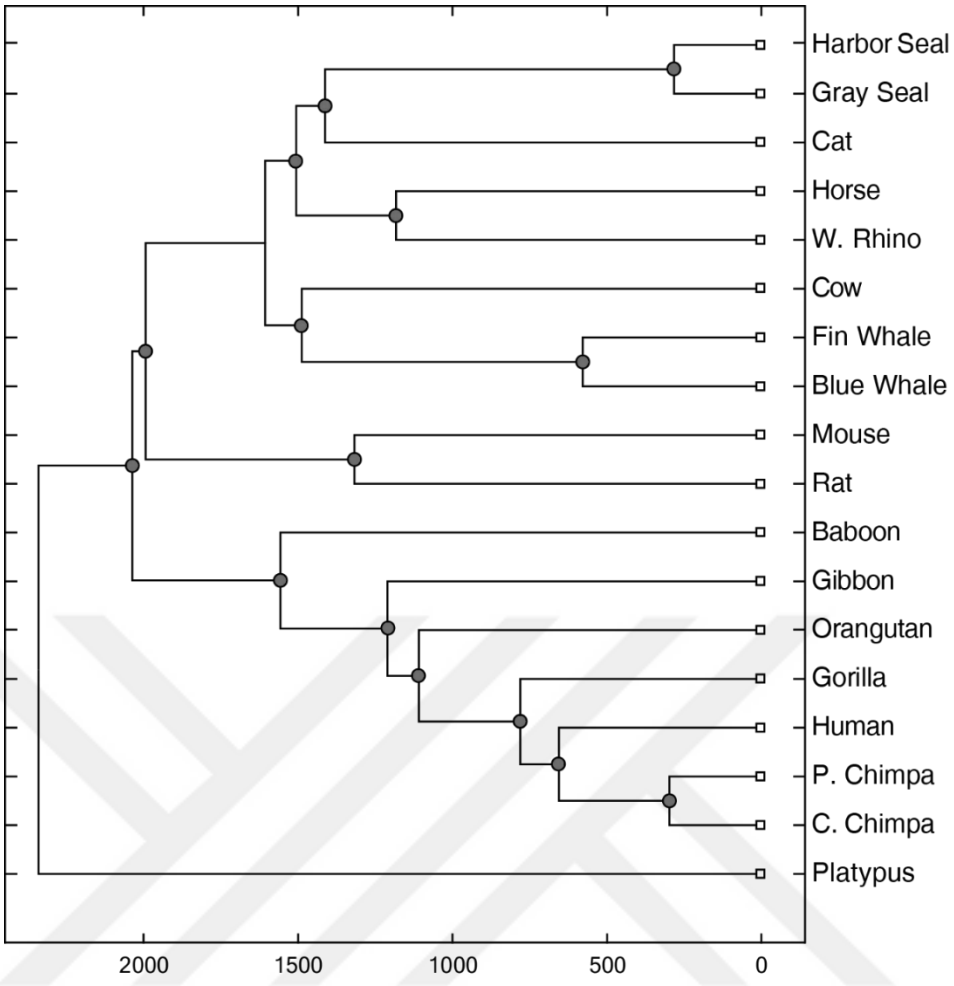


Şekil 5.18. 12 türe ait 4-boyutlu vektörlerinin iki temel bileşen PC1 ve PC3'ten oluşan 2-B uzaya projeksiyonu

Tablo 3.3'te verilen 18 eteneli memeliye ait tüm mitokondriyal DNA veri setine önerilen metodun uygulanması ile gri-seviye görüntüler oluşturulmuş ve bu görüntülerin histogramları çıkartılmıştır. Histogram metriklerinden elde edilen vektörler arasında daha sonra Öklid uzaklığı kullanılarak benzerlik hesaplaması yapılmıştır. Bu hesaplamadan elde edilen filogenetik ağaç Şekil 5.19'da verilmiştir. Aynı DNA dizilerinin analizi MEGA7 yazılımında hizalama tabanlı ClustalW metodu ile yapılmış referans filogenetik ağaç oluşturmak için de UPGMA metodu kullanılmıştır. Referans filogenetik ağaç Şekil 5.20'de verilmiştir.

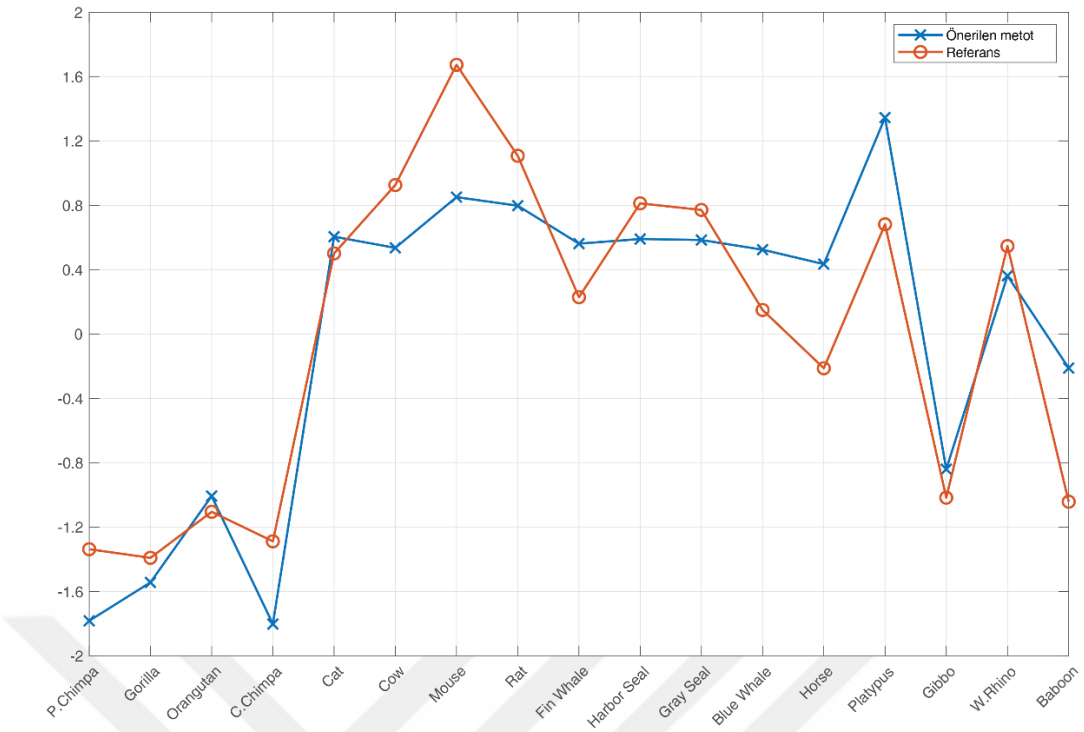


Şekil 5.19. Önerilen Metot ile Elde Edilen Filogenetik Ağaç

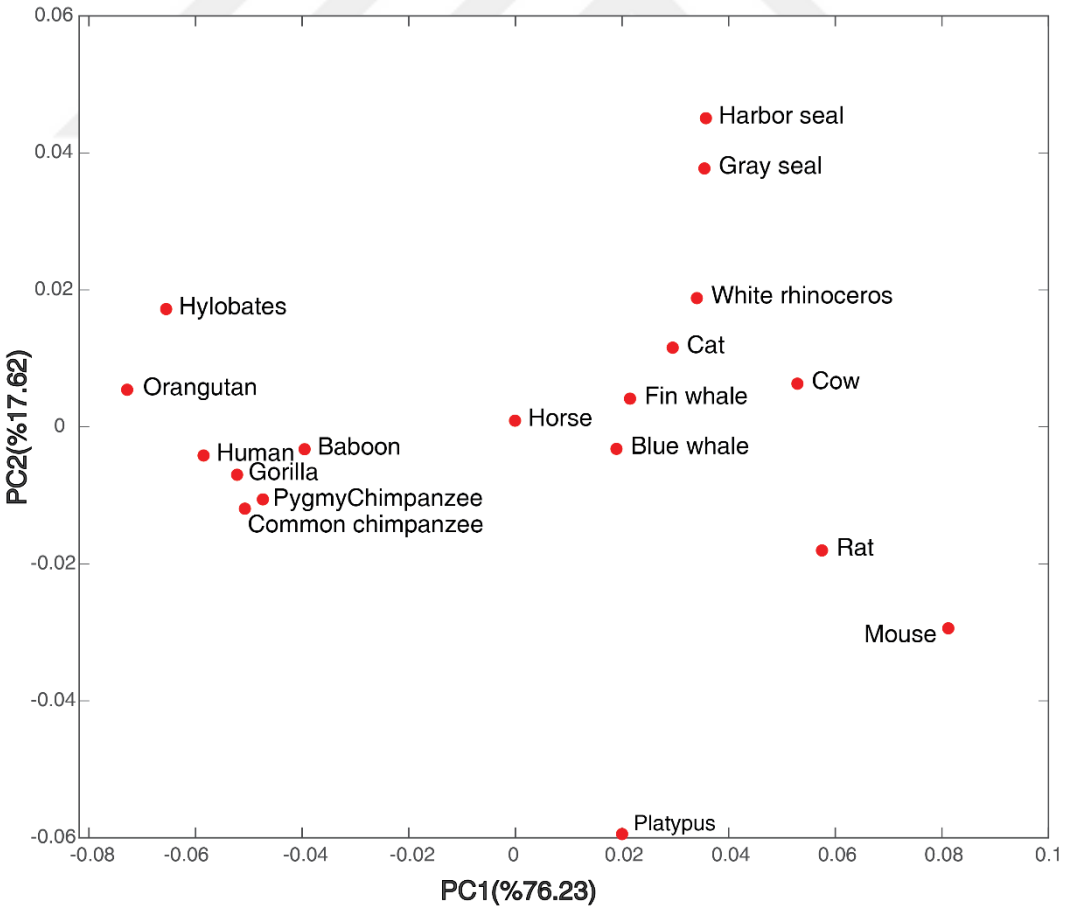


Şekil 5.20. MEGA7 ile ClustalW Hizalama ve UPGMA Metodu ile Elde Edilen Filogenetik Ağaç

Önerilen metotla ve MEGA7 ile üretilen filogenetik ağaçların niteliksel uyumluğu görülmektedir. Bu durumun daha netleştirilmesi ve görselleştirilmesi adına, Şekil 5.21'deki grafik ile rastgele seçilen *Human* türü ve diğer türler arasındaki uzaklıklar, önerilen metodun ve referans analizin sonuçları 0'a merkezlenerek gösterilmiştir. Bu grafikteki çizgisel uyum, önerilen metodun sonucuyla referans ağacın sonucunun uyumuna işaret etmektedir. 18 özellik vektörüne ait PCA analizi yapılmış, vektörlerin iki ana bileşeni PC1 ve PC2'den oluşan 2-boyutlu özellik uzayına projeksiyonu Şekil 5.22'de sunulmuştur. Bu projeksiyonla, henüz uzaklık hesaplamasına tabi tutulmamış vektörlerin uzayda doğru bir şekilde konumlandığı ve yukarıdaki sonuçlarla uyumlu olduğu genel olarak görülebilmektedir. İki bileşen, bu veri setinin 4-boyutlu vektörün toplam ataletinin %93.85'sini içermektedir.



Şekil 5.21. *Human* ve 17 tür arasındaki uzaklık eğrisi



Şekil 5.22. 18 türe Ait 4-boyutlu vektörlerinin iki temel bileşen PC1 ve PC3'ten oluşan 2-B uzaya projeksiyonu

Bilgisayar biliminde sınıflandırma, tanıma veya bölümleme gibi uygulamalarda benzerlik analizi kullanılmaktadır. Önerilen bu metotta farklı benzerlik analizi uygulamalarının DNA benzerliğine uygulanabilirliğini değerlendirmek amacıyla yeni bir yaklaşıma yer verilmiştir. Görüntü işleme alanının bir alt çalışma alanı olan doku analizi DNA benzerlik analizine uyarlanmıştır. İçeriğinde belirgin şekiller barındırmayan dokularda düzenli ve tekrarlı özellikteki desenler dokunun karakteristiğini belirlemektedir. Benzer durum da DNA dizileri içerisindeki tekrarlı bölgeler için geçerlidir. Nükleotidlerin oluşturduğu bu desenler DNA dizilerini karakterize etmiş ve referans sonuçlarla tutarlı çıktılar elde edilmiştir.

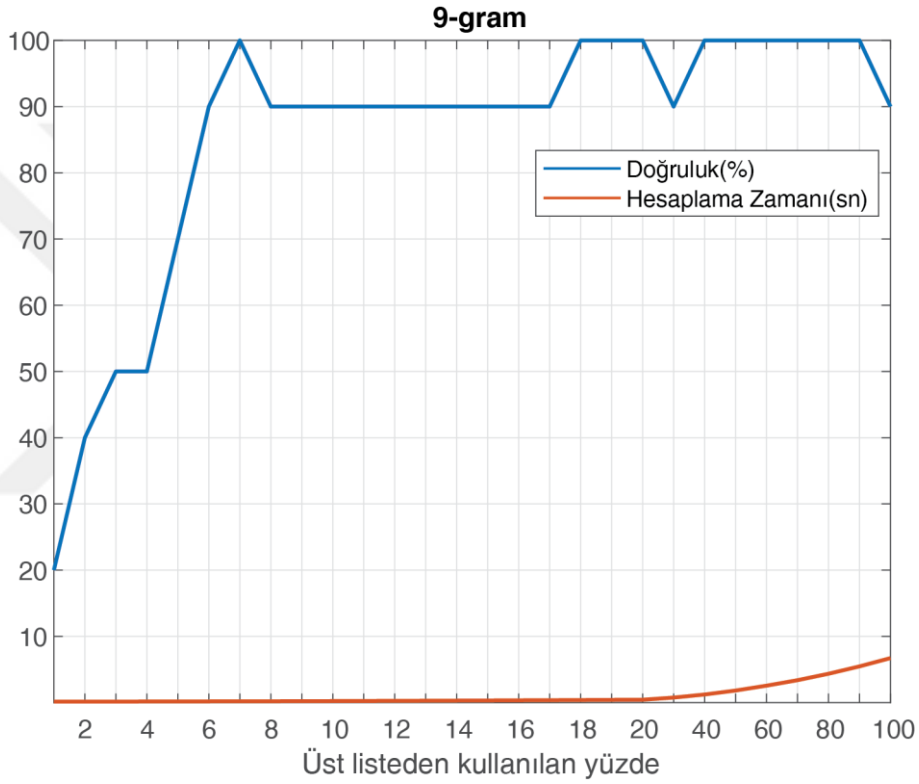
5.3. Üst-*k* *n*-gram Eşleşmesi Tabanlı Benzerlik Analizi Sonuçları

Bölüm 4.3'te önerilen metodun uygulamasında farklı n-gram ve üst-*k* değerleri kullanılmıştır. Yöntem parametreleri olarak *3*-gram ile *20*-gram arası çalıştırılmıştır. Daha yüksek değerlerde hesaplama maliyeti çok yüksek olacağından deneme aralığına alınmamıştır. Yüksek frekanslara sahip gramların DNA dizisini karakterize etmede etkili ve yeterli olacağı öngörüsüyle ve işlem hacmini azaltmak adına üst-k aralığı %1 - %20 aralığında birer, %20 - %100 aralığında 10'ar artış ile denenmiştir.

