

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GEN İFADESİ VERİLERİNDEN DERİN ÖĞRENME
YAKLAŞIMIYLA GENLER ARASI DOLAYLI
ETKİLEŞİMLERİN ORTAYA ÇIKARILMASI

Gülce ÇELEN

YÜKSEK LİSANS TEZİ
Biyomühendislik Anabilim Dalı
Biyomühendislik Programı

Danışman
Doç. Dr. Alper YILMAZ

Şubat, 2022

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**GEN İFADESİ VERİLERİNDEN DERİN ÖĞRENME YAKLAŞIMIYLA
GENLER ARASI DOLAYLI ETKİLEŞİMLERİN ORTAYA ÇIKARILMASI**

Gölce ÇELEN tarafından hazırlanan tez çalışması 25.02.2022 tarihinde aşğıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Biyomühendislik Anabilim Dalı Biyomühendislik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Alper YILMAZ
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Alper YILMAZ, Danışman
Yıldız Teknik Üniversitesi

Doç. Dr. Cem Bülent ÜSTÜNDAĞ, Üye
Yıldız Teknik Üniversitesi

Dr. Öğr. Üyesi Muhammed Erkan KARABEKMEZ, Üye
İstanbul Medeniyet Üniversitesi

Danışmanım Doç. Dr. Alper YILMAZ sorumluluğunda tarafımca hazırlanan Gen İfadesi Verilerinden Derin Öğrenme Yaklaşımıyla Genler Arası Dolaylı Etkileşimlerin Ortaya Çıkarılması başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Gölce ÇELEN

İmza

*"We must believe that we are gifted for something,
and that this thing, at whatever cost, must be attained."*

Marie Skłodowska-Curie



TEŞEKKÜR

Akademik eğitimim ve tez çalışmam süresince bana yol gösteren, yalnızca öğretmeyi amaçlamayıp öğrenmeyi de öğreten ve öğrencisini cesaretlendiren, yardımlarını ve desteklerini esirgemeyen değerli hocam sayın Doç. Dr. Alper YILMAZ'a teşekkür ederim.

Yüksek lisans sürecimde güzel bir bilimsel tartışma ortamı oluşturduğumuza inandığım Alper Yılmaz Research Group üyeleri Selcen ARI-YUKA, Nilay YÖNET, Muhammet ÇELİK, Oğuzhan KALYON, Elif İrem KELEŞ ile lisans dönemimizden beri beyin fırtınası yapmaktan büyük keyif aldığım Büşra Nur DARENDELİ'ye teşekkür ederim.

Hayatımın her alanında beni her zaman destekleyen, kendimi şanslı hissetmemi sağlayan kıymetli aileme teşekkür ederim.

Gülce ÇELEN

İÇİNDEKİLER

KISALTMA LİSTESİ	vi
ŞEKİL LİSTESİ	viii
TABLO LİSTESİ	ix
ÖZET	x
ABSTRACT	xii
1 GİRİŞ	1
1.1 Literatür Özeti	1
1.2 Tezin Amacı	2
1.3 Hipotez	2
2 GEN İFADESİ VE TRANSKRİPTOM ANALİZİ	4
2.1 Mikrodizi Teknolojisi	5
2.2 Yeni Nesil Dizileme ve RNA-Seq	6
2.3 Transkripsiyon Faktörleri	11
2.3.1 Transkripsiyon Faktörlerinin Yapısı ve İşlevi	11
2.3.2 Transkripsiyon Faktörü-Hedef Gen Çiftlerinin Bulunması	12
2.3.3 Gen Düzenleyici Ağlar	13
3 DERİN ÖĞRENME	20
3.1 Yapay Sinir Ağları	22
3.1.1 Evrişimli Sinir Ağları	25
3.2 CNN ile Transkripsiyon Faktörü-Hedef Gen Çiftlerinin Bulunması	27
4 YÖNTEM	29
5 SONUÇ VE ÖNERİLER	34
KAYNAKÇA	40
TEZDEN ÜRETİLMİŞ YAYINLAR	47

KISALTMA LİSTESİ

AtGenExpress	Expression Atlas of Arabidopsis Development
AtRegNet	Arabidopsis thaliana Regulatory Network
cDNA	Tamamlayıcı DNA
ChIP	Kromatin İmmüno-Çöktürme
ChIP-Seq	Kromatin İmmüno-Çöktürme Dizileme
CNN	Evrişimli Sinir Ağı
DNA	Deoksiribonükleik Asit
ESTs	İfade Edilmiş Dizi Etiketleri
FFNN	İleri Beslemeli Sinir Ağı
GEO	Gene Expression Omnibus
GTEx	Genotype Tissue Expression Project
LIME	Local Interpretable Model-Agnostic Explanation
MLP	Çok Katmanlı Algılayıcı
mRNA	Mesajcı RNA
NGS	Yeni Nesil Dizileme
ReLU	Doğrultulmuş Lineer Birim Fonksiyonu
RNA	Ribonükleik Asit
RNA-Seq	RNA dizileme
RNN	Özyinelemeli Sinir Ağı
RPKM	Milyon Başına Kilobaz Başına Transkript Okuması
SAGE	Gen İfadesinin Seri Analizi
SNP	Tek Nükleotid Polimorfizmi
TAFSIID	TATA'ya Bağlanan Proteinle İlgili IID Faktörleri

TATABS	TATA Bağlanma Bölgesi
TCGA	The Cancer Genome Atlas
TRRD	Transcription Regulatory Region Database
TRRUST	Transcriptional Regulatory Relationships Unraveled by Sequence-based Text mining
UASBS	Upstream Aktivasyon Dizisi Bağlanma Bölgesi
UniPROBE	Universal PBM Resource for Oligonucleotide Binding Evaluation



ŞEKİL LİSTESİ

Şekil 2.1	Mikrodizi ile gen ifadesi analizinde tek renk ve çift renk floresan yöntemleri	5
Şekil 2.2	Illumina platformunun temel iş akışı	9
Şekil 2.3	Alternatif uçbirleştirmenin tespiti için RNA-Seq yönteminin kullanılması	10
Şekil 2.4	Transkripsiyon faktörünün yapısı	12
Şekil 2.5	Gen düzenleyici ağın şematik olarak gösterimi	14
Şekil 2.6	Moleküller arası kompleks etkileşimlerin gen düzenleyici ağ olarak modellenmesi	16
Şekil 2.7	Ağ türlerine örnek olarak sırasıyla yönlü, yönsüz ve ağırlıklı ağlar . .	16
Şekil 2.8	Biyolojik veriden gen düzenleyici ağların soyut model olarak çıkarımı	17
Şekil 2.9	Gen düzenleyici ağların çıkarımı için izlenen aşamalar	18
Şekil 3.1	Programlama ve makine öğrenmesi arasındaki farklar	21
Şekil 3.2	Basit bir yapay sinir birimi	22
Şekil 3.3	Biyolojik sinir ve kolon vektör formuyla birlikte yapay sinir diyagramları	23
Şekil 3.4	Derin öğrenmede sıklıkla kullanılan farklı sinir ağı yapıları	24
Şekil 3.5	Evrişimli sinir ağlarının yapısı ve işleyişi	26
Şekil 3.6	Evrişimli sinir ağlarını oluşturan temel yapılar	26
Şekil 3.7	Çalışmada kullanılan CNN modelinin temel şematik gösterimi	28
Şekil 5.1	TCGA verileri kullanılarak eğitilen model	36
Şekil 5.2	GTEX verileri kullanılarak eğitilen model	37
Şekil 5.3	LIME algoritmasının görüntü verisinin işlendiği tahmin modelinde uygulaması	39

TABLO LİSTESİ

Tablo 4.1	Bilinen transkripsiyon faktörü-hedef gen çiftleri veri kümesi	30
Tablo 4.2	GTEx projesinden normal örneklerden gen ifadesi kesiti	30
Tablo 4.3	Model ve kullanılan parametrelerin kodları	31
Tablo 4.4	GTEx verileriyle eğitilen CNN'in yapısı ve model parametreleri . . .	32
Tablo 4.5	TCGA verileriyle eğitilen CNN'in yapısı ve model parametreleri . . .	32
Tablo 5.1	Çalışmada kullanılan verilerin özet tablosu	34
Tablo 5.2	CNN kullanılarak bir gen için elde edilen çift tahminleri	38

Gen İfadesi Verilerinden Derin Öğrenme Yaklaşımıyla Genler Arası Dolaylı Etkileşimlerin Ortaya Çıkarılması

Gülce ÇELEN

Biyomühendislik Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Alper YILMAZ

Gen düzenleyici ağlar, transkripsiyon faktörleri ve bunların hedef genler arasındaki karmaşık düzenleyici ilişkileri açığa çıkarmak için kullanılan çizge tabanlı matematiksel modellerdir. Gen çiftlerinin sayısı düşünüldüğünde her bir transkripsiyon faktörü-hedef çifti için deneysel olarak gen düzenleyici ağların belirlenmesi mümkün değildir. Bu nedenle, farklı türde biyolojik verilerden gen düzenleyici ağların oluşturulması için birçok hesaplamalı yaklaşım geliştirilmiştir. Ancak gen ifadesi verilerinden düzenleyici ağların daha doğru olarak ortaya çıkarılması hesaplamalı biyolojide hala zorlu bir iştir. Derin öğrenme yöntemlerinin bir sınıfı olan evrişimli sinir ağı (CNN), gen ifadesi verisinden transkripsiyon faktörü-hedef çifti tahmininde kullanılabilmektedir. Bu çalışmada benzer bir yaklaşım, normal örneklerden ve tümör örneklerinden insan transkripsiyon faktörü-hedef çiftlerinin tahmin edilmesi amacıyla insan RNA-Seq verisi üzerinde uygulanmıştır. İnsan transkripsiyon faktörü-hedef etkileşimleri TRRUST veri tabanından ve RNA-Seq verisi ise TCGA ve GTEx projelerinden sırasıyla insan tümör örnekleri ile insan normal örneklerini içeren UCSC Toil'den alınmıştır. %80 eğitim verisi, %10 validasyon verisi ve %10 test verisi olarak oluşturulması amacıyla her transkripsiyon faktörü-hedef çifti için gen ifadesi değerleri çıkartılmıştır. Normal örnekler ve tümör örnekleri için iki farklı model oluşturulmuştur. Sonuç olarak normal örneklerde ve tümör örneklerinde transkripsiyon faktörü-hedef çiftlerinin tahmini için iki ayrı evrişimli sinir ağı modeli geliştirilmiştir. Tümör örnekleriyle eğitilen CNN modeli %92 validasyon doğruluğu gösterirken normal örneklerle eğitilen diğer CNN modeli %91 validasyon doğruluğu göstermiştir. Yaklaşımımız var olan ve artmaya devam eden insan gen

düzenleyici ağlarının genişlemeye devam etmesi için bir potansiyele sahiptir. Aynı zamanda, normal örnekler ve tümör örnekleri için ayrı transkripsiyon faktörü-hedef tahminleri, tümör mekanizmalarında bir role sahip olan transkripsiyon faktörü-hedef etkileşimlerini ortaya çıkarabilir.

Anahtar Kelimeler: Gen düzenleyici ağlar, RNA-seq, gen ifadesi, derin öğrenme, evrişimli sinir ağları



Discovery of Indirect Interactions Between Genes by Deep Learning Using Gene Expression Data

Gülce ÇELEN

Department of Bioengineering

Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Alper YILMAZ

Gene regulatory networks are graph based mathematical models to elucidate complex regulatory relationships between transcription factors and their target genes. Considering the number of gene pairs, identifying gene regulatory networks for every transcription factor-target pair is experimentally infeasible. Therefore, many computational approaches have been developed to reconstruct gene regulatory networks from different types of biological data. However, uncovering regulatory interactions from gene expression data more accurately is still a challenge in computational biology. Convolution neural network (CNN) which is a class of deep learning methods can be used to predict transcription factor-target pairs from gene expression data. In this study a similar approach is applied to human RNA-Seq data to predict human transcription factor-target pairs from normal samples and tumor samples. Human transcription factor-target interactions were retrieved from TRRUST database and human RNA-Seq data for tumor and normal samples were retrieved from UCSC Toil which contains data from TCGA and GTEx projects, respectively. For each transcription factor-target pair, expression values were extracted to assemble data as 80% for training, 10% for validation and 10% for test data. Two separate models were generated for normal and tumor samples. As a result, two separate convolutional neural network models were developed to predict transcription factor-target pairs in normal and tumor samples. CNN model which is trained on tumor samples demonstrated 92% validation accuracy, while the other CNN model which is trained on normal samples demonstrated 91% validation accuracy. Our approach has potential to extend the existing and ever growing human gene regulatory networks. Also,

individual transcription factor-target predictions for normal and tumor samples may reveal transcription factor-target interactions which has role in tumor mechanisms.

Keywords: Gene regulatory networks, RNA-seq, gene expression, deep learning, convolutional neural networks



1.1 Literatür Özeti

Biyolojik sistemlerin kompleks yapısı ve elde edilen biyolojik verilerin yüksek boyutlu olması düşünüldüğünde, bu verilerden anlamlı sonuçların elde edilmesi oldukça zorlu bir işlemdir. Biyolojik verilerin analizi için güvenilir yaklaşımlara ihtiyaç duyulmaktadır. Bu amaçla biyolojik verideki örüntülerin keşfi, sınıflandırılması ve tahmini için birçok makine öğrenmesi yöntemi geliştirilmiştir. Makine öğrenmesinin bir alt dalı olan derin öğrenme teknikleri ile büyük veri kümelerindeki, veri için belirli karakteristik özellikler olan özniteliklerin çıkarımı, makine öğrenmesi yöntemlerinden farklı olarak insan müdahalesine gerek duyulmadan algoritma tarafından yapılabilmektedir. Ham veriden kompleks örüntülerin çıkarımını yapabilmesi, derin öğrenmeyi biyoinformatik dahil olmak üzere birçok farklı alanda avantajlı hale getirmektedir. Derin öğrenme algoritmalarından biri olan ve özellikle görüntü analizi uygulamalarında başarılı olan evrişimli sinir ağları, mikroRNA'dan (miRNA) hücre görüntüsüne kadar birçok farklı biyolojik verinin analizinde kullanılmıştır [1].

Mikrodizi ya da RNA dizileme (RNA-Seq) gibi yöntemler aracılığıyla transkriptom analizinden elde edilen gen ifadesi verileri, özellikle hastalıkların teşhisi ve tahmini gibi uygulamalar için farklı yöntemlerle değerlendirilmektedir. Literatürde gen ifadesi verilerinin analizi için evrişimli sinir ağlarının ve diğer sinir ağı modellerinin evrişimli sinir ağı ile birlikte kullanıldığı birçok çalışma bulunmaktadır [2]. Örneğin, meme kanseri türlerinin sınıflandırılması amacıyla yapılan bir çalışmada; RNA-Seq teknolojisi ile elde edilen gen ifadesi verileri kullanılmış ve bu veriler iki boyutlu görüntü benzeri veri tiplerine dönüştürülerek bir evrişimli sinir ağı modeli geliştirilmiştir [3]. 33 tümör türünün sınıflandırılması amacıyla yapılan bir başka çalışmada ise; RNA-Seq ile elde edilen gen ifadesi verileri ile evrişimli sinir ağı modeli kullanılmış ve biyobelirteçlerin belirlenmesinde önem arz eden genlerin bu sınıflandırmadaki etkisine bakılmıştır. Fonksiyonel analiz sonucu, seçilen en önemli genlerin tümöre özgü yollarla ilişkilendirilmesi ile modelin etkililiği vurgulanmıştır

[4]. Mikrodizi ile elde edilen gen ifadesi verilerinin kullanıldığı bir başka çalışmada, akciğer adenokarsinomu için risk sınıflandırılması modeli geliştirilmesi amaçlanmıştır. Bu amaçla ortak gen ifadesi ağı analizi ile sağkalımla ilişkili ağlar belirlenmiş ve bu ağlarda bulunan belirli genler aracılığıyla risk sınıflandırılması modelini oluşturan bir evrişimli sinir ağı elde edilmiştir [5].

Genler arası etkileşimlerin analizi, genler ve gen ürünlerinin işlevi hakkında önemli biyolojik bilgileri açığa çıkartmakta, bu bilgiler ışığında sağlıklı ve hastalıklı fenotiplerin altında yatan mekanizmalar anlaşılabilir [6]. Genler arası etkileşimlerin modellenmesi amacıyla oluşturulan gen düzenleyici ağlar, hesaplamalı biyolojinin önemli bir konusudur. Genetik malzeme aynı olsa dahi, her dokuda farklı hücre tipleri bulunmakta ve bu hücrelerde farklı genler ifade edilmektedir. Benzer şekilde, aynı hücre için farklı koşullarda farklı genlerin ifadesi söz konusudur. Bu nedenle, gen ifadesi verileri kullanılarak gen düzenleyici ağların çıkarımı için farklı modeller geliştirilmiştir [7]. Bu modeller arasından sinir ağı yapıları, verideki örüntüleri tanıyabilmesi, farklı veri yapıları ve herhangi bir fonksiyonel ilişkiyi modelleyebilme kabiliyeti gibi avantajlarla ön plana çıkmaktadır [8]. Evrişimli sinir ağlarının gen ifadesi verilerini işleyebilme kapasitesi göz önüne alındığında; genler arası etkileşimlerin ortaya çıkarılması için evrişimli sinir ağı modeli kullanılarak gen ifadesi verilerinin işlenmesi ve gen düzenleyici ağların çıkarımının yapılması biyolojik bilgi birikimi açısından önemli bir katkı sağlayacaktır.

1.2 Tezin Amacı

Bu çalışmanın amacı, tümör örneklerinin ve normal örneklerin RNA-Seq analizlerinden elde edilen gen ifadesi verilerinin ve bilinen transkripsiyon faktörü-hedef gen çiftleri bilgisinin işlenmesiyle transkripsiyon faktörü-hedef gen çifti tahmini sağlayacak iki farklı evrişimli sinir ağı modeli geliştirilmesidir.

1.3 Hipotez

RNA-Seq yöntemini de içinde barındıran yeni nesil dizileme teknolojisi sayesinde, günümüzde daha hızlı şekilde veri üretilmeye başlanmış ve üretilen verinin hacmi artmıştır. Bu hızlı artış, üretilen büyük hacimli transkriptomik verilerin biyolojik çıkarımlar yapılabilmesi amacıyla kullanılabilmesi için etkili yöntemlerin geliştirilmesini ve bu ham verilerin işlenmesini gerektirmiştir. Derin öğrenme yöntemleri, günümüzde büyük veri kümelerinin işlenmesinde sıklıkla tercih edilen ve biyolojik verilerin analizinde başarılı sonuçlar veren yapay sinir ağı algoritmalarını içermektedir. Yapay sinir ağı yapılarından biri olan evrişimli sinir ağları, uygun

şekilde işlenmiş RNA-Seq verilerinden öznitelikleri ve parametreleri öğrenerek dizi içerisindeki örüntüleri tanıyabilir [9]. Evrişimli sinir ağlarının bu özelliğinden yararlanılarak; gen düzenleyici ağları oluşturan transkripsiyon faktörü-hedef gen çiftlerinin daha doğru ve etkili tahmini için gen ifadesi verilerini içeren tümör örnekleri ve normal örnekler kullanılarak iki farklı evrişimli sinir ağı modeli geliştirilebilir.