Metodun uygulamasında ilk olarak Tablo 3.1'de verilen 12 türe ait NADH dehidrogenaz alt birim 4 genleri üzerinde analiz gerçekleştirilmiştir. Analizde *3*-gram ile *20*-gram arası ve üst-*1* ile üst-*100* arası yüzdelikler uygulanmıştır. Uygulamanın sonucunda en yüksek doğruluk oranını elde eden ilk 5 beş sonuç Tablo 5.1'de verilmiştir. Testlerin yapıldığı aralıkta en yüksek doğruluk oranına erişmiş başka sonuçlar da söz konusudur. Ancak buradaki amaç, en kısa hesaplama süresinde en yüksek doğruluğa ulaşan durumları belirlemektir. En iyi sonucu veren durum için çıkan doğruluk-hesaplama zamanı grafiği Şekil 5.23'te verilmiştir. Doğruluk değeri nRF skoruna göre yüzdelik karşılığa dönüştürülmüştür. Test edilen diğer tüm üst-*k* *n*-gram durumları Ek-1'de verilmiştir.

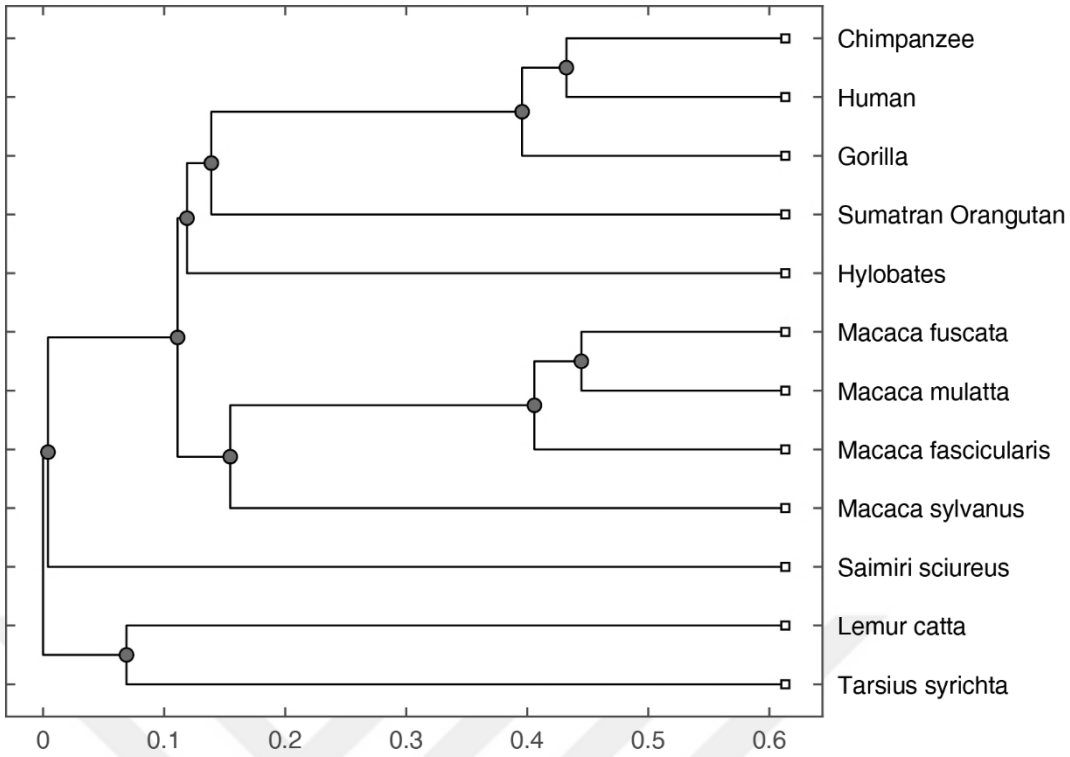
Tablo 5.1. 12 türe ait NADH dehidrogenaz alt-birim 4 genleri için üst-*k* *n*-gram doğruluk ve hesaplama zaman grafiği

n-gram	Üst-k (yüzde)	Doğruluk	Hes. Zamanı (sn)
9	7	100	0.1878
10	8	100	0.2023
13	8	100	0.2040
10	9	100	0.2143
13	9	100	0.2149

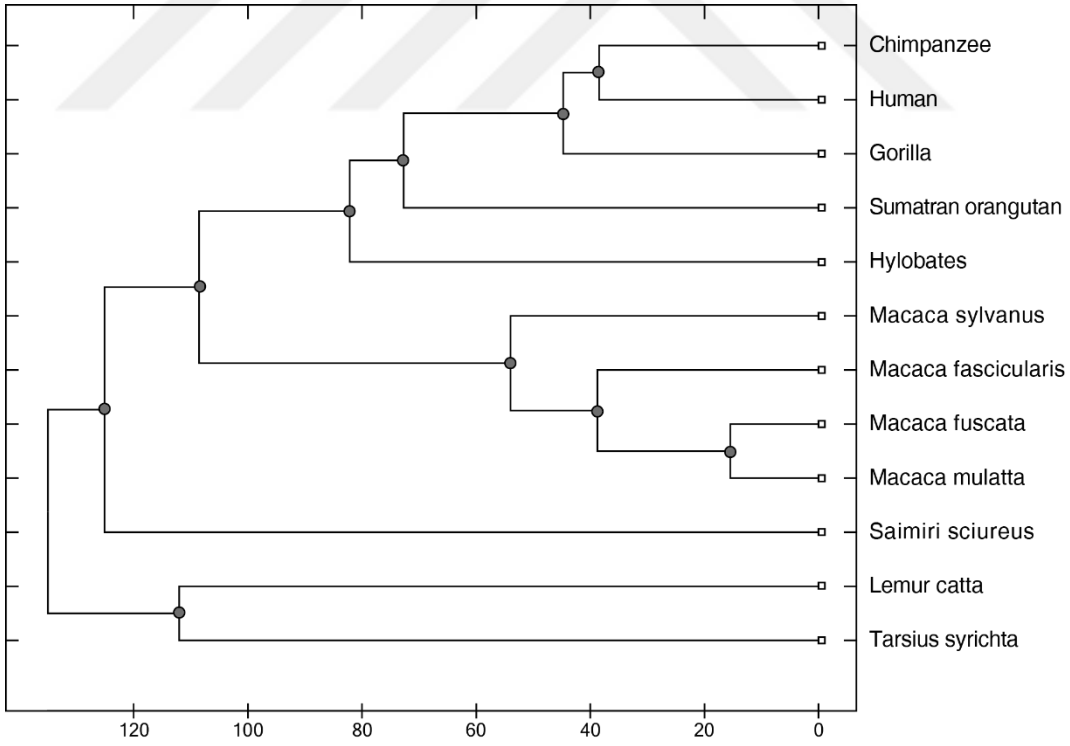


Şekil 5.23. En iyi sonucu üreten 9-gram'a ait doğruluk-hesaplama zamanı grafiği

Bulunan en iyi sonuç ile elde edilen özellik vektörlerinden türler arasındaki benzerlik hesaplandıktan sonra bu benzerliğe dayanarak filogenetik ağaç oluşturulmuş ve Şekil 5.24'te verilmiştir. Şekil 5.25'te de ClustalW hizalaması ve UPGMA yöntemiyle MEGA7 kullanarak elde edilen referans filogenetik ağaç sunulmuştur.



Şekil 5.24. Bölüm 4.3'te önerilen metot ile elde edilen filogenetik ağaç



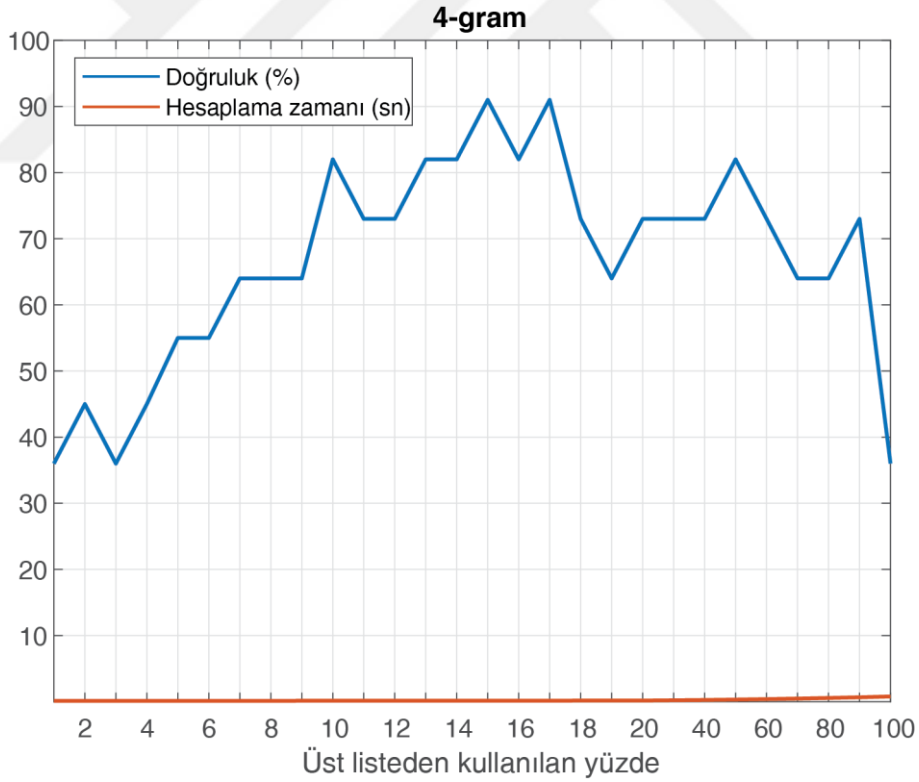
Şekil 5.25. MEGA7 ile ClustalW hizalama ve UPGMA metodu ile elde edilen filogenetik ağaç

Önerilen metot için ikinci test uygulaması Tablo 3.2'de verilen 13 bakteriye ait 16S ribozomal DNA dizileri üzerinde yapılmıştır. Bu veri seti için de metot *3*-gram ile

20-gram aralığında ve üst-*1* ile üst-*100* yüzdelik dilimleri için çalıştırılmıştır. En yüksek doğruluk oranına erişen sonuçlar Tablo 5.2'de verilmiştir. Doğruluk oranı değerleri nRF skorlarının yüzdelik şekilde ifadesidir. En iyi sonucu veren *n*-gram için doğruluk-çalışma zamanı grafiği Şekil 5.26'da verilmiştir. Diğer tüm üst-*k* *n*-gram değerleri için elde edilen grafikler EK-1'de verilmiştir.

Tablo 5.2. 13 bakteriye ait 16s ribozomal DNA veri seti için üst-*k* *n*-gram doğruluk ve hesaplama zaman grafiği

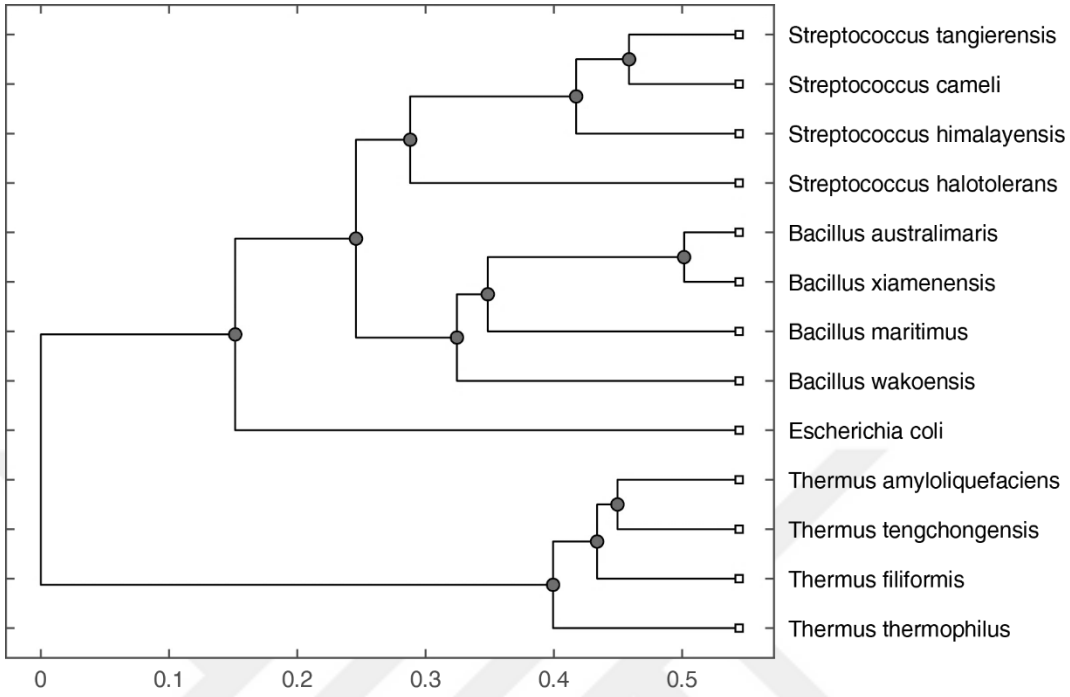
n-gram	Üst-k (yüzde)	Doğruluk	Hes. Zamanı (sn)
4	15	91	0.1461
4	17	91	0.1517
14	1	91	0.2470
16	1	91	0.2494
17	1	91	0.2517