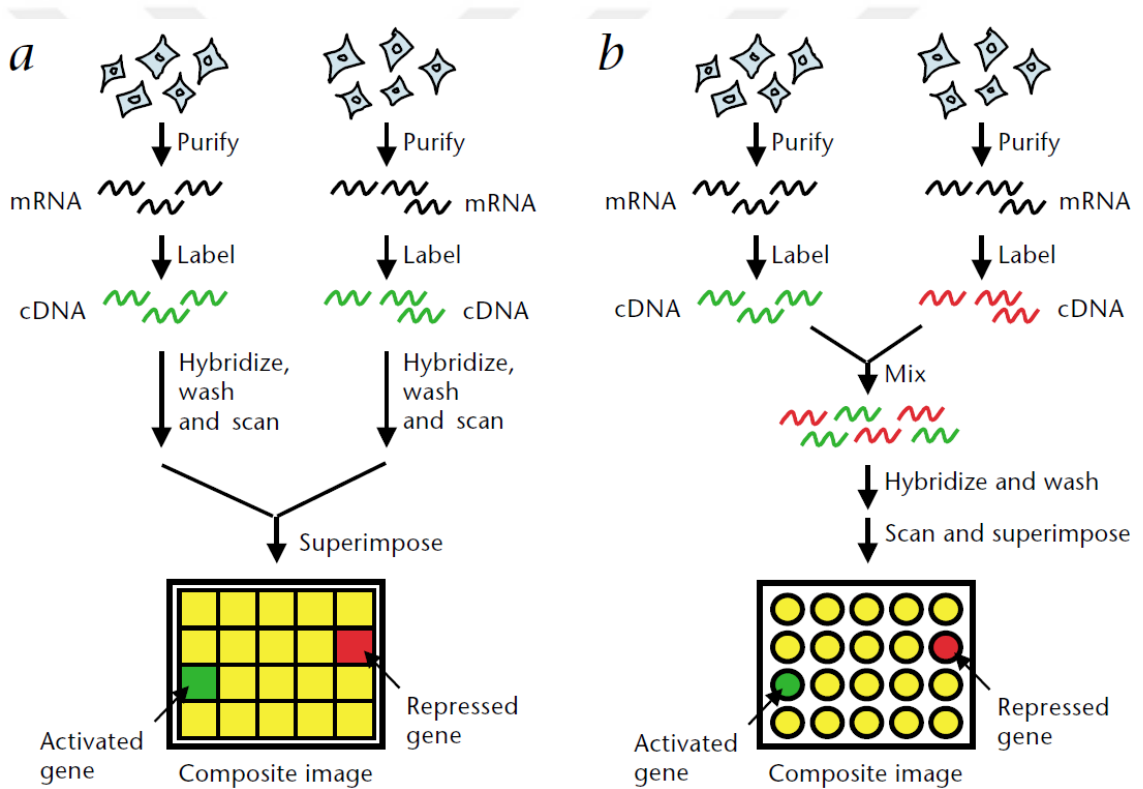
2 GEN İFADESİ VE TRANSKRİPTOM ANALİZİ

Gen ifadesi; DNA'daki bilginin işlevsel bir protein ya da RNA (mRNA veya kodlama yapmayan RNA (İng. noncoding RNA)) oluşturmak üzere işlendiği, denetimli bir süreçtir. Gen ifadesindeki değişimler hücrenin yaşam döngüsü boyunca hücre içi dengesini ve çevreye uyumunu sağlayarak hücrel değişim ve gelişim için önemli bir katkı sağlarken, benzer şekilde patolojik süreçlerde de kritik bir rol oynar [10]. Söz konusu değişimler genlerin farklı koşullar için ifadesinin değişmesiyle olduğundan, gen ifadesi seviyesinin her bir farklı koşul için ölçümünün yapılması gerekmektedir. Bu ölçümler için genlerin ürettiği RNA'lar ya da proteinlerin seviyesine bakılır. Her bir gen için ya da tüm genom düzeyinde gen ifadesi seviyesi ölçümü yapılabilmektedir. Bir hücrenin belirli bir koşuldaki tüm RNA transkriptlerinin toplamı transkriptom olarak adlandırılmaktadır. Transkripsiyon analizinde genellikle odak nokta mRNA'lar olsa da aynı anda transkripsiyonu yapılmış tüm RNA'ların tek bir anlık ölçümü hedeflenmektedir [11].

RNA transkriptlerinin gözlenmesi için birçok yöntem geliştirilmiştir. Northern blot tekniği ile RNA hibridizasyonunu içeren deney protokollerini, tamamlayıcı DNA (İng. complementary DNA) (cDNA) klonlama tekniklerinin gelişmesi sayesinde ifade edilmiş dizi etiketleri (İng. expressed sequence tags) (ESTs) ve gen ifadesinin seri analizi (İng. serial analysis of gene expression) (SAGE) yöntemleri izlemiştir [12]. 1990'lı yılların sonunda geliştirilen mikrodizi teknolojisi transkriptom analizinde kullanılan en popüler teknik olurken, günümüzde yerini bir yeni nesil dizileme (İng. next generation sequencing) (NGS) teknolojilerinden RNA dizilemeye (İng. RNA sequencing) (RNA-Seq) bırakmaktadır [12, 13]. RNA-Seq günümüzde transkripsiyon analizi için altın standart olarak kabul görmektedir [12]. Transkriptom analizi amacıyla kullanılan tekniklerin her biri için iş akışı (1) RNA saflaştırılması, (2) dizileme ve (3) veri analizi basamaklarını içermektedir [11].

2.1 Mikrodizi Teknolojisi

Mikrodizi; bir örnekteki spesifik RNA'larla hibritleşen milyonlarca cDNA parçası ya da prob olarak da adlandırılan oligonükleotid içeren [14], içerisinde referans genomdan elde edilmiş bilinen gen dizilerinden oluşturulmuş her bir diziye özgü probun mikro hücre içerisinde katı bir yüzeyde bulunduğu fiziksel bir cihazdır [15]. Analizi yapılan örnekten elde edilen ve floresanla işaretlenmiş RNA molekülleri, tamamlayıcı olan proba hibritleşir ve gözlenen floresan yoğunluğu ile hedef RNA'nın çokluğu belirlenmiş olur [12]. Mikrodizi verisinde gürültü (İng. noise) oluşmasına sebep olan birçok faktör bulunduğundan, analizden önce ham mikrodizi verisinin normalizasyonu için çeşitli yöntemler kullanılmaktadır [14]. Başta Affymetrix olmak üzere farklı şirketler mikrodizi teknolojisini standardize ederek geliştirmiş ve bu teknolojinin transkriptom analizinde günümüzde hala kullanılıyor olmasına olanak sağlamıştır [12].



Şekil 2.1 Mikrodizi ile gen ifadesi analizinde tek renk ve çift renk floresan yöntemleri. Her iki yöntemde de sarı renkler iki farklı örnekte eşit şekilde ifade edilen genleri göstermektedir. (a) Tek renk floresan analizinde iki ya da daha fazla örnek için iki çip kullanılarak ifade profili oluşturulur. Aktifleştirilmiş (yeşil) ve baskılanmış (kırmızı) genler farklı çiplerden elde edilen görüntülerin üst üste yerleştirilmesiyle belirlenir. (b) Çift renk floresan analizinde iki farklı hücre örneği için tek çip kullanılır. Aktifleştirilmiş ve baskılanmış genler tek mikrodizi üzerindeki farklı kanallardan elde edilen görüntülerin üst üste yerleştirilmesiyle belirlenir [16]

Mikrodizi ile nicel gen ifadesinin eldesi için tek renk floresan (İng. one-colour fluorescent) ya da çift renk floresan (İng. two-colour fluorescent) ile analiz yapılabilir. Tek renkli analizde her bir örnek için ifade profilleri tek floresanla işaretli olarak farklı çipler üzerinde hazırlanır ve sonucunda farklı görüntüler karşılaştırılır. Çift renkli analizde ise iki farklı RNA örnekleri farklı floresanla işaretlenerek tek mikrodizide hibridizasyonu yapılır. Ardından iki farklı kanaldan floresan görüntülerinin oluşturulması için taranır. Şekil 2.1’de yöntemler ve aşamaları gösterilmiştir. Her iki yaklaşım da farklı dokuların karşılaştırılması ya da zamana veya farklı koşullara dayalı örneklerin analizi için kullanılabilir [16].

2.2 Yeni Nesil Dizileme ve RNA-Seq

Çok büyük ölçekte paralel dizileme (İng. massively parallel sequencing) ya da derin dizileme (İng. deep sequencing) terimleriyle de ilişkilendirilen NGS [17], DNA veya RNA dizilerinin temel bilimler, uygulamalı bilimler ve klinik araştırmalarda çeşitli biyolojik olayların çalışılması amacıyla belirlenmesini sağlayan teknolojilerdir [18]. Sanger dizileme gibi önceki teknolojilerle kıyaslandığında NGS ile dizileme işlemleri çok daha kısa sürede, daha ucuza, ve daha hızlı bir şekilde yapılabilmektedir [18]. Gen ifadesi üzerine yapılan çalışmalar, mikrodizi ve SAGE gibi teknikler kullanıldığında yalnızca insan ile fare ve meyve sineği gibi model organizmalarla sınırlıyken, büyük ölçekli paralel dizileme sayesinde tüm genom ya da tüm transkriptom analizleri model olmayan organizmalar için de çalışılabilir hale gelmiştir [19]. NGS teknolojisinde, kullanılan platforma göre parçaların uzunluğu değişmek üzere, milyonlarca kısa DNA parçası aynı anda ve tek seferde okunarak analiz edilebilmektedir. Okuma sonucu elde edilen yüksek çıktılı verilerin işlenebilmesi için çeşitli algoritmalar kullanılmakta ve bu algoritmaların geliştirilmesine ihtiyaç duyulmaktadır. Benzer şekilde, genom dizilemenin yanı sıra kromatin immüno-çöktürme (İng. chromatin immunoprecipitation) (ChIP), tek nükleotid polimorfizmi (İng. single nucleotide polymorphism) (SNP), mutasyon analizi, DNA metilasyonu haritalaması gibi çeşitli amaçlarla dizilemede kullanılmak üzere NGS teknolojisi içerisinde sınıflandırılan birçok platform geliştirilmiştir ve ticari şirketler bu platformların varyasyonlarını geliştirmeye devam etmektedir [20].

Numune malzemesinin cDNA olduğu dizileme sonucunda çıkarılan gen ifadesi profili de RNA-Seq olarak ifade edilebilmektedir [21]; ancak birinci nesil dizileme olarak tanımlanan Sanger dizilemeden sonra geldiği için, literatürde ikinci nesil dizileme olarak da adlandırılabilen [18] NGS teknolojileri platformları içerisinde dizilemenin RNA’ya uygulandığı temel ticari platformlara pyrosequencing esasına dayanan Roche 454, ligasyon tabanlı dizilemeye dayanan SOLiD ve polimeraz tabanlı sentez yoluyla

dizileme esaslı Illumina-Solexa örnek olarak verilebilir [15]. Klinik laboratuvarlarda en sık kullanılan platformlar, her ikisi de sentez yoluyla dizileme yaklaşımını kullanan Illumina ve Ion Torrent platformlarıdır [18, 22].

RNA-Seq'in temel prensibi; cDNA'ların rastgele kesilmesi, milyonlarca kısa parçanın okunarak dizilenmesi, her kısa dizinin genoma hizalanması ve böylece ifade edilen transkriptlerin ve ifade düzeylerinin belirlenerek tamamıyla dizilenmiş transkriptomun elde edilmesi üzerine kuruludur. RNA-Seq milyonlarca okumayı aynı anda yapması, önceden anotasyonu yapılan transkriptlere dayanmaması ve hem nicel hem nitel oluşuyla diğer yöntemlerden ayrılmaktadır. Tüm transkriptom analizinin yanı sıra spesifik genlere odaklanabilmek için farklı stratejilere başvurulabilir. Örneğin; ilgilenilen genin yakalanabilmesi için hedefli RNA-Seq (İng. targeted RNA-Seq), yalnızca kodlanan genlerin seçimi için mRNA-Seq seçim stratejisi olarak kullanılabilir [13]. Tüm NGS işlemleri için izlenen basamaklar; numunenin hazırlanması, DNA'nın parçalara ayrılması, kütüphanenin hazırlanması, dizileme ve veri analizi şeklindedir [18]. RNA-Seq deneylerinde kullanılan cihazlar DNA tabanlı dizileme için üretildiğinden, RNA-Seq analizleri için genellikle cDNA kütüphanelerinin hazırlanması tercih edilse de RNA moleküllerinin doğrudan dizilenmesi de mümkündür [23]. Benzer şekilde RNA-Seq aracılığıyla transkriptom analizi için izlenen iş akışı; örneklerden RNA ekstraksiyonu, cDNA'ların oluşturulması, belli uzunlukta cDNA'ların amplifikasyonu ve dizilenmesi, okuma (İng. reads) olarak da adlandırılan elde edilmiş bu dizilerin bilinen bir transkriptom ya da genomla haritalandırılması ve toplamının eldesi için haritalanan okumaların transkriptlere montajlanmasını (İng. assembly) içerir. Buradaki temel prensip, ilgili transkript miktarıyla, ilgili genin haritalanan cDNA parçalarının sayısının birbiriyle orantılı olması fikridir; ancak dizi dağılımı ve ikincil yapılar transkriptlerin parçalara ayrılmasını ve ters transkripsiyonunu etkileyebilmekte, ökaryotlarda ise alternatif gen uçbirleştirme gibi biyolojik fenomenlerden dolayı okunan parçaların transkriptlere montajlanması çetrefilli olabilmektedir [12]. Dizilenen okumaların hizalanması veya montajlanması, transkriptle çakışan okumanın nicel analizinin yapılması, örneklerin filtrelenmesi ve normalize edilmesi gibi aşamalar bilgisayar üzerinde yapılmaktadır [24]. Transkriptlerin ifade edilme seviyesini tasvir edebilmek için transkripte haritalanan okumaların toplam sayısını belirten milyon başına kilobaz başına transkript okuması (İng. reads per kilobase of transcript per million) (RPKM) gibi bir birim kullanılmalıdır. Birden fazla transkriptle eşlenebilen okumalar, transkriptlerle benzersiz bir şekilde eşlenen okumaların toplam sayısına göre tekrar dağıtılabilmektedir [14].

Günümüzde en çok kullanılan dizileme platformları Illumina tarafından üretilmektedir. Sentez yoluyla dizileme esasına dayanan Illumina teknolojisi,

temelde RNA'dan elde edilen DNA moleküllerinin akış hücresinde katı bir yüzey üzerinde sınırlandırılmış amplifikasyonu üzerine kuruludur. Yüzeye yerleştirilerek amplifikasyonu yapılmış DNA'lar kalıp olarak kullanılarak floresanla işaretlenmiş nükleotidler, DNA polimerazlar yardımıyla, oluşturulan DNA zincirine katılırlar. Her bir nükleotidin katılımı, floresan sinyalinin karakteristik olarak yayılımı sayesinde optik okuyucu ile gözlenir ve bu şekilde sentezlenen dizi tekrar oluşturulabilir. Bu işlem her bir noktada aynı anda ve döngüsel bir şekilde gerçekleştirildiğinden, kütüphanelerin oluşturulması çok büyük ölçekte paralel olarak gerçekleştirilebilmektedir [12].

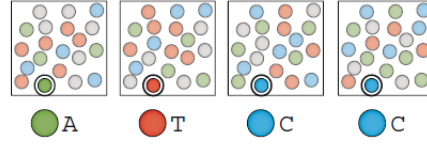
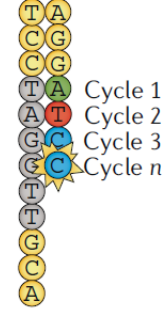
Illumina platformu kullanılarak kurgulanan deneylerde ilk adım, analizi yapılacak olan RNA örneklerinin izole edilmesi ve saflaştırılması, ardından bu RNA'lardan cDNA kütüphanelerinin hazırlanmasıdır [15]. Kütüphane hazırlandıktan sonra, cDNA'lar konsantre edilerek elektroforetik çip yardımıyla kalite değerlendirmesi (parça uzunluğu ve dağılımının ölçülmesi) yapıldıktan sonra [15] her bir cDNA molekülü bir akış hücresi üzerinde kümelenir. Şekil 2.2'de görüldüğü üzere; 3' ucu devre dışı bırakılan ve floresanla işaretlenmiş nükleotidler, cDNA'yı sentez yoluyla dizilemek için kullanılır ve 4 farklı florofor sayesinde sentezlenen DNA ipliğinin dizisi saptanabilir. Her bir döngüde 50-500 baz çifti uzunluğunda okumalar elde edilebilmektedir [24]. RNA-Seq analizinin çıktısı içerisinde okumaları ve kalite değerlerini barındıran, genelde .bam ya da .fastq uzantılı bir fasta dosyasıdır [15].

Deney tasarımında dikkate alınması gereken ana parametreler istenen okuma uzunluğu (İng. read length) ve tek uçlu (İng. single-end) ya da çift uçlu (İng. paired-end) seçenekleridir [15]. Kısa okumalar (İng. short-read) genellikle 100-300 baz çifti arasındayken 500 baz çiftine kadar çıkabilen, parçalarına ayrılmış mRNA okumalarını belirtir. Uzun okumalar (İng. long-read) tam uzunluktaki ya da neredeyse tam uzunluktaki mRNA'ları ifade eden 1000 baz çiftinin üzerine çıkabilen okumaları belirtir [24]. Kısa okumalar yanlış montajlanmış ve parçalarına ayrılmış genomlara neden olabildiğinden ve birçok hastalıkla ilişkili geniş yapısal varyasyonların tanımlanmasını ve karakterizasyonunu zorlaştırdığından, görece kısa okumalar Illumina platformunun en büyük dezavantajlarından [18]. Ancak Illumina'nın kısa okumalar üretebilmesi, degrades olmuş mRNA'ların analizinde avantaj sağlamaktadır. Bununla birlikte kısa okumalarda uzun okumalara göre yüksek çıktılı veri eldesi daha iyidir ve hata oranı daha düşüktür [24]. Tek uçlu okumalar cDNA parçasının sonundan başlanarak dizilenmesine dayalıdır [24]. Çift uçlu okumalar aynı DNA parçası üzerinde bulunan iki farklı bölgenin dizilenmesi yaklaşımı üzerine kuruludur ve görece kısa okumalardan kaynaklı bazı dezavantajların giderilmesini sağlar [25].