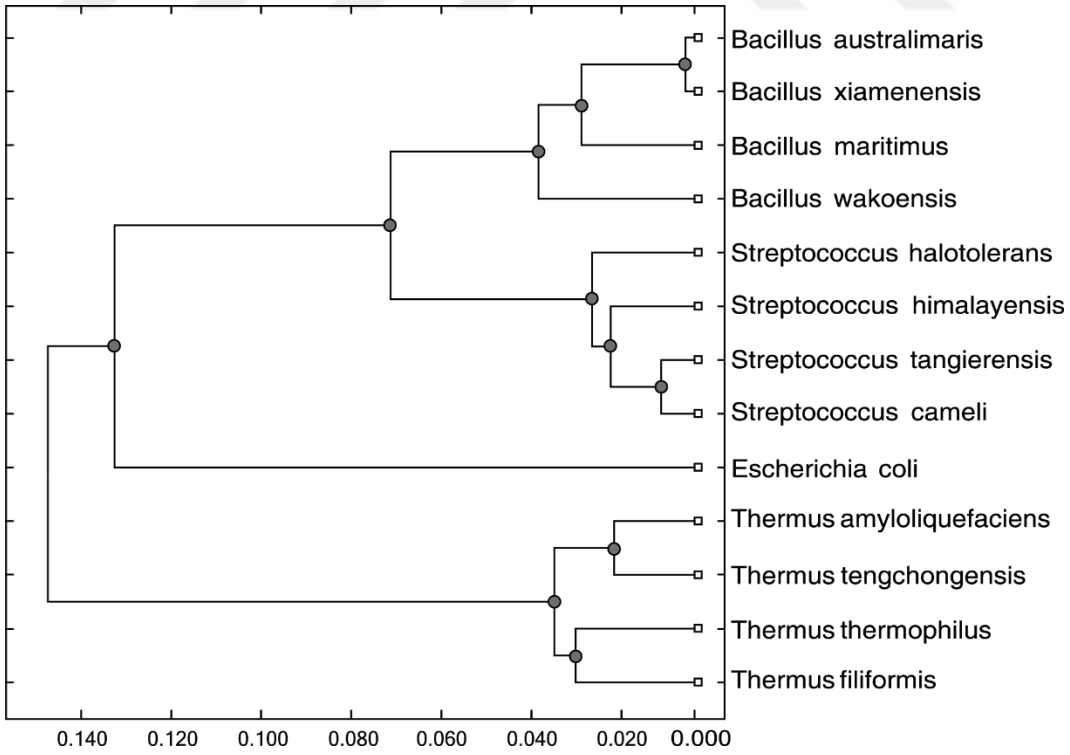
Şekil 5.26. En iyi sonucu üreten 4-gram'a ait doğruluk-hesaplama zamanı grafiği

Bulunan en iyi sonuç ile elde edilen özellik vektörlerinden türler arasındaki benzerlik hesaplandıktan sonra bu benzerliğe dayanarak filogenetik ağaç oluşturulmuş ve

Şekil 5.27'de verilmiştir. Şekil 5.28'de de ClustalW hizalaması ve UPGMA yöntemiyle MEGA7 kullanarak elde edilen referans filogenetik ağaç sunulmuştur.



Şekil 5.27. Bölüm 4.3'te önerilen metot ile elde edilen filogenetik ağaç

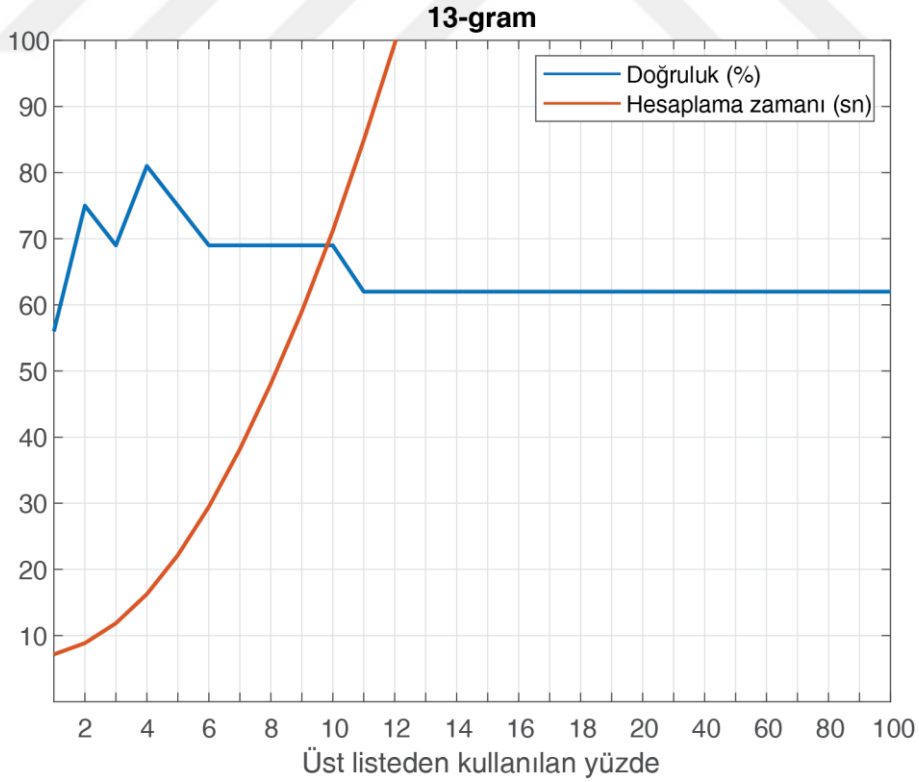


Şekil 5.28. MEGA7 ile ClustalW hizalama ve UPGMA metodu ile elde edilen filogenetik ağaç

Üçüncü test datası olarak, metot uygulaması Tablo 3.3'te verilen 18 eteneli memeliye ait tüm mitokondriyal DNA dizileri üzerinde yapılmıştır. Bu veri seti için de metot, *3*-gram ile *20*-gram aralığında ve üst-*1* ile üst-*100* yüzdelik dilimleri için çalıştırılmıştır. En yüksek doğruluk oranına erişen sonuçlar Tablo 5.3'te verilmiştir. Doğruluk oranı değerleri nRF skorlarının yüzdelik şekilde ifadesidir. En iyi sonucu veren *n*-gram için doğruluk-çalışma zamanı grafiği Şekil 5.29'da verilmiştir. Diğer tüm üst-*k* *n*-gram değerleri için elde edilen grafikler EK-1'de verilmiştir.

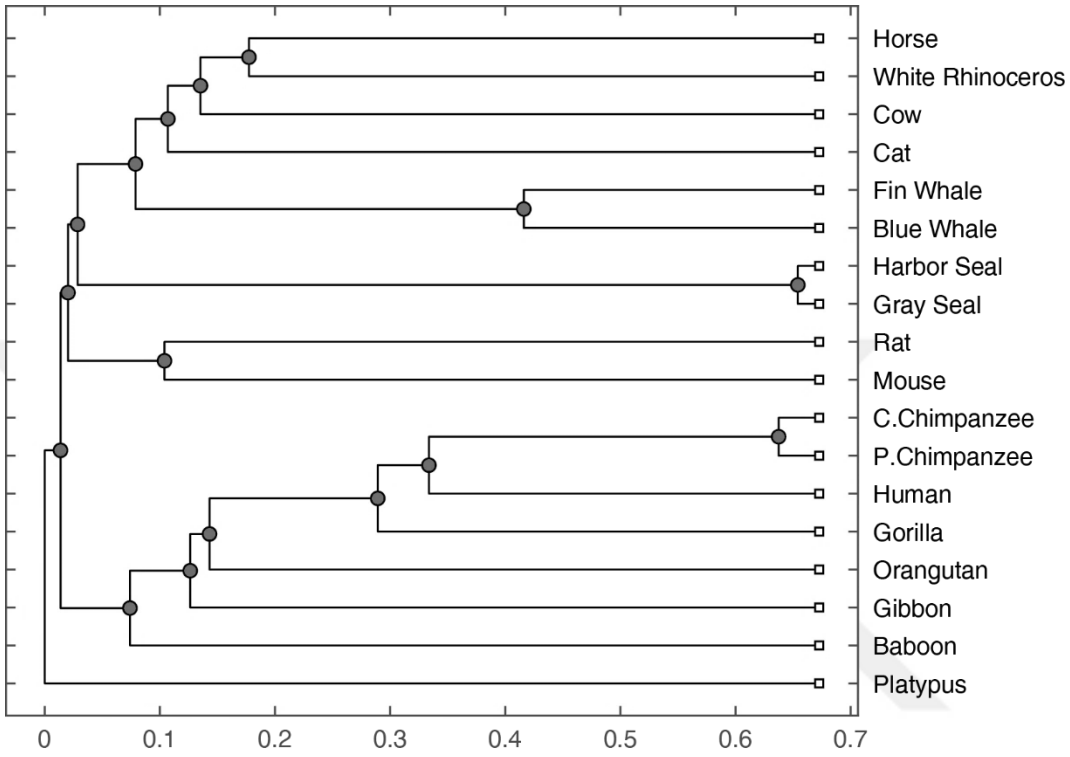
Tablo 5.3. 18 eteneli memeliye ait tüm mitokondriyal DNA veri seti için üst-*k* *n*-gram doğruluk ve hesaplama zamanı grafiği

n-gram	Üst-k (yüzde)	Doğruluk	Hes. Zamanı (sn)
13	4	81	16.2565
12	5	81	21.9551
13	2	75	8.8509
7	8	75	12.8548
7	9	75	15.2619

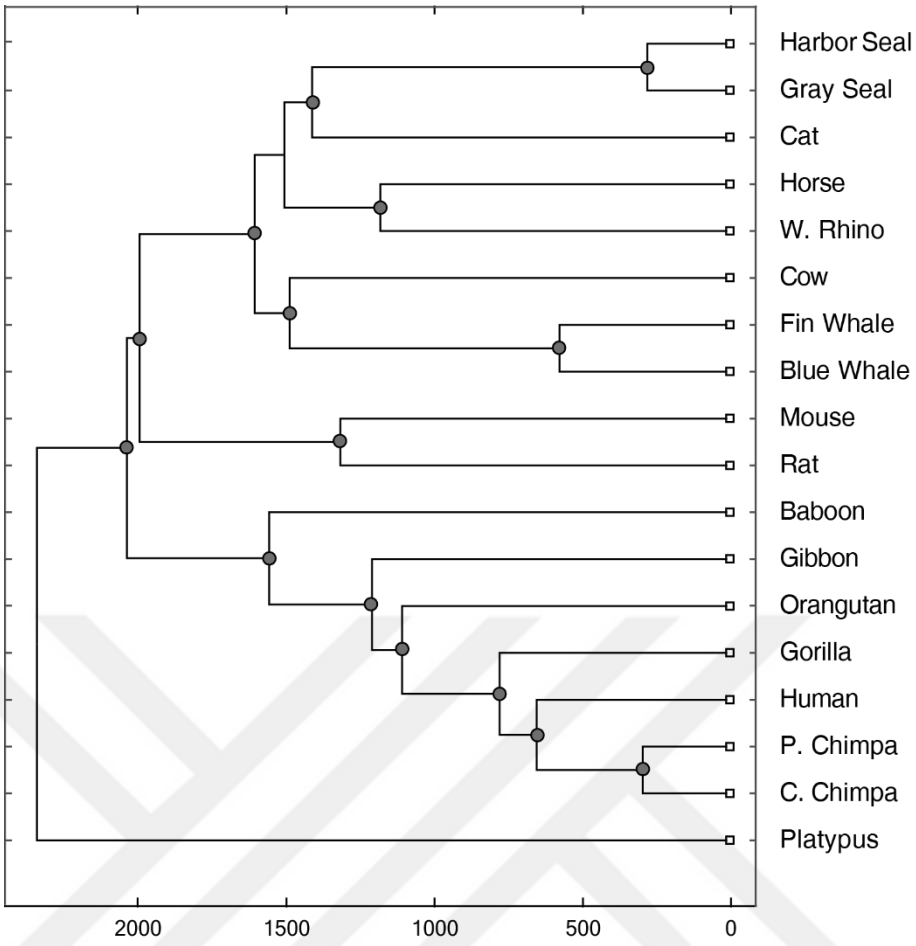


Şekil 5.29. En iyi sonucu üreten 4-gram'a ait doğruluk-hesaplama zamanı grafiği

Bulunan en iyi sonuç ile elde edilen özellik vektörlerinden türler arasındaki benzerlik hesaplandıktan sonra bu benzerliğe dayanarak filogenetik ağaç oluşturulmuş ve Şekil 5.30'da verilmiştir. Şekil 5.31'de de ClustalW hizalaması ve UPGMA yöntemiyle MEGA7 kullanarak elde edilen referans filogenetik ağaç sunulmuştur.



Şekil 5.30. Bölüm 4.3'te önerilen Metot ile Elde Edilen Filogenetik Ağaç



Şekil 5.31. MEGA7 ile ClustalW hizalama ve UPGMA metodu ile elde edilen filogenetik ağaç

Kelime tabanlı metotlar farklı uzunluklardaki DNA dizileri üzerinde doğruluk bakımından başarılı metotlardır. Ancak bu metotların hesaplama maliyeti ciddi problemler oluşturmaktadır. Önerilen bu metotta DNA dizisinden elde edilen tüm gramlar yerine yüksek frekanslı sık tekrar eden gramların kullanılması sayısal karakterizasyon yapılırken DNA dizisinin kullanılmasını sağlamaktadır. Böylece, Şekil 5.29'da da görüldüğü üzere *n*-gramların tamamının kullanımından dolayı ortaya çıkacak hesaplama yükü hafifletilmektedir. DNA dizilerindeki sık tekrar eden desenlerin tüm diziyi temsil etme konusundaki gücü ortaya konmuştur.

6. SONUÇLAR VE ÖNERİLER

6.1. Sonuçlar

Bu tez çalışmasında DNA dizi analizinde hizalamasız karşılaştırma işlemleri için yeni yaklaşımlar sunulmuştur. Bilgisayar bilimindeki benzerlik hesaplaması içerikli çalışmalar temel alınarak tasarlanan metotlarla, makine öğrenimi ve veri madenciliği alanlarının önemli başlıklarından olan özellik çıkarımı uygulamalarının DNA benzerlik analizine sunacağı katkılara dikkat çekilmek istenmiştir. Dört harfli alfabesi olan bir metin parçası olarak ele alınan DNA dizilerinde, dizinin karakteristiğini ortaya koyabilecek tanımlayıcıların istatistiksel metriklerinin kullanımı, metin sınıflandırmada kullanılan bir metodun kullanımı ve görüntü işlemede kullanılan metriklerin kullanımı gibi uygulama ve uyarlamalarla, özellik çıkarımında yaygın kullanımı olan konuların DNA benzerliğine uygulanabilirliği ortaya konmuştur.

DNA dizileri arasında benzerlik analizi yapılırken temel anlayış bu dizilerin aslında aynı olduğu, mutasyonlarla farklılaşan bölgelerin eşleşmedeki etkisini azaltmak üzere benzer kalmış bölgelerin maksimum eşleştirilmesi hedeflenerek kalan benzeşme oranları ile akrabalık ilişkilerini ortaya koymaktır. Hizalamasız analiz metotları da bu amaçlara uygun şekilde farklı özellik çıkarımı yöntemleri denemektedir. Çalışmada önerilen ilk metotta, DNA dizilerinin benzer kalmış bölgelerinin diziler arasındaki benzerliği de tesis edeceği öngörüsüyle, aynı kimyasal özelliklerdeki grupta olan (kendisi veya aynı gruptaki diğer nükleotid) nükleotide geçiş ve değişiklik olmayan duplikasyon bölgelerinin ortalama konumları ile nükleotidlerin dizideki frekanslarının özellik vektörüne eklenmesiyle 7 uzunluklu bir vektörün kullanımını içermektedir. Hesaplaması oldukça kolay olan bu metriklerle bir DNA dizisi, uzunluğundan bağımsız olarak hızlı ve etkili bir şekilde temsil edilebilmiştir.

Önerilen ikinci metotta görüntü analizinde kullanılan doku analizi metriklerinin DNA dizilerini temsil edebileceği düşüncesiyle, görüntünün temel taşı olan pikselleri, DNA dizisinin temel taşı olan nükleotidlerin ikili grupları ile modelleyip, atanan değerlerle elde edilen gri-seviye yoğunluk histogramından metrikleri hesaplanmıştır. Bu metotta da yine 4-boyutlu kısa bir özellik vektörü ile DNA dizisi temsil edilmiş ve başarılı sonuçlar elde edilmiştir.

Son olarak metin madenciliği ve doğal dil işleme uygulamalarında kullanılan *n*-gram metodunun üst-*k* *n*-gram versiyonu ile “DNA dizisindeki sık tekrar eden parçaların,

birbirine çok benzeyen DNA dizilerinde de sık tekrar etmesi gerekir" öngörüsü ispatlanmaya çalışılmıştır. *N*-gramların tamamının alınması hesapsal açıdan maliyetli olacağından, tekrarlama sıklığı en yüksek olan belirli miktardaki *n*-gramın temsil kabiliyetinin yüksek olacağı düşünülmüş ve uygulanmıştır. Bu metotta da beklenen başarılı sonuçlar elde edilmiştir.

Metotların test edilmesi için literatürde daha önce kullanılmış veri setleri tercih edilmiş ve bu çalışmalardaki sonuçlarla karşılaştırma imkânı sunulmuştur. Veri setlerinin farklı uzunluklarda seçilmesi önerilen metotların uzunluğa bağımlı olmadığını, farklı uzunluklarda da başarı gösterebildiğini ortaya çıkarmıştır. Önerilen metotlarla elde edilen sonuçlar karşılaştırılırken, hesaplamalı biyoloji ve biyoinformatik disiplinindeki uygulamalarda sıklıkla kullanılan ve akademik alt yapı ile donatılmış MEGA yazılımı, referans ağaçların üretilmesi için kullanılmıştır. Böylece bilgisayar bilimi ile müdahil olunan DNA dizi benzerlik analizi çalışmalarında, hizalamasız analiz metotları olarak önerilen metotların sonuçlarının bu disiplinde kabul gören sonuçlarla doğrudan karşılaştırılması imkânı sunulmuştur. Sonuçlar, üretilen ağaçlar üzerinden değerlendirilebilirken, ek olarak verilen grafiklerle ve temel bileşen analizi sonuçları ile görselleştirilmiştir. Böylece sonuçların gerek elde edilen vektörlerin henüz herhangi bir uzaklık hesaplaması yapılmamış halinin 2-B koordinat sistemindeki kümelenme biçimiyle başarılı bir vektörel dönüşüm yaptığını göstermesiyle, gerekse uzaklık hesaplaması sonuçlarına göre elde edilen değerlerin grafik üzerinde referans hesaplamalarla örtüştüğünü göstermesi açısından daha belirgin karşılaştırma ortamı sunulmuştur.

Geleneksel benzerlik analiz metotları olan hizalama tabanlı benzerlik analiz metotlarının temel problemlerinden biri hesaplama maliyetidir. Hizalamasız benzerlik analizi metotlarının ana çıkış noktalarından birisi, bu maliyeti kabul edilebilir boyutlara indirgemektir. Önerilen metotlarda özellik vektörlerinin çıkarılmasında bu husus göz önünde bulundurulmuştur. Önerilen ilk metotta özellik vektörünün oluşturulması için DNA dizileri yalnızca bir kez taranarak çok hızlı bir çözüm sunulmuştur. İkinci metotta da doku dönüşümünde bir kez ve dokudan özelliklerin çıkarımı sürecinde bir kez olmak üzere diziler iki taramayla vektör boyutu teşkil eden metriklere dönüştürülmüştür. Son metotta hesaplama maliyeti yüksek olan n-gram hesaplamasında ise, n-gramların en sık tekrar eden ve toplam miktara göre çok küçük bir kısmı kullanılarak karakterizasyon gerçekleştirilmiştir. Bu şekilde bu metodun da hesaplama maliyeti sorunundan başarıyla çıkması sağlanmıştır.

Algoritma tasarımları yapıldığında hesaplama zamanı performansının test edilmesi daha çok büyük verilerle karşılaşıldığında anlamlı olmaktadır. Küçük verilerde doğrusal veya karesel karmaşıklıkta olan metotlar benzer sonuçları üretirken, zaman grafiğinde makasın açılması büyük verilerle kendini göstermektedir. Bu çerçevede tez çalışmasında önerilen metotların MEGA7 yazılımında kullanılan hizalama tabanlı metotlarla hesaplama zamanı karşılaştırmasının kullanılan veri setlerinden 16,572 bp ortalama uzunluklu 18 eteneli memeliye ait tüm mitokondriyal DNA dizisi veri seti ile yapılması daha anlamlı olacaktır. Tablo 6.1'de önerilen üç metodun ve MEGA7'de sıklıkla kullanılan hizalama tabanlı ClustalW ve Muscle algoritmalarının hesaplama zamanları karşılaştırılmıştır. Bölüm 4.3'te verilen metot Tablo 5.3'te verilen sonuçların en kısa hesaplama süresine sahip n-gram ve yüzdelik dilim parametreleri ile hesaplanmıştır.

Tablo 6.1. Tezde önerilen metotlarla MEGA7 yazılımında bulunan metotların Bölüm 3.6.5'te tanımlanan veri setini analizdeki çalışma zamanları

	Önerilen Metotlar			MEGA7	
	Bölüm 4.1	Bölüm 4.2	Bölüm 4.3	ClustalW	Muscle
Hesaplama Zamanı (sn.)	25.22	27.18	16.26	4528.05	2877.58

Hesaplama zamanları incelendiğinde önerilen metotların tümünün hizalamaya dayalı metotlara göre üstünlüğü görülebilmektedir. Ayrıca önerilen metotlar kendi aralarında incelendiğinde önemli bir husus söz konusu olmuştur. Bölüm 4.3'te önerilen metodun, karmaşıklığı daha düşük olan Bölüm 4.1 ve 4.2'deki metotlardan daha kısa hesaplama zamanına sahip olduğu görülmektedir. Hesaplama maliyeti yüksek olan kelime tabanlı analiz metotlarının yalnızca sık tekrar eden parçaları ile yapılan bir analizin işlem yükünü düşük karmaşıklıklı metotların altına kadar indirebildiği görülmüştür. Bu da metodun ayrıca başarısını gösteren bir unsur olmuştur.

6.2. Öneriler

DNA dizi benzerlik analizi genetik ve biyoinformatikte önemli bir araştırma konusudur. Son yıllarda artan talepler doğrultusunda çok sayıda benzerlik analiz metodu önerilmiştir. Buna karşın biyologların ve tıp bilimcilerin ihtiyaçları tam olarak

karşılanamamaktadır. Bu doğrultuda, sık kullanılan grafik tabanlı analiz metotların dışına çıkılarak yeni yaklaşımlara ağırlık verilmelidir. Önerilen yaklaşımlarda özellik çıkarımı yapılırken dizilerin doğası doğru şekilde yansıtılmalıdır. Giderek ve hızla büyüyen genetik verilerin analizlerinde performansın çok önemli bir başlık olduğu göz önünde bulundurulmalı, çalışılan metotlarda doğruluk ile performansın paralel olarak değerlendirilmesi gerekmektedir. Ayrıca büyük veri tabanlarındaki biyolojik dizilerin kayıpsız sıkıştırılması ile elde edilebilecek, gerek depolama gerekse hesaplama maliyetleri daha küçük boyutlara indirgeyecek çalışmalar da dikkate alınmalıdır.

KAYNAKLAR

Aggarwal, N. ve Agrawal, R., 2012, First and second order statistics features for classification of magnetic resonance brain images.

Alhanahnah, M., Lin, Q. C., Yan, Q. B., Zhang, N. ve Chen, Z. X., 2018, Efficient Signature Generation for Classifying Cross-Architecture IoT Malware, *2018 Ieee Conference on Communications and Network Security (Cns)*.

Alobaidli, S., McQuaid, S., South, C., Prakash, V., Evans, P. ve Nisbet, A., 2014, The role of texture analysis in imaging as an outcome predictor and potential tool in radiotherapy treatment planning, *The British Journal of Radiology*, 87 (1042), 20140369.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. ve Lipman, D. J., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, 25 (17), 3389-3402.

Baichoo, S. ve Ouzounis, C. A., 2017, Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment, *Biosystems*, 156-157, 72-85.

Bao, J. P., Yuan, R. Y. ve Bao, Z., 2014, An improved alignment-free model for dna sequence similarity metric, *Bmc Bioinformatics*, 15.

Bharati, M. H., Liu, J. J. ve MacGregor, J. F., 2004, Image texture analysis: methods and comparisons, *Chemometrics and Intelligent Laboratory Systems*, 72 (1), 57-71.

Blaisdell, B. E., 1986, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc Natl Acad Sci U S A*, 83 (14), 5155-5159.

Chen, W. Y., Liao, B. ve Li, W. W., 2018, Use of image texture analysis to find DNA sequence similarities, *Journal of Theoretical Biology*, 455, 1-6.

Cheng, B. Y., Carbonell, J. G. ve Klein-Seetharaman, J., 2005, Protein classification based on text document classification techniques, *Proteins*, 58 (4), 955-970.

Choudhuri, S., 2014, Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools, Elsevier, p.

Chowdhury, B. ve Garai, G., 2017, A review on multiple sequence alignment from the perspective of genetic algorithm, *Genomics*, 109 (5), 419-431.

Cohen, I., 2019, Optimizing Feature generation, https://towardsdatascience.com/optimizing-feature-generation-dab98a049f2e: [13.07.2020].

Compeau, P. ve Pevzner, P., 2015, Bioinformatics Algorithms: An Active Learning Approach. Vol.1, CA: Active Learning Publishers, p.

Dahal, S., 2015, Effect of different distance measures in result of cluster analysis.

Dai, Q., Liu, X. Q. ve Wang, T. M., 2006, A novel 2D graphical representation of DNA sequences and its application, *Journal of Molecular Graphics & Modelling*, 25 (3), 340-344.

Dai, Q., Yang, Y. C. ve Wang, T. M., 2008, Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison, *Bioinformatics*, 24 (20), 2296-2302.

Darling, A. E., Mau, B. ve Perna, N. T., 2010, progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement, *Plos One*, 5 (6), e11147.

Delibaş, E. ve Arslan, A., 2020, DNA sequence similarity analysis using image texture analysis based on first-order statistics, *Journal of Molecular Graphics and Modelling*, 99, 107603.

Delibaş, E., Arslan, A., Şeker, A. ve Diri, B., 2020, A Novel Alignment-free DNA Sequence Similarity Analysis Approach Based on Top-k n-gram Match-up, *Journal of Molecular Graphics and Modelling*, 100, 107693.

Dong, G. ve Liu, H., 2018, Feature engineering for machine learning and data analytics, CRC Press, p.

Durbin, R., Eddy, S. R., Krogh, A. ve Mitchison, G., 1998, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, *Cambridge*, Cambridge University Press, p.

Earl, D., Nguyen, N., Hickey, G., Harris, R. S., Fitzgerald, S., Beal, K., Seledtsov, I., Molodtsov, V., Raney, B. J., Clawson, H., Kim, J., Kemena, C., Chang, J.-M., Erb, I., Poliakov, A., Hou, M., Herrero, J., Kent, W. J., Solovyev, V., Darling, A. E., Ma, J., Notredame, C., Brudno, M., Dubchak, I., Haussler, D. ve Paten, B., 2014, Alignathon: a competitive assessment of whole-genome alignment methods, *Genome research*, 24 (12), 2077-2089.

Eddy, S. R., 2004, What is dynamic programming?, *Nature Biotechnology*, 22 (7), 909-910.

Edgar, R. C., 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, 32 (5), 1792-1797.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. ve Punta, M., 2013, Pfam: the protein families database, *Nucleic Acids Research*, 42 (D1), D222-D230.

Fraz, M. M., Barman, S. A., Remagnino, P., Hoppe, A., Basit, A., Uyyanonvara, B., Rudnicka, A. R. ve Owen, C. G., 2012, An approach to localize the retinal blood vessels using bit planes and centerline detection, *Computer Methods and Programs in Biomedicine*, 108 (2), 600-616.

Ganapathiraju, M. K., Mitchell, A. D., Thahir, M., Motwani, K. ve Ananthasubramanian, S., 2012, Suite of Tools for Statistical N-Gram Language Modeling for Pattern Mining in Whole Genome Sequences, *Journal of Bioinformatics and Computational Biology*, 10 (6).

Gomez, W., Pereira, W. C. A. ve Infantosi, A. F. C., 2012, Analysis of Co-Occurrence Texture Statistics as a Function of Gray-Level Quantization for Classifying Breast Ultrasound, *IEEE Transactions on Medical Imaging*, 31 (10), 1889-1899.

Gotoh, O., 1996, Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments, *Journal of Molecular Biology*, 264 (4), 823-838.

Guo, Y. ve Wang, T. M., 2008, A new method to analyze the similarity of the DNA sequences, *Journal of Molecular Structure-Theochem*, 853 (1-3), 62-67.

Hayasaka, K., Gojobori, T. ve Horai, S., 1988, Molecular Phylogeny and Evolution of Primate Mitochondrial-DNA, *Molecular Biology and Evolution*, 5 (6), 626-644.