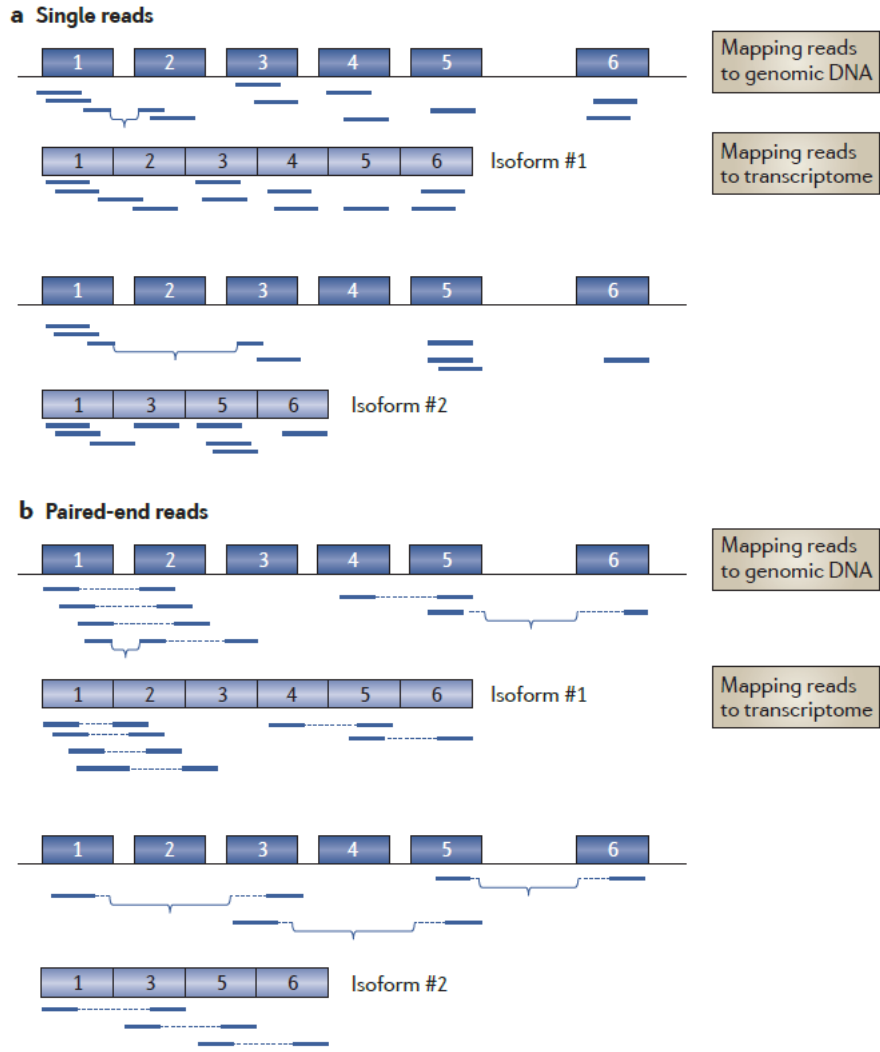
Illumina

Flowcell



Şekil 2.2 Illumina platformunun temel iş akışı [24]

Alternatif uçbirleştirme, hücresel gelişime ve değişime katkıda bulunan genetik bilginin ifadesi bakımından önemli bir süreçtir [25]. Hücresel süreçlerin düzenlenmesinde görev alır [23]. Sağlıklı hücrelerdeki işlevinin yanı sıra, alternatif uçbirleştirmeyi etkileyen mutasyon gibi olaylar sonucu bu sürecin nasıl değiştiği ve hastalık oluşumunda nasıl bir rol aldığı, sağlıklı ve hastalıklı hücrelerin çalışma mekanizmasını aydınlatacağından, biyolojik bilgi birikimi açısından önem arz eden bir araştırma konusudur. RNA-Seq platformları kullanılarak farklı yaklaşımlarla alternatif uçbirleştirmenin tespiti mümkündür [25]. Illumina platformuyla tek uçlu ya da çift uçlu okumalar üretilebilmektedir [18]. Şekil 2.3'te görüldüğü üzere; alternatif uçbirleştirmenin tespiti için hem tekli okuma hem çift uçlu okuma ile dizileme yöntemleri kullanılabilir. Dizileme tek uçlu okumaya göre yapıldığında (a); okunan diziler genomik DNA'ya ya da referans transkriptoma haritalanır. Böylece bir RNA transkriptinin alternatif izoformları bulunabilmektedir. Haritalama, her ekzondaki okuma sayısı temel alınarak yapılmaktadır. Genomik bölgeye haritalanan hiçbir okuma olmaması, transkriptte genomik ekzon olmadığı anlamına gelmektedir. Dizileme çift uçlu okumaya göre yapıldığında (b); çift uçlu okumalar ekzonik uçbirleştirme olayları hakkında daha fazla bilgi verir. İlk okumanın ekzonla eşleşmesi ile ikinci okumanın altakım ekzonla eşleşmesi sayesinde transkriptin yapısı ortaya çıkarılır [25].



Şekil 2.3 Alternatif uçbirleştirmenin tespiti için (a) tek uçlu okumada ve (b) çift uçlu okumada RNA-Seq yönteminin kullanılması [25]

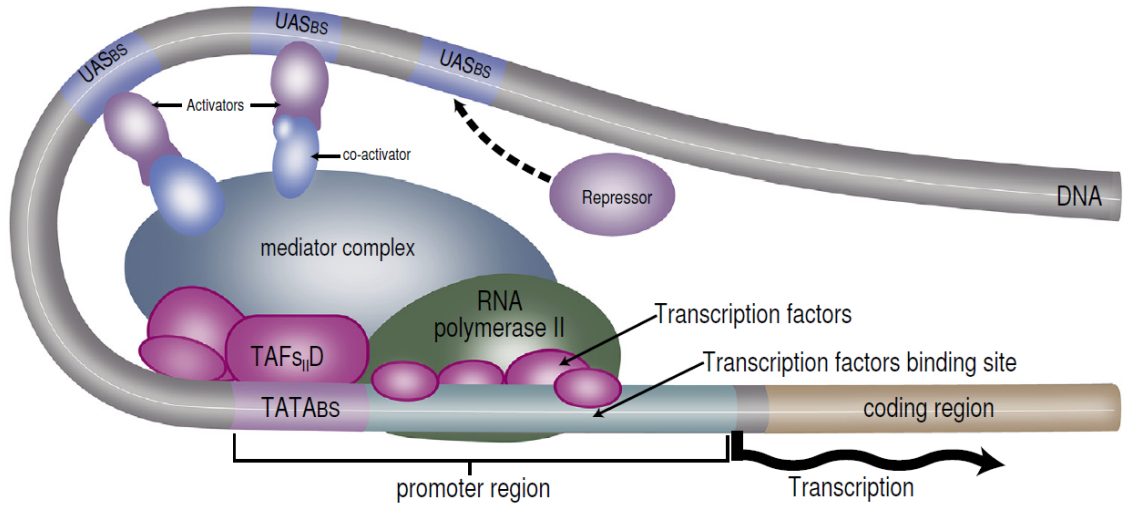
2.3 Transkripsiyon Faktörleri

Transkripsiyon faktörleri; transkripsiyon faktörü bağlanma bölgesi veya gen promotor bölgesi olarak adlandırılan spesifik DNA dizilerine bağlanarak transkripsiyonun kontrolünü sağlayan proteinlerdir [15]. Her ne kadar alternatif uçbirleştirme gibi transkripsiyon sonrası düzenleme mekanizmaları olsa da genlerin transkripsiyonu düzeyinde transkripsiyonunun düzenlenmesi hem dokuya özgü gen ifadesi hem de spesifik bir uyarana karşı gen aktivitesinin düzenlenmesi için bir başlangıç noktasıdır [26]. Biyolojik süreçlerin doğru işleyebilmesi belirli genlerin doğru düzeyde ifadesine bağlı olduğundan, transkripsiyonel düzenlemenin aksaması hücre içi metabolik ve fizyolojik dengenin bozularak çeşitli hastalıkların oluşmasına neden olmaktadır [27]. Örneğin transkripsiyon faktörlerinin onkogenler düzeyinde normal dokulara oranla aşırı ifade edildiği ve bunun kansere yol açtığı bilinmektedir [28]. Transkripsiyon faktörleri veya transkripsiyon faktörlerinin bağlanma bölgelerinde oluşan mutasyonlar da hastalığa sebep olan faktörlerdendir [29].

Transkripsiyon faktörlerinin düzenleyici bölgeleri, protein dizileri ve fizyolojik işlevleri çok hücreli hayvanlarda korunmuştur [30]. Bu durum ortak düzenleyici ağların varlığına ve korunduğuna işaret etse de transkripsiyon faktörlerinin uzun zaman aralıkları incelendiğinde farklılaştığı söylenebilir [29]. Bununla birlikte, bir transkripsiyon faktörü farklı genlerin ifadesinin düzenlenmesinde rol alabilmektedir [31]. Bu nedenle de transkripsiyon faktörlerinin oluşturduğu düzenleyici ağlar dinamiktir ve transkripsiyon faktörlerinin hem bireysel hem de birlikte çalışarak bu dinamizmi devam ettirdiği bir süreçtir [29]. Aynı zamanda, DNA'nın hedef bölgesine bağlanmayan; fakat transkripsiyon faktörleriyle kompleks oluşturarak, transkripsiyon faktörlerinin DNA'ya bağlanmasına eşlik eden düzenleyici proteinler de bulunmaktadır. Böylece, transkripsiyon faktörleri DNA'dan RNA'ya genetik bilgi akışını kontrol edebilmektedir [15].

2.3.1 Transkripsiyon Faktörlerinin Yapısı ve İşlevi

Transkripsiyon faktörleri DNA'ya bağlanabilen bir diziye ve dolayısıyla yapıya sahip olmalarıyla karakterizedir. Transkripsiyon faktörlerinin yapısı ve kromatin ve transkripsiyon kontrolü için oluşturduğu kompleks Şekil 2.4'te gösterilmiştir [32]. Çoğu transkripsiyon faktörü modüler bir yapıya sahiptir. Spesifik DNA bölgesini tanıyan bağlanma bölgesinin (İng. domain) yanı sıra aktivasyon ve baskılayıcı bölgeler de içerebilir. Promotor bölgesi içerisindeki spesifik DNA dizilerine bağlanan bu transkripsiyon faktörü kompleksleri, aktiveleştirici ya da baskılayıcı görev üstlenerek bir genin ifadesini arttırabilir ya da azaltabilir [33].



Şekil 2.4 Transkripsiyon faktörünün yapısı. Transkripsiyon faktörleri oluşturdukları kompleksler ile hedef geni aktive edebilir veya inhibe edebilir. TAFSIIID, TATA'ya bağlanan proteinle ilgili IID faktörlerini; TATABS, TATA bağlanma bölgesini; UASBS upstream aktivasyon dizisi bağlanma bölgesini ifade etmektedir [32]

Spesifik DNA dizilerinin tanınması, DNA bağlanma bölgesindeki nükleotidlerin ve transkripsiyon faktörünün amino asit yan zincirinin etkileşimi sayesinde olur. Bu etkileşim genellikle DNA çift helikal yapısı içerisindeki nükleotid bazlar, yani büyük oluk (İng. major groove) ile gerçekleştirilse de küçük oluk (İng. minor groove) ile de etkileşime giren transkripsiyon faktörleri bulunmaktadır [33].

Transkripsiyon faktörleri ve DNA bağlanma bölgelerinin özgüllüğü sıklıkla motif olarak adlandırılan modellerle ifade edilir. Bu motifler transkripsiyon faktörünün işlevi hakkında bilgi verir ve daha detaylı araştırmalar için ilk basamağı oluşturur [29]. Transkripsiyon faktörleri ve transkripsiyon faktörü-DNA etkileşimleri ile ilgili bilgilerin bir araya getirildiği birçok veri tabanı bulunmaktadır. Bunlara TRANSFAC [34], JASPAR [35], UniPROBE [36] ve TRRD [37] gibi veri tabanları örnek olarak gösterilebilir.

2.3.2 Transkripsiyon Faktörü-Hedef Gen Çiftlerinin Bulunması

DNA bağlanma bölgelerinin deneysel olarak belirlenmesi, bu bölgelerin transkripsiyon faktörü tarafından düzenlenen geni ifade ettiği anlamına gelmemektedir. Transkripsiyon faktörünün bağlandığı spesifik konumun ve o bölgenin işlevsel sonucunun, ilgili genin ifadesinin çıktısı olup olmadığı belirsizdir [38]. Genom düzeyinde yapılan deneyler, transkripsiyon faktörlerinin bağlandığı gen sayısının mümkün olan doğrudan hedef gen sayısını aştığını göstermiştir. Bu nedenle işlevsel ve işlevsel olmayan bağlanmaların ayırt edilmesi gerekmektedir. Bu ayırımın yapılması transkripsiyonel düzenleme çalışmaları için önemli bir sorundur [39].

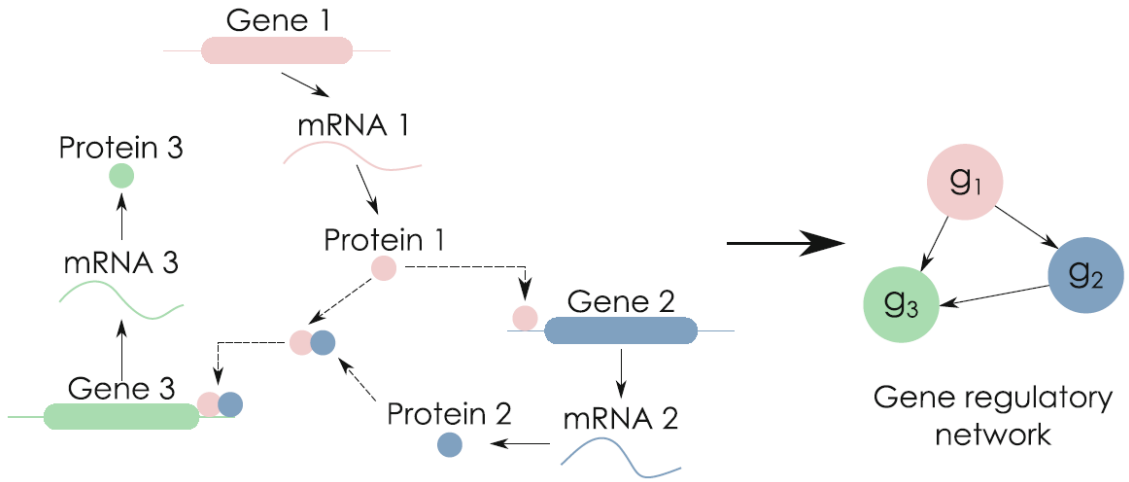
ChIP-seq deneyleri transkripsiyon faktörlerinin genom düzeyinde bağlandığı bölgeleri tek deneyde bulmayı mümkün kılarsa da bu bilgi transkripsiyonel düzenlemenin çıkarımı için yeterli değildir [40]. Örneğin gen transkripsiyonunu hızlandırıcı (İng. enhancer) bölgelerin bulunduğu gen, transkripsiyon faktörünün bağlandığı başka bir gen tarafından düzenleniyor olabilir [41]. Bu nedenle hem hesapsal hem deneysel olarak transkripsiyonel ağların ortaya çıkarılması için birçok farklı yöntem üzerinde çalışılmıştır [27].

Transkripsiyon faktörleri tarafından düzenlenen gen gruplarının ortaya çıkarılması için birçok biyoenformatik analizi yöntemi geliştirilmiştir. Bunların bir kısmı genlerin promotor dizisine ve transkripsiyon bağlanma bölgesinin dizisine dayalı olarak yapılan tahminleri sonuç olarak verir [15]. Bununla birlikte, bir transkripsiyon faktörü tarafından düzenlenen genler gen ifadesi ölçümünde devamlılık sağlamalıdır varsayımıyla, transkripsiyon bağlanma bölgelerini hedef genlerle eşleştirmek için gen ifadesi verisi de kullanılabilir [40].

Transkripsiyon faktörleri ve hedef genler arasındaki düzenleyici etkileşimler gen düzenleyici ağlar olarak adlandırılır ve transkripsiyon faktörleri ile hedef genler arasındaki bu ilişkinin ortaya çıkarılması büyüme, gelişme ve çevresel etmenlere karşı cevap verebilme gibi biyolojik süreçlerin anlaşılması için de elzemdir [42]. Hücrelerin iç ve dış uyaranlara karşı gen ifadelerini düzenlemeleri, birçok transkripsiyon faktörü tarafından kontrol edilen düzenleyici ağlar sayesinde olur ve bu ağlar çok sayıda geni içerebilir [32].

2.3.3 Gen Düzenleyici Ağlar

En genel haliyle bir ağ, düğümler (İng. node) ve kenarlardan (İng. edge) oluşan bir modeldir. Düğümlerin ve kenarların temsil ettiği biyolojik birimler ağın türünü belirlemektedir. Metabolik ağlar, hücre sinyali ağları ve gen düzenleyici ağlar gibi çeşitli biyolojik ağlar bulunmaktadır. Gen düzenleyici ağlar temelde hücresel düzenleyici olaylara katılan genler ve bunların hedef genleri arasındaki etkileşimi tanımlayan ve tahmin eden modellerdir. Şekil 2.5'te görüldüğü üzere; bir gen düzenleyici ağda düğümler genleri, kenarlar ise düzenleyici ilişkileri ifade etmektedir. [43].



Şekil 2.5 Gen düzenleyici ağın şematik olarak gösterimi [44]

Gen düzenleyici ağlar; genel olarak gen ifadesi verisinin analizi sonucu çıkarımı yapılan [45] ve sinyal iletimi ile transkripsiyonel düzenleme gibi hücresel düzenleyici işlevlerde görev alan genlerin arasındaki ilişkiyi açıklayan ağlardır [46]. Hücresel biyolojik süreçlerinin ve bunların hücrenin devamlılığındaki etkinliğinin anlaşılması için, genlerin tek tek ele alınması yerine bütüncül bir yaklaşımla incelenmesiyle gen düzenleyici ağların çıkarımının yapılması biyolojik olarak önemli bilgilerin açığa çıkmasını sağlar [8].