Hogeweg, P. ve Hesper, B., 1984, The alignment of sets of sequences and the construction of phyletic trees: an integrated method, *Journal of Molecular Evolution*, 20 (2), 175-186.

Hotelling, H., 1933, Analysis of a complex of statistical variables into principal components, *Journal of educational psychology*, 24 (6), 417.

Hou, W. B., Pan, Q. H. ve He, M. F., 2014, A novel representation of DNA sequence based on CMI coding, *Physica a-Statistical Mechanics and Its Applications*, 409, 87-96.

Jafarzadeh, N. ve Iranmanesh, A., 2012, A Novel Graphical and Numerical Representation for Analyzing DNA Sequences Based on Codons, *Match-Communications in Mathematical and in Computer Chemistry*, 68 (2), 611-620.

Jafarzadeh, N. ve Iranmanesh, A., 2013, C-curve: A novel 3D graphical representation of DNA sequence based on codons, *Mathematical Biosciences*, 241 (2), 217-224.

Jin, X., Nie, R. C., Zhou, D. M., Yao, S. W., Chen, Y. Y., Yu, J. F. ve Wang, Q., 2016, A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and Huffman coding, *Physica a-Statistical Mechanics and Its Applications*, 461, 325-338.

Jin, X., Jiang, Q., Chen, Y., Lee, S.-J., Nie, R., Yao, S., Zhou, D. ve He, K., 2017, Similarity/dissimilarity calculation methods of DNA sequences: A survey, *Journal of Molecular Graphics and Modelling*, 76 (Supplement C), 342-355.

Kantorovitz, M. R., Robinson, G. E. ve Sinha, S., 2007, A statistical method for alignment-free comparison of regulatory sequences, *Bioinformatics*, 23 (13), I249-I255.

Kassambara, A., 2017, Practical Guide To Principal Component Methods in R (Multivariate Analysis), STHDA, p.

Katoh, K., Misawa, K., Kuma, K. i. ve Miyata, T., 2002, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, 30 (14), 3059-3066.

Kolekar, P. S., Kale, M. M. ve Kulkarni-Kale, U., 2010, 'Inter-Arrival Time' Inspired Algorithm and its Application in Clustering and Molecular Phylogeny, *International Conference on Modeling, Optimization, and Computing*, 1298, 307-+.

Kumar, R., Mishra, B. K. ve Lahiri, T. e. a., 2016a, PCV: An Alignment Free Method for Finding Homologous Nucleotide Sequences and its Application in Phylogenetic Study, *Interdiscip Sci Comput Life Sci*, 1-11.

Kumar, S., Tamura, K. ve Nei, M., 1994, Mega - Molecular Evolutionary Genetics Analysis Software for Microcomputers, *Computer Applications in the Biosciences*, 10 (2), 189-191.

Kumar, S., Stecher, G. ve Tamura, K., 2016b, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, *Molecular Biology and Evolution*, 33 (7), 1870-1874.

Lakshmi, N., Gavarraju, P., Jeevana, J. ve Karteeka, P., 2016, A literature survey on multiple sequence alignment algorithms, *Int. J. Adv. Res. Comput. Sci*, 6 (3).

Landy, M. S. ve Graham, N., 2004, The Visual Perception of Texture, *The visual neurosciences*, 1106.

Lassmann, T. ve Sonnhammer, E. L., 2005, Kalign–an accurate and fast multiple sequence alignment algorithm, *Bmc Bioinformatics*, 6 (1), 1-9.

Levine, M., 1985, Vision in Man and Machine, McGraw-Hill, p.

Liao, B. ve Wang, T. M., 2005, Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, *Chemical Physics Letters*, 388 (1-3), 195-200.

Liao, B., Zhang, Y., Ding, K. Q. ve Wang, T. M., 2005, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, *Journal of Molecular Structure-Theochem*, 717 (1-3), 199-203.

Lipman, D. J., Altschul, S. F. ve Kececioglu, J. D., 1989, A Tool for Multiple Sequence Alignment, *Proceedings of the National Academy of Sciences of the United States of America*, 86 (12), 4412-4415.

Liu, N. ve Wang, T.-m., 2005, A relative similarity measure for the similarity analysis of DNA sequences, *Chemical Physics Letters*, 408 (4), 307-311.

Liu, X. Q., Dai, Q., Xiu, Z. L. ve Wang, T. M., 2006, PNN-curve: A new 2D graphical representation of DNA sequences and its application, *Journal of Theoretical Biology*, 243 (4), 555-561.

Luo, J., Li, R. ve Zeng, Q. Q., 2008, A novel method for sequence similarity analysis based on the relative frequency of dual nucleotides, *Match-Communications in Mathematical and in Computer Chemistry*, 59 (3), 653-659.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., Simons, M. ve Stanley, H. E., 1994, Linguistic Features of Noncoding DNA-Sequences, *Physical Review Letters*, 73 (23), 3169-3172.

Mapayi, T., Viriri, S. ve Tapamo, J. R., 2015, Adaptive Thresholding Technique for Retinal Vessel Segmentation Based on GLCM-Energy Information, *Computational and Mathematical Methods in Medicine*.

Materka, A. ve Strzelecki, M., 1998, Texture Analysis Methods – A Review, *Technical University of Lodz, Institute of Electronics*.

Mjelva, K., 2018, Rapid and Sensitive Alignment-free DNA Sequence Comparison.

NCBI, 2020, https://www.ncbi.nlm.nih.gov/home/about/: [01.08.2020].

Needleman, S. B. ve Wunsch, C. D., 1970, A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins, *Journal of Molecular Biology*, 48 (3), 443-+.

Ng, P. C. ve Henikoff, S., 2001, Predicting deleterious amino acid substitutions, *Genome research*, 11 (5), 863-874.

Nguyen, T. ve Kwoh, C., 2015, Efficient agglomerative hierarchical clustering for biological sequence analysis, *TENCON 2015 - 2015 IEEE Region 10 Conference*, 1-3.

Notredame, C., Higgins, D. G. ve Heringa, J., 2000, T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton, *Journal of Molecular Biology*, 302 (1), 205-217.

Osmanbeyoglu, H. U. ve Ganapathiraju, M. K., 2011, N-gram analysis of 970 microbial organisms reveals presence of biological language models, *Bmc Bioinformatics*, 12.

Pearson, W. R. ve Lipman, D. J., 1988, Improved Tools for Biological Sequence Comparison, *Proceedings of the National Academy of Sciences of the United States of America*, 85 (8), 2444-2448.

Pham, T. A., 2010, Optimization of Texture Feature Extraction Algorithm, Master of Science, *Delft University of Technology*, i.

Phillips, A., Janies, D. ve Wheeler, W., 2000, Multiple Sequence Alignment in Phylogenetic Analysis, *Molecular Phylogenetics and Evolution*, 16 (3), 317-330.

Polyanovsky, V. O., Roytberg, M. A. ve Tumanyan, V. G., 2011, Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences, *Algorithms for molecular biology : AMB*, 6 (1), 25-25.

Pratt, W., 1991, Digital Image Processing, Wiley, p.

Qi, X. Q., Wen, J. ve Qi, Z. H., 2007, New 3D graphical representation of DNA sequence based on dual nucleotides, *Journal of Theoretical Biology*, 249 (4), 681-690.

Qi, X. Q., Wu, Q., Zhang, Y. S., Fuller, E. ve Zhang, C. Q., 2011, A Novel Model for DNA Sequence Similarity Analysis Based on Graph Theory, *Evolutionary Bioinformatics*, 7, 149-158.

Randic, M., Vracko, M., Lers, N. ve Plavsic, D., 2003, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, *Chemical Physics Letters*, 371 (1-2), 202-207.

Randić, M., Guo, X. ve Basak, S. C., 2001, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *Journal of chemical information and computer sciences*, 41 (3), 619-626.

Reinert, G., Chew, D., Sun, F. ve Waterman, M. S., 2009, Alignment-Free Sequence Comparison (I): Statistics and Power, *Journal of Computational Biology*, 16 (12), 1615-1634.

Rieck, K. ve Laskov, P., 2008, Linear-time computation of similarity measures for sequential data, *Journal of Machine Learning Research*, 9, 23-48.

Robbins, R. J., Benton, D. ve Snoddy, J., 1995, Informatics and the Human-Genome-Project, *Ieee Engineering in Medicine and Biology Magazine*, 14 (6), 694-701.

Robinson, D. F. ve Foulds, L. R., 1981, Comparison of Phylogenetic Trees, *Mathematical Biosciences*, 53 (1-2), 131-147.

Sadovsky, M. G., 2003, The method to compare nucleotide sequences based on the minimum entropy principle, *Bulletin of Mathematical Biology*, 65 (2), 309-322.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. ve Miller, W., 2003, Human-mouse alignments with BLASTZ, *Genome research*, 13 (1), 103-107.

Schwende, I. ve Pham, T. D., 2013, Pattern recognition and probabilistic measures in alignment-free sequence analysis, *Briefings in Bioinformatics*, 15 (3), 354-368.

Sellers, P. H., 1974, On the Theory and Computation of Evolutionary Distances, *SIAM Journal on Applied Mathematics*, 26 (4), 787-793.

Shapiro, L. ve Stockman, G., 2001, Computer vision prentice hall, *Inc., New Jersey*.

Sheneman, L. ve Foster, J. A., 2006, Estimating the destructiveness of crossover on binary tree representations, *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, 1427-1428.

Shi, L. ve Huang, H. L., 2012, DNA Sequences Analysis Based on Classifications of Nucleotide Bases, *Affective Computing and Intelligent Interaction*, 137, 379-384.

Smith, T. F. ve Waterman, M. S., 1981, Identification of Common Molecular Subsequences, *Journal of Molecular Biology*, 147 (1), 195-197.

Srinivasan, G. N. ve Shobha, G., December 2008, Statistical Texture Analysis. Proceedings Of World Academy Of Science, Engineering And Technology. 36**:** 1264-1269.

Taboada, H. A. ve Coit, D. W., 2007, Data Clustering of Solutions for Multiple Objective System Reliability Optimization Problems, *Quality Technology & Quantitative Management*, 4 (2), 191-210.

Thompson, J. D., Higgins, D. G. ve Gibson, T. J., 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, 22 (22), 4673-4680.

Thompson, J. D., Koehl, P., Ripp, R. ve Poch, O., 2005, BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark, *Proteins: Structure, Function, and Bioinformatics*, 61 (1), 127-136.

Tomović, A., Janičić, P. ve Kešelj, V., 2006, n-Gram-based classification and unsupervised hierarchical clustering of genome sequences, *Computer Methods and Programs in Biomedicine*, 81 (2), 137-153.

Tuceryan, M. ve Jain, A. K., 1998, Texture Analysis, In: The Handbook of Pattern Recognition and Computer Vision Eds, p. 207-248.

Vazquez-Fernandez, E., Dacal-Nieto, A., Martin, F. ve Torres-Guijarro, S., 2010, Entropy of gabor filtering for image quality assessment, *International Conference Image Analysis and Recognition*, 52-61.

Vinga, S. ve Almeida, J., 2003, Alignment-free sequence comparison - a review, *Bioinformatics*, 19 (4), 513-523.

Vinga, S., 2014a, Information theory applications for biological sequence analysis, *Briefings in Bioinformatics*, 15 (3), 376-389.

Vinga, S., 2014b, Editorial: Alignment-free methods in computational biology, *Briefings in Bioinformatics*, 15 (3), 341-342.

Wagner, A., Bicer, V., Tran, T. ve Studer, R., 2014, Holistic and Compact Selectivity Estimation for Hybrid Queries over RDF Graphs, *Semantic Web - Iswc 2014, Pt Ii*, 8797, 97-113.

Wang, D. Z., Wei, L., Li, Y., Reiss, F. ve Vaithyanathan, S., 2011, Selectivity estimation for extraction operators over text data, *2011 IEEE 27th International Conference on Data Engineering*, 685-696.

Wang, J. ve Zhang, Y., 2006, Characterization and. similarity analysis of DNA sequences grounded on a 2-D graphical representation, *Chemical Physics Letters*, 423 (1-3), 50-53.

Wang, J. ve Zheng, X. Q., 2008, WSE, a new sequence distance measure based on word frequencies, *Mathematical Biosciences*, 215 (1), 78-83.

Wang, S., Tian, F., Qiu, Y. ve Liu, X., 2010, Bilateral similarity function: A novel and universal method for similarity analysis of biological sequences, *Journal of Theoretical Biology*, 265 (2), 194-201.

Win, T., Miles, K. A., Janes, S. M., Ganeshan, B., Shastry, M., Endozo, R., Meagher, M., Shortman, R. I., Wan, S., Kayani, I., Ell, P. J. ve Groves, A. M., 2013, Tumor Heterogeneity and Permeability as Measured on the CT Component of PET/CT Predict Survival in Patients with Non–Small Cell Lung Cancer, *Clinical Cancer Research*, 19 (13), 3591.

Wu, T. J., Hsieh, Y. C. ve Li, L. A., 2001, Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition, *Biometrics*, 57 (2), 441-448.

Wu, T. J., Huang, Y. H. ve Li, L. A., 2005, Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences, *Bioinformatics*, 21 (22), 4125-4132.

Xie, X. J., Guan, J. H. ve Zhou, S. G., 2015, Similarity evaluation of DNA sequences based on frequent patterns and entropy, *Bmc Genomics*, 16.

Yamada, S., Gotoh, O. ve Yamana, H., 2006, Improvement in accuracy of multiple sequence alignment using novel group-to-group sequence alignment algorithm with piecewise linear gap cost, *Bmc Bioinformatics*, 7 (1), 524.

Yang, A. C. C., Goldberger, A. L. ve Peng, C. K., 2005, Genomic classification using an information-based similarity index: Application to the SARS coronavirus, *Journal of Computational Biology*, 12 (8), 1103-1116.

Yang, X. ve Wang, T., 2013, Linear regression model of short k-word: a similarity distance suitable for biological sequences with various lengths, *Journal of Theoretical Biology*, 337, 61-70.

Yao, Y. H., Nan, X. Y. ve Wang, T. M., 2005, Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, *Chemical Physics Letters*, 411 (1-3), 248-255.

Yao, Y. H., Nan, X. Y. ve Wang, T. M., 2006, A new 2D graphical representation - Classification curve and the analysis of similarity/dissimilarity of DNA sequences, *Journal of Molecular Structure-Theochem*, 764 (1-3), 101-108.

yourgenome.org, 2020, https://www.yourgenome.org/facts/what-types-of-mutation-are-there: [12.08.2020].

Yu, J. F., Sun, X. ve Wang, J. H., 2009, TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, *Journal of Theoretical Biology*, 261 (3), 459-468.

Yu, J. F., Wang, J. H. ve Sun, X., 2010, Analysis of Similarities/Dissimilarities of DNA Sequences Based on a Novel Graphical Representation, *Match-Communications in Mathematical and in Computer Chemistry*, 63 (2), 493-512.

Zhang, Y. S. ve Chen, W., 2007, New invariant of DNA sequences, *Match-Communications in Mathematical and in Computer Chemistry*, 58 (1), 197-208.

Zhang, Y. S., 2008, A simple method to construct the similarity matrices of DNA sequences, *Match-Communications in Mathematical and in Computer Chemistry*, 60 (2), 313-324.

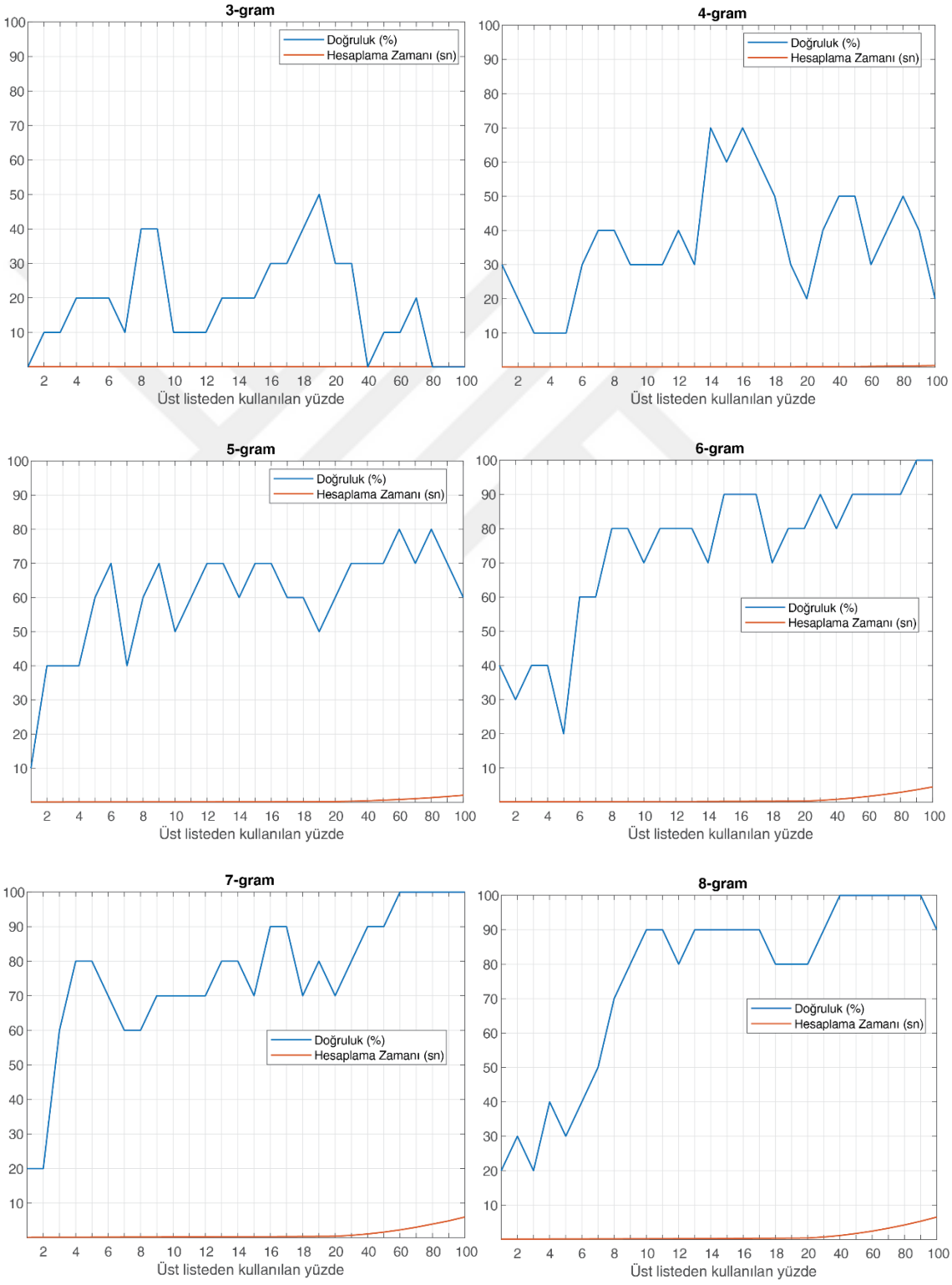
Zhou, J., Zhong, P. ve Zhang, T., 2016, A Novel Method for Alignment-free DNA Sequence Similarity Analysis Based on the Characterization of Complex Networks, *Evolutionary Bioinformatics*, 2016:12, 229-235.

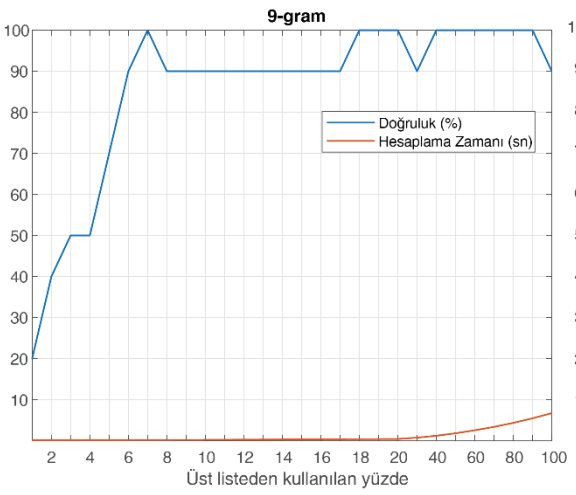
Zielezinski, A., Vinga, S., Almeida, J. ve Karlowski, W. M., 2017, Alignment-free sequence comparison: benefits, applications, and tools, *Genome Biology*, 18 (1), 186.

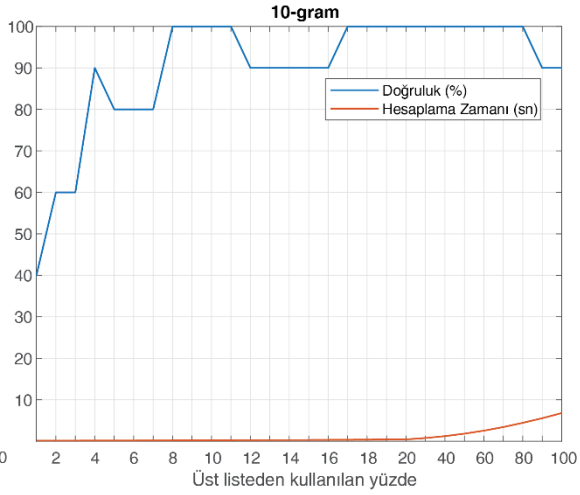
EKLER

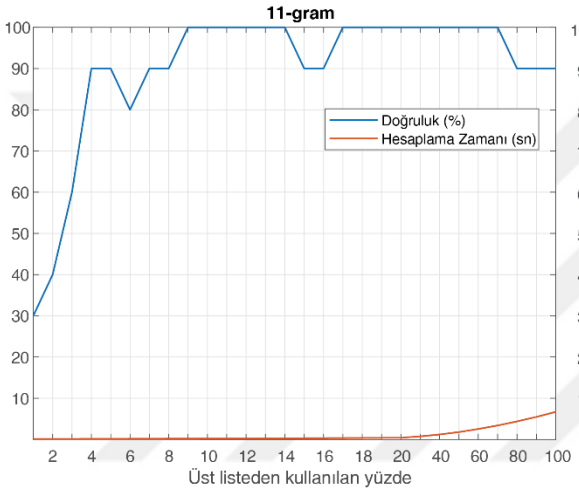
EK-1 Top-k N-gram Eşleşmesi Tabanlı Benzerlik Analizinde Test Edilen n-gramlara Ait Tüm Grafikler

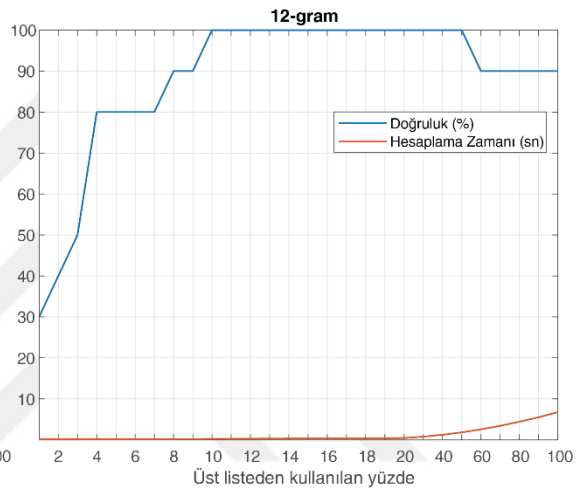
12 türe ait NADH dehidrogenaz alt-birim 4 geni için 3-gram ile 20-gram aralığındaki üst-k n-gram grafikleri.