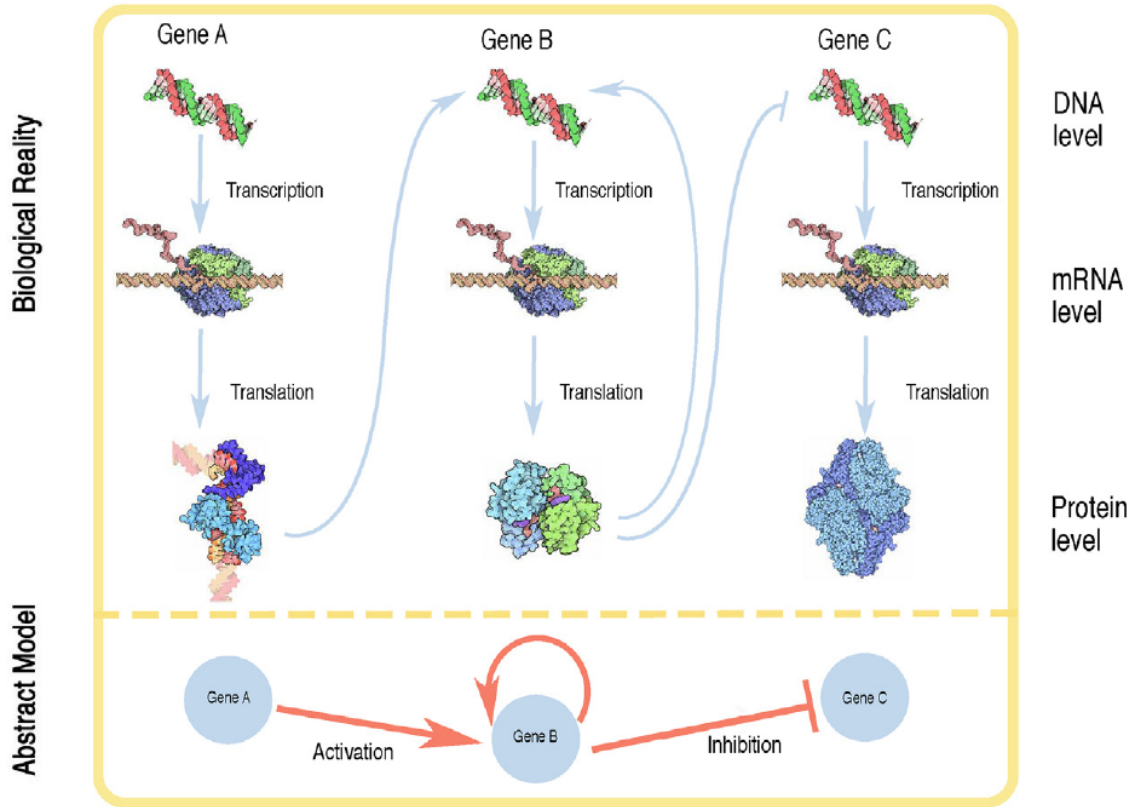
İnsan genomunda yaklaşık 20000 kadar gen olduğu göz önüne alındığında her bir gen çifti için deney düzeneği oluşturularak genler arasındaki ilişkinin deneysel olarak tespit edilmesi hem zor hem de elverişsiz bir işlem olduğundan, gen ifadesi verilerinin nicel analizi yapılmakta ve çeşitli algoritmalar yardımıyla gen düzenleyici ağların çıkarımı yapılmaktadır [47]. Bununla birlikte söz konusu gen düzenleyici ağlar sağladıkları biyolojik bilgi sayesinde yeni deney tasarımlarının yapılmasını sağlayabilir [45]. Gen düzenleyici ağlar, gen-gen etkileşimi gibi moleküler etkileşimler açısından önemli bir model oluşturduğundan [8] diğer moleküler etkileşimler arasında yeni hipotezler oluşturulması için de kullanılmaktadır [48].

Gen düzenleyici ağlar, özellikle kanser gibi bir tek gen yerine yollarla karakterize olan karmaşık hastalıklar için biyobelirteç (İng. biomarker) olabilme potansiyeli taşımaktadır; fakat ağ tabanlı olarak oluşturulan biyobelirteçlerin geliştirilmesi oldukça karmaşık bir işlemdir [45]. Gelişen teknolojiyle birlikte, gen düzenleyici ağların kanserin ilerlemesi ve tedavisinin izlenmesinde biyobelirteç olarak kullanıldığı çalışmalar yapılmıştır [48, 49]. Belirli bir hastalık veya duruma çözüm oluşturmak için yapılan ilaç keşfi çalışmalarında, öncelikli olarak hedef moleküllerin veya yolların tanımlanması amaçlanır. Seçilen aday ilaçlar, *Escherichia coli* gibi bir konakta üretilir. İlaç keşfi ve üretim prosesinin geliştirilmesinde gen düzenleyici ağlar kullanılmaktadır

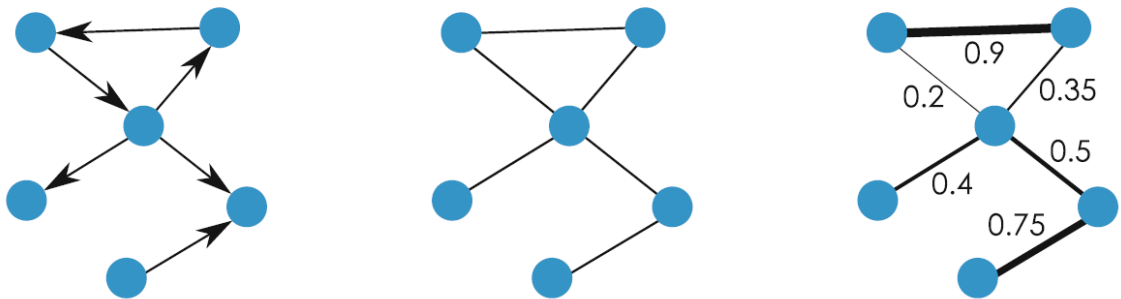
[50]. Gen düzenleyici ağların modellenmesi ile bitkilerin adaptasyon mekanizmaları açığa çıkarılabilir, yüksek oranda metabolit üreten ve çevresel strese karşı dayanıklı mahsuller oluşturulabilir ve önemli tarımsal bitkilerin veriminin artması sağlanabilir [51]. Endüstriyel ve tıbbi açıdan önemli kimyasalların ve malzemelerin sürdürülebilir bir biçimde üretimi için konağın biyolojik ağlarının mühendisliği üzerine yoğunlaşan metabolizma mühendisliği, bilgisayarlı analizlere dayalı modellemelerin yapılmasında farklı türde biyolojik ağları kullanmaktadır. Düzenleyici ağların modellenmesi diğer biyolojik ağ türlerine kıyasla teknik anlamda daha zor olduğundan, sistem metabolizma mühendisliği yaklaşımında düzenleyici ağların etkisi sınırlı kalmıştır; fakat teknolojik gelişmeler sayesinde düzenleyici ağların modellenmesinin daha doğru hale gelmesi ve gen düzenleyici ağların üretim yapılacak konağın genetik manipülasyonu için doğrudan olmasa da dolaylı olarak hedef manipülasyonu açığa çıkarmasından dolayı önemli bilgiler vermesi göz önüne alındığında, gen düzenleyici ağların çıkarımı sistem metabolizma mühendisliği için de önemli bir potansiyel taşımaktadır [52].

Şekil 2.6'da görüldüğü üzere; gen düzenleyici ağ modellenmesi soyutlama ile oluşturulur ve biyolojik süreçlerin farklı seviyelerde işlemesiyle moleküler bileşenler arası etkileşimler ağ modelleri aracılığıyla temsil edilir. Bu modellerin oluşturulması için tersine mühendislik prensipleri kullanılmaktadır [48]. Gen düzenleyici ağların yeniden oluşturulması, sistem biyolojisinin en çok çalışılan konularından biri olmuştur [44].

Gen ifadesi dağılımı, konumlanması ve zamanlaması üzerinden gen düzenleyici ağ çıkarımı, bir genin gen ifadesi düzeyinin bir ya da daha fazla genin fonksiyonu şeklinde modellenebileceği varsayımı üzerine kuruludur [50]. Gen düzenleyici ağların modellenmesinde kullanılan ağ yapılarında da düğümler genel kabul olarak genlerin gen ifadesi seviyesini temsil etmektedir. Genler arası etkileşimleri temsil eden kenarların ağ yapısı içerisinde belirttiği anlam sistemin matematiksel modellenmesine göre değişmektedir [44]. Gen düzenleyici ağ modeli kurulurken, bir genin diğer geni kontrol ettiği bilindiğinde model içerisinde bu iki gen arasında bir bağlantı oluşturulur. Genin kontrol ettiği geni aktive ya da inhibe edip etmediği veya kontrolün yönü bilinmeyebilir. Burada kenarların yönlü olup olmamasına göre ağlar farklı isimler almaktadır [50]. Şekil 2.7'de görüldüğü üzere; kenarların düğümler arası etkileşimler temsil edilirken yönlü ağlar (İng. directed networks), yönsüz ağlar (İng. undirected networks) ve ağırlıklı ağlar (İng. weighted networks) gibi farklı topolojiler kullanılmaktadır. Ağırlıklı ağlar yönlü ya da yönsüz olarak da ifade edilebilir ve ağırlıklar genlerin birbirine bağılılıklarını temsil etmek üzere gen düzenleyici ağların çıkarımında sıklıkla tercih edilir [44].