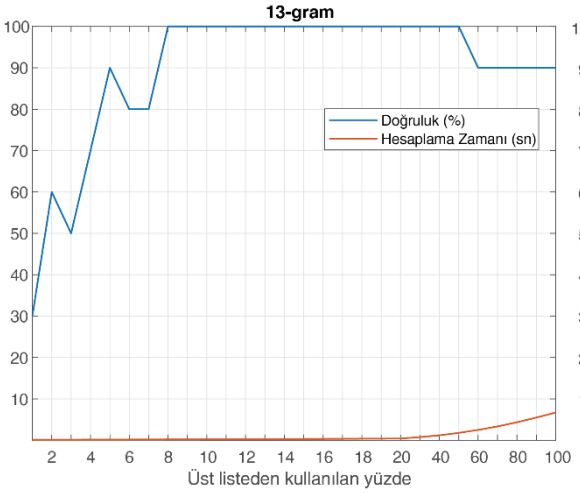


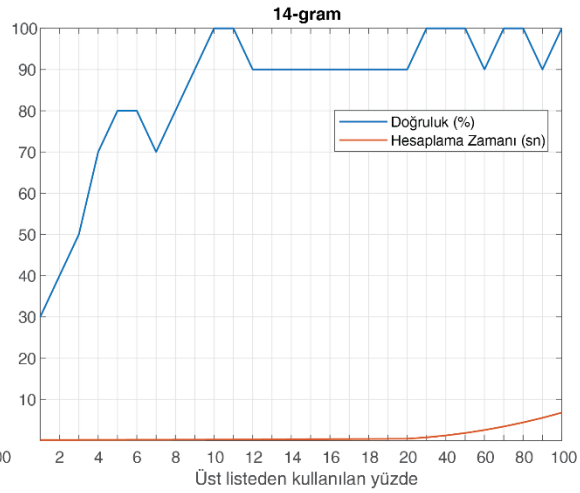

9-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

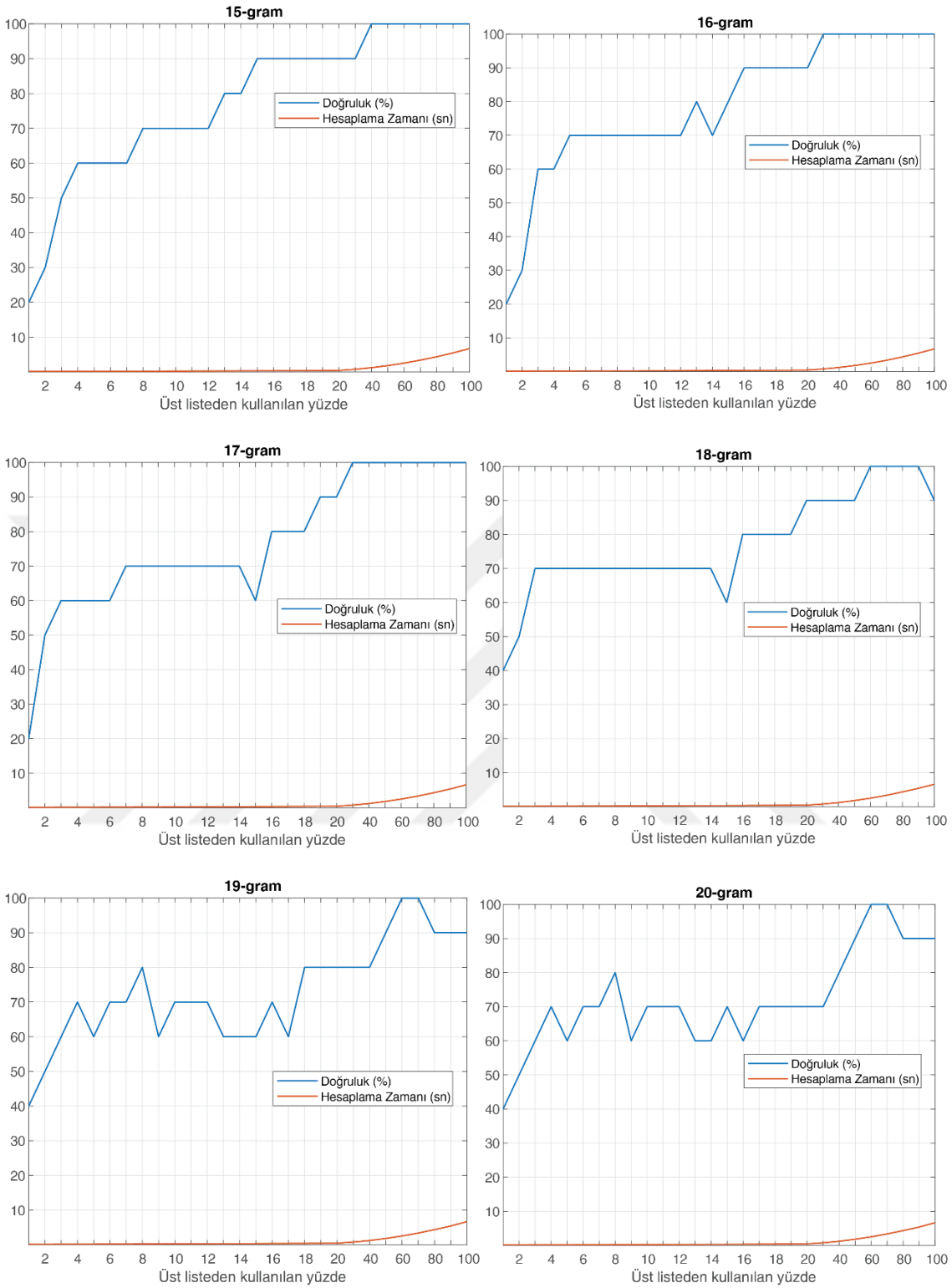

10-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


11-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

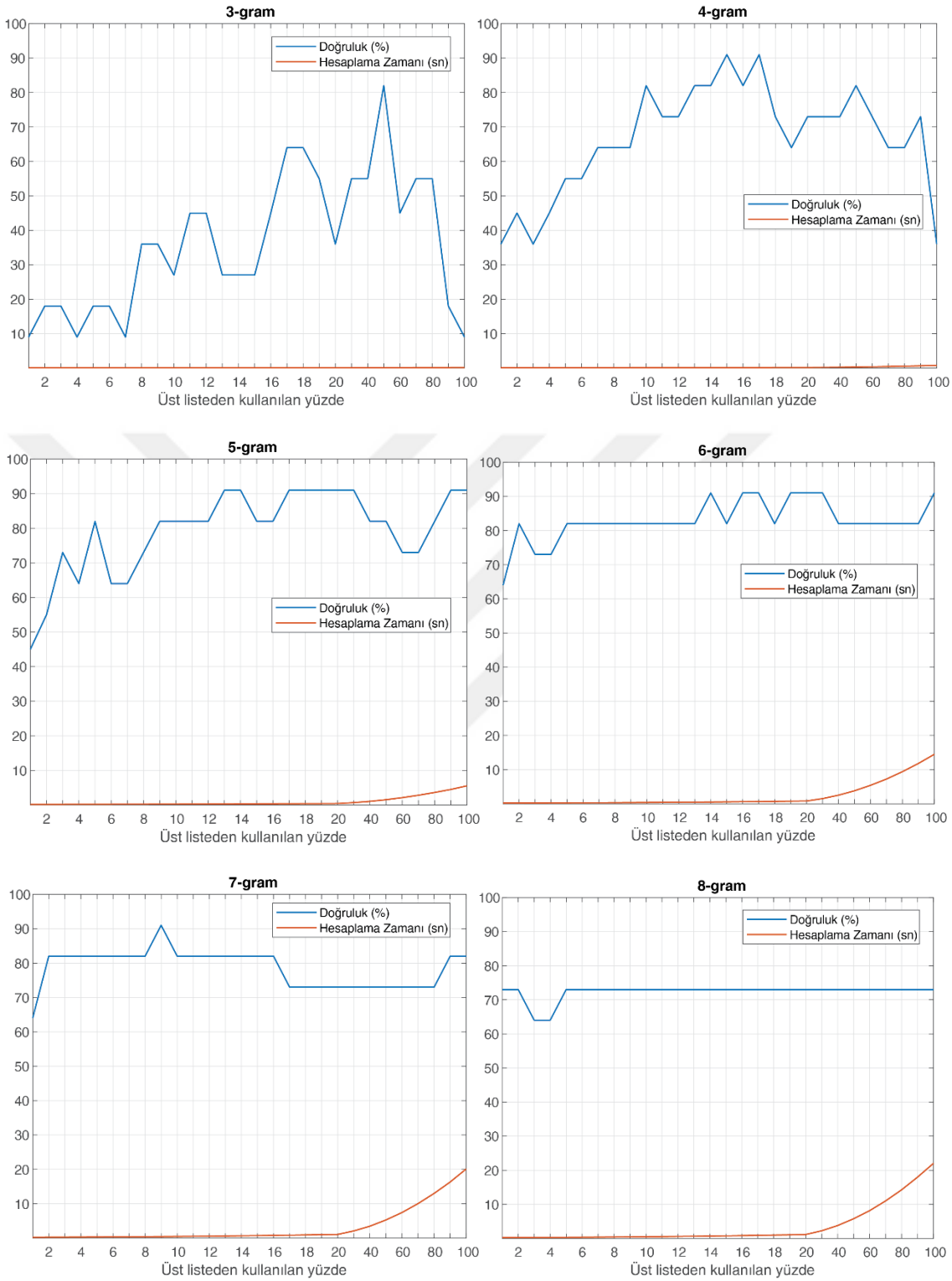

12-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

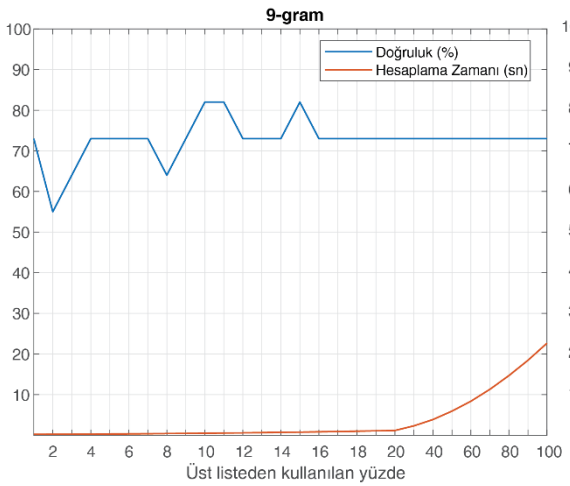

13-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

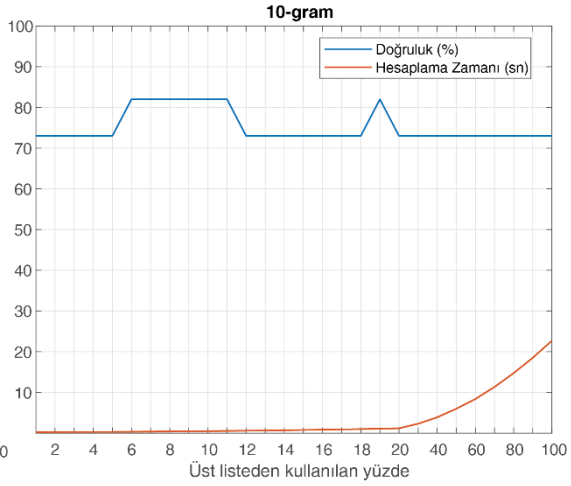

14-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

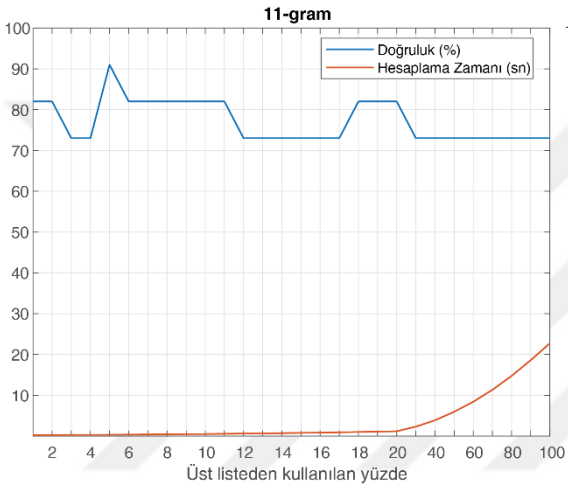

15-gram
16-gram
17-gram
18-gram
19-gram
20-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

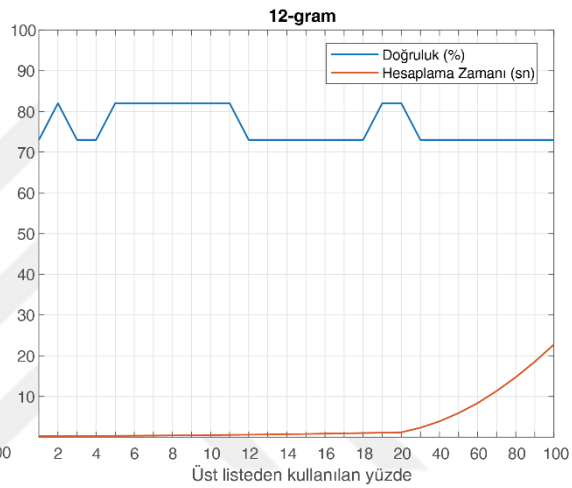
13 bakteriye ait 16S ribozomal DNA dizisi için 3-gram ile 20-gram aralığındaki üst-k n-gram grafikleri.

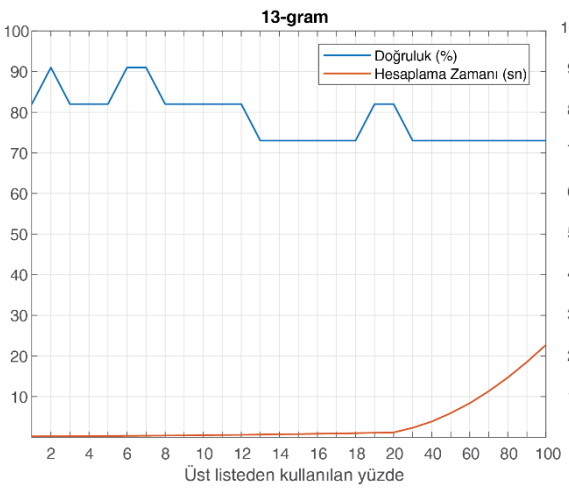


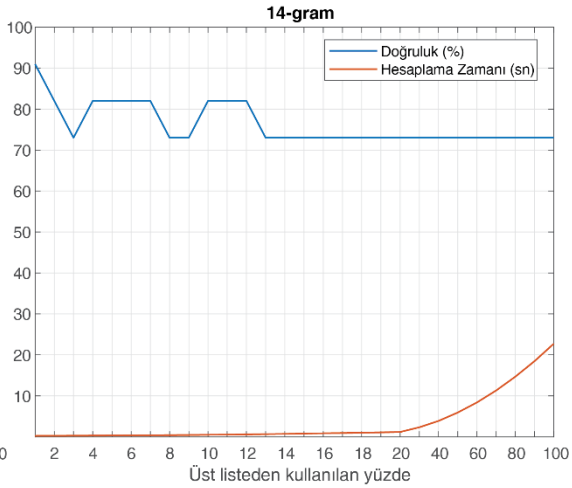

9-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

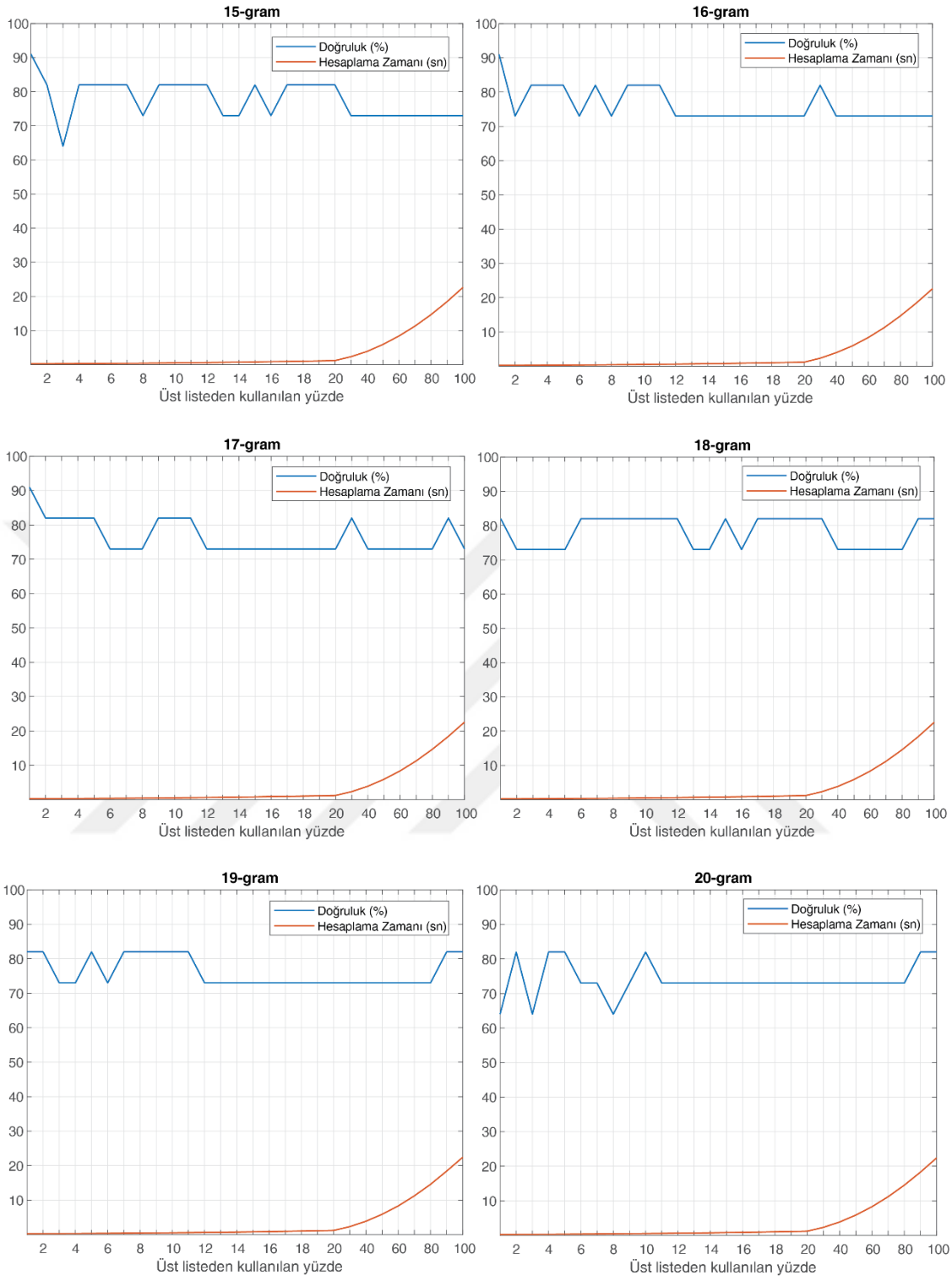

10-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


11-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

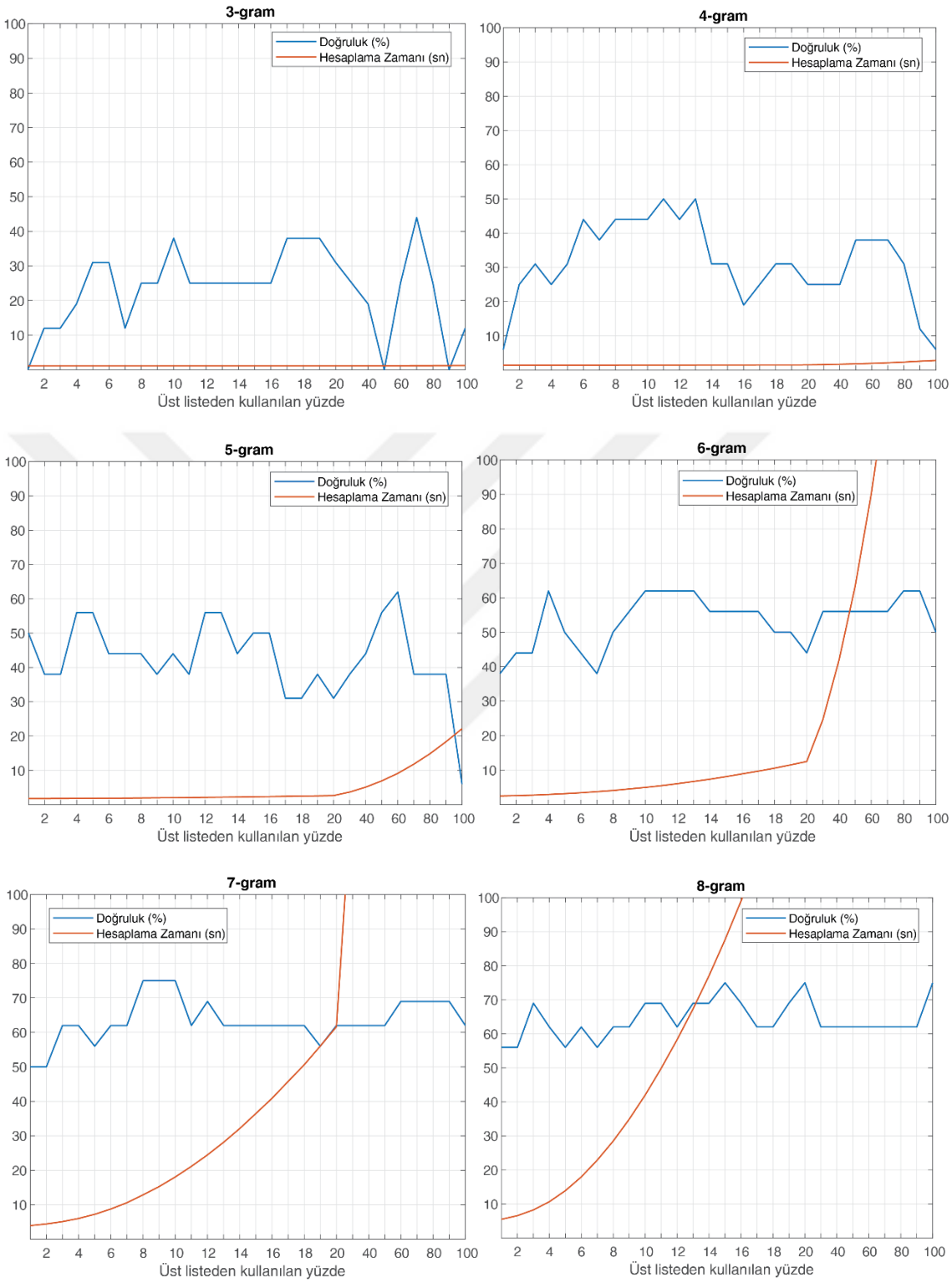

12-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

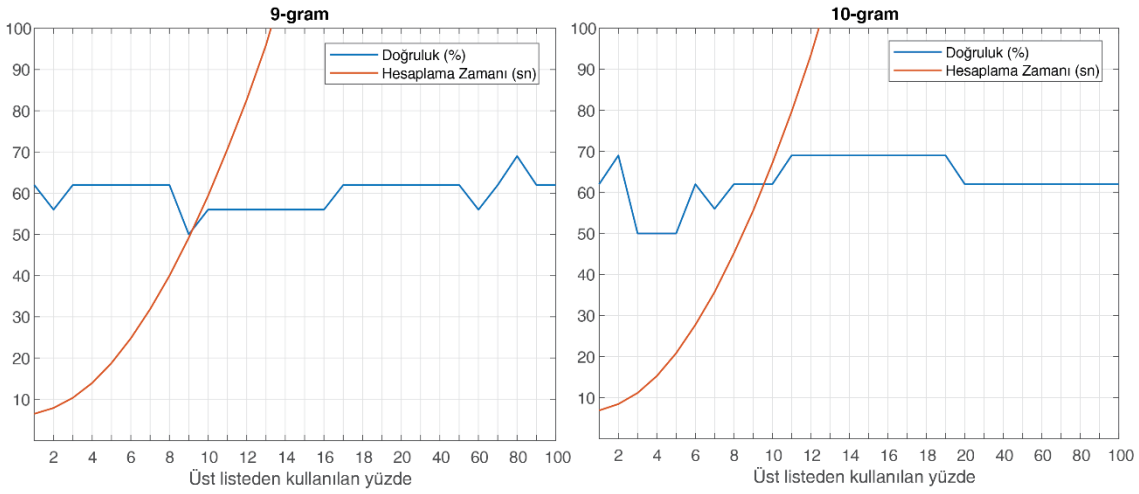

13-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


14-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

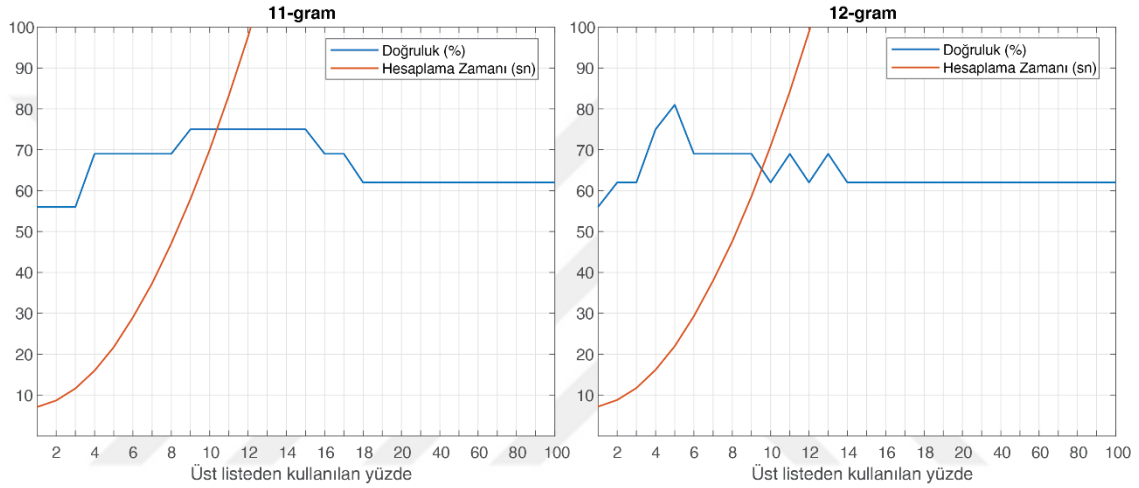

15-gram
16-gram
17-gram
18-gram
19-gram
20-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

18 eteneli memeliye ait tüm mitokondriyal DNA dizisi için 3-gram ile 20-gram aralığındaki üst-*k* *n*-gram grafikleri.

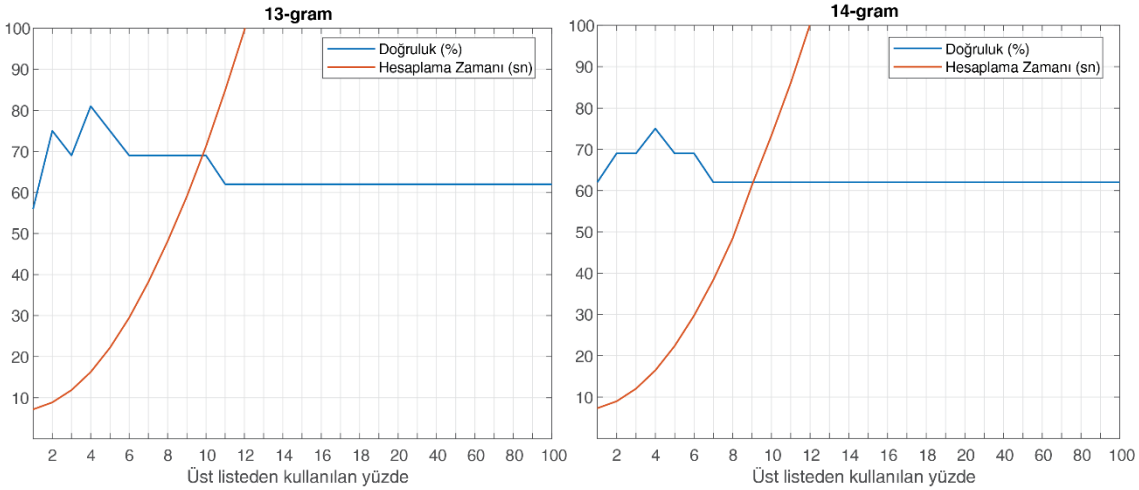



9-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

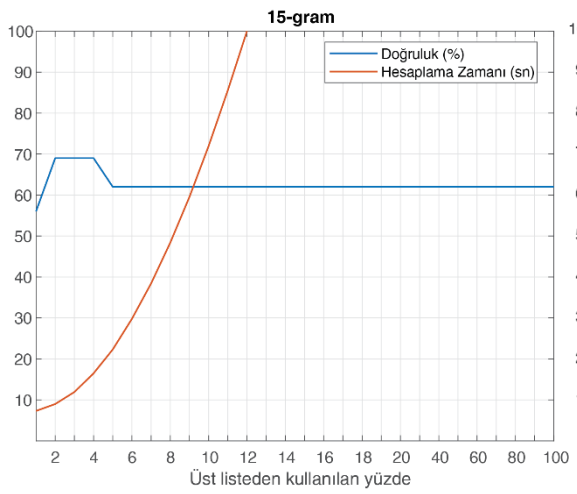
10-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

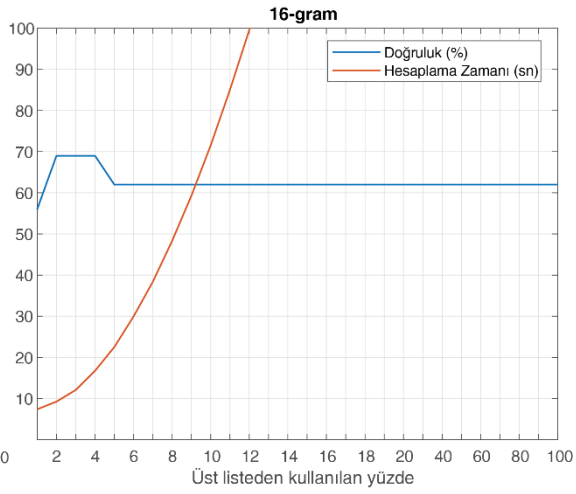

11-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

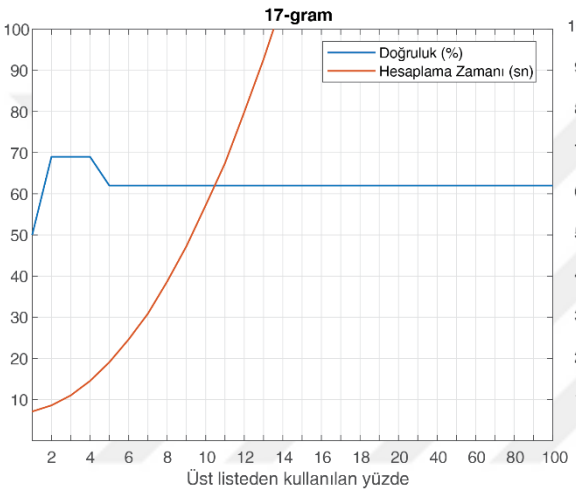
12-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

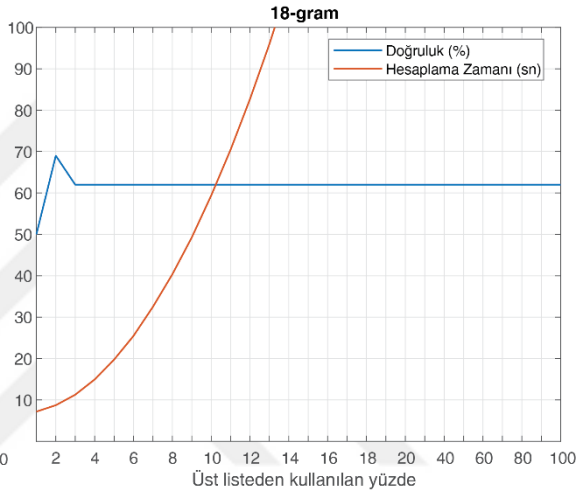

13-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

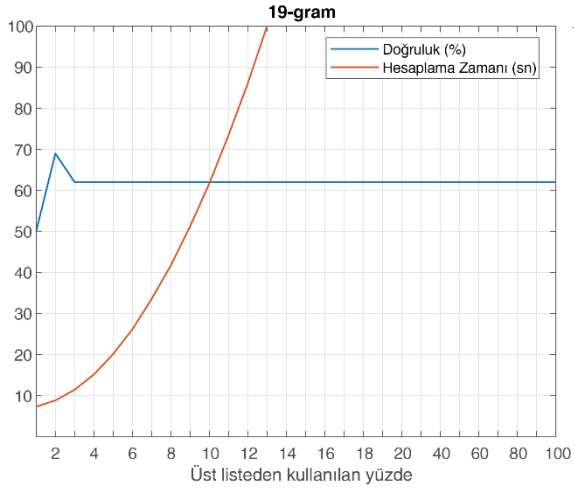
14-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde

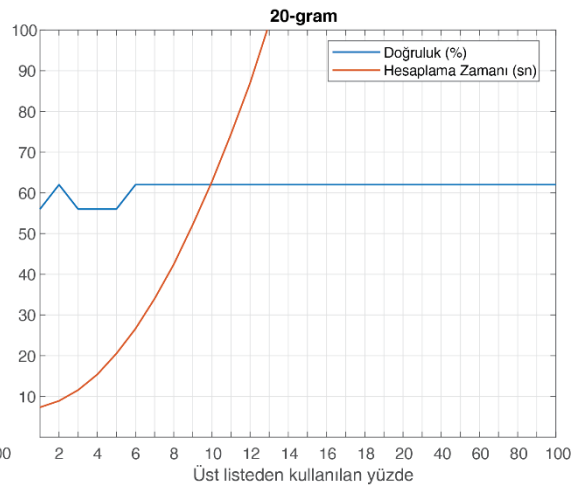

15-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


16-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


17-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


18-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


19-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde


20-gram
Doğruluk (%)
Hesaplama Zamanı (sn)
Üst listeden kullanılan yüzde