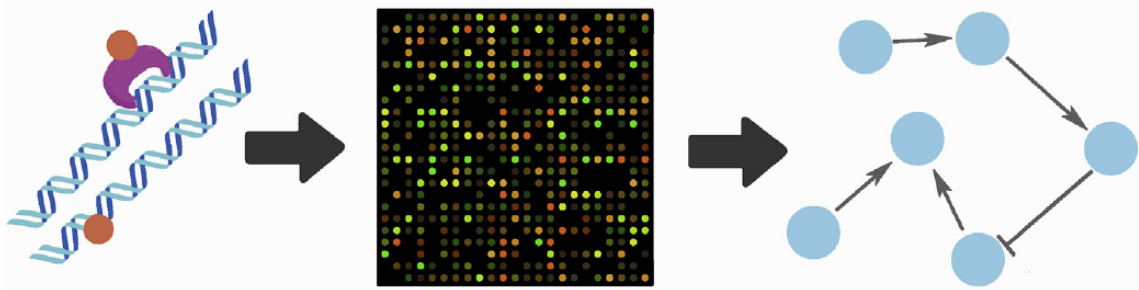


Şekil 2.6 Moleküller arası kompleks etkileşimlerin gen düzenleyici ağ olarak modellenmesi [48]



Şekil 2.7 Ağ türlerine örnek olarak sırasıyla yönlü, yönsüz ve ağırlıklı ağlar. Ağırlıklar kenarların kalınlığıyla da ifade edilebilir ve ağırlıklı ağlar yönlü ya da yönsüz olabilir [44]

Gen düzenleyici ağların çıkarımı mikroRNA'lardan, düzenleyici dizi motiflerinden, gen ifadesi profillerinden [8], ChIP verisinden, transkripsiyonel komplekslerin incelendiği protein-protein etkileşimi verilerinden [53] yapılabildiği gibi gen ifadesi verisinin yanı sıra DNA dizisi verisi gibi farklı biyolojik verilerin entegre edilmesiyle de yapılabilmektedir [8]. Modellemenin genel çıkarımı şematik olarak Şekil 2.8'de gösterilmiştir.

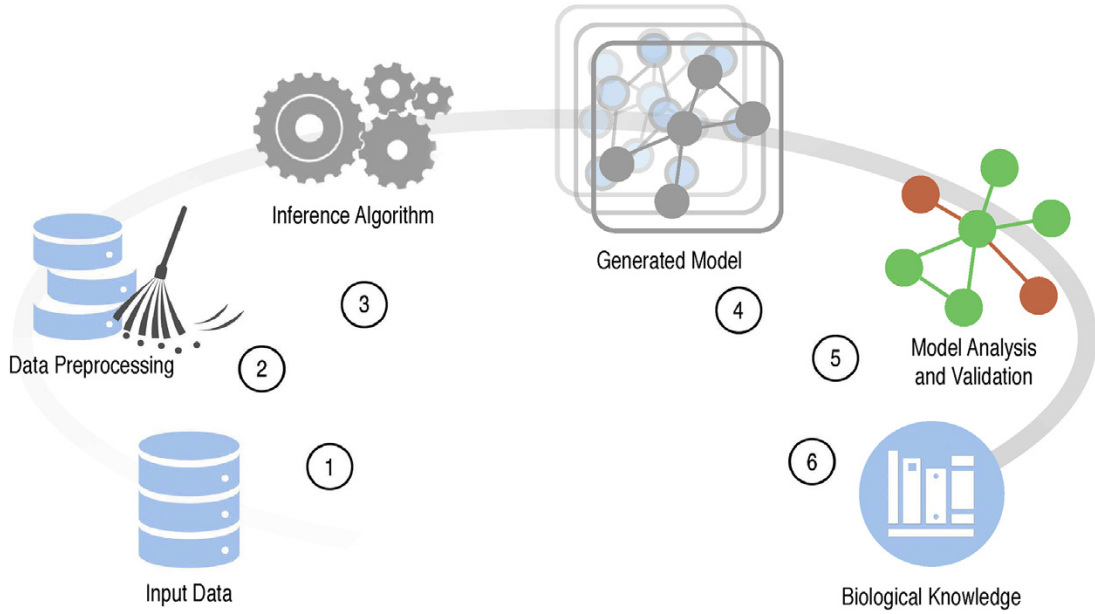


Şekil 2.8 Biyolojik veriden gen düzenleyici ağların soyut model olarak çıkarımı [48]

Uygun çıkarım yönteminin seçimi, çıkarımı yapılacak gen düzenleyici ağın uygulama alanı ve farklı biyolojik kısıtlamalar dikkate alınarak yapılmalıdır [50]. Her ne kadar ChIP ve kütle spektrometresi gibi tekniklerle deneysel olarak protein-protein veya protein-DNA etkileşimlerini doğrudan saptamak mümkün olsa da gen düzenleyici ağların çıkarımı için genellikle mikrodizi veya RNA-Seq aracılığıyla elde edilen gen ifadesi verisi kullanılmaktadır ve bu veriler çeşitli istatistiksel yöntemlerle veya algoritmalarla işlenmektedir [54].

Gen düzenleyici ağların bilgisayarlı çıkarımı bilgi veri tabanı keşfi iş akışı (İng. Knowledge Database Discovery Workflow) ile yapılabilir. Şekil 2.9'da iş akışına uygun olarak gen düzenleyici ağların çıkarımı için izlenen basamaklar görülmektedir. Süreç ilk olarak deneysel ya da veri tabanlarından elde edilen verinin seçilmesi ile başlar [48]. Mikrodizi ve RNA-Seq verilerini barındıran halka açık bir veri tabanı olan Gene Expression Omnibus (GEO) [55] gibi veri tabanlarından veya deneysel olarak elde edilen verilerden gen ifadesi veri kümeleri alınır. Girdi olarak kullanılacak ham veri seçildikten sonra veri kümesi uygun bir hesapsal yöntem ile ön işleme alınır. İşlenen veri için uygun bir algoritma kullanılarak düzenleyici ağ modeli ortaya çıkarılır [48]. Bu aşamada kullanılan yaklaşım istatistiksel veya makine öğrenmesi temelli olabilir [46]. Son aşamada, elde edilen gen düzenleyici ağın optimizasyonu ve validasyonu, bilinen ağlarla karşılaştırılarak yapılır [48].

Gen ifadesi verisi, gen düzenleyici ağların oluşturulması için kritik öneme sahiptir. Burada kullanılan veriler SAGE, mikrodizi ve RNA-Seq gibi çeşitli tekniklerden elde edilir ve zaman serisi (İng. time series) ve pertürbasyon deneylerinden elde edilen statik veri olarak iki farklı gen ifadesi verisi bulunmaktadır [8]. Çevresel veya



Şekil 2.9 Gen düzenleyici ağların çıkarımı için izlenen aşamalar. (1) Girdi verisi deneysel olarak ya da veri tabanlarından elde edilir. (2) Ham veriye uygun bir yöntemle ön işlem uygulanır ve kullanılacak algoritma için hazır hale getirilir. (3) Uygun hale getirilen veri çıkarım yapacak algoritmaya verilir. (4) Algoritmanın veriyi işlemesi ile ağ modeli üretilir. (5) Gen düzenleyici ağ modelinin analizi ve optimizasyonu yapılır. (6) Elde edilen gen düzenleyici ağ modeli biyolojik bilgi birikimi olarak literatürde yerini alır [48]

moleküler düzeyde değişiklik yapılacak şekilde biyolojik sisteme müdahale edildikten sonra, gen ifadesinin tek bir anda ölçülmesiyle elde edilen veri statik veri; birbirini izleyen zaman dilimlerinde ölçülerek elde edilen veri ise zaman serisi verisi olarak adlandırılmaktadır. Zaman serisi verisi, sistem dinamiği hakkında daha fazla bilgi sağlasa da özünde daha karmaşıktır ve daha fazla bilgisayar gücü gerektirmektedir [50]. Zaman serisi verisi biyolojik ağlardaki zaman örüntülerinin araştırılması için daha uygun sonuçlar verirken statik veriler ağ içi etkileşim ve bu etkileşimin yönü hakkında bilgi verir [8]. Bu nedenle, genler arası etkileşimin ve gen ağına dahil olan gizli genlerin ortaya çıkarılmasında statik verinin kullanılması önemli bir avantaj sağlayacaktır. Bununla birlikte, gen ifadesi verisinin ölçülmesinde kullanılan farklı teknolojiler de yapılacak analizde ve bu analiz için uygulanacak yöntemin seçilmesinde önem teşkil etmektedir. Mikrodizi teknolojisinin yüksek gürültü sinyal oranına sahip olması önemli bir dezavantajken yeni nesil dizileme teknolojileri paralel işleme yaptığından bazı klonlama sorunlarının üstesinden gelebilmektedir [50].

Transkriptom verisinin işlenmesiyle gen düzenleyici ağların çıkarımı için çeşitli istatistikî yöntemler ve makine öğrenmesi yöntemleri geliştirilmiştir [46]. Korelasyon katsayısı, Öklit uzaklığı gibi enformasyon teorisi modelleri [48], Boolean ağları [51], regresyon tabanlı modeller [50], Bayesian teorisi modelleri [44], diferansiyel denklem

modelleri [54], rassal orman (İng. random forest), destek vektör makinesi (İng. support vector machine) ve yapay sinir ağıları (İng. artificial neural networks) gibi makine öğrenmesi yöntemleri [52] örnek olarak verilebilir.

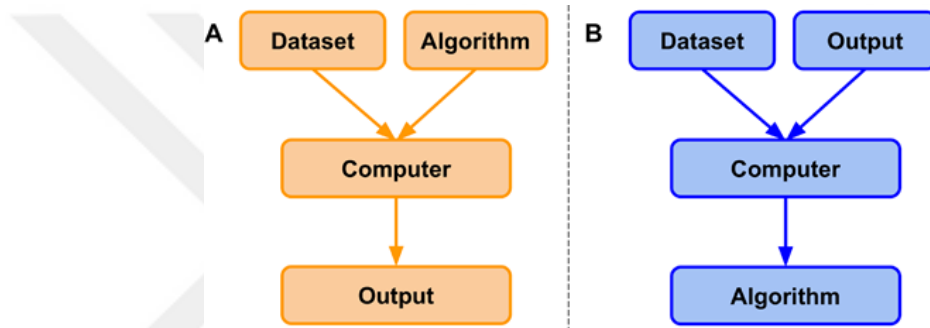


3 DERİN ÖĞRENME

Günümüzde doğal dil işleme ile dillerin çevrilmesinden cerrahi operasyonların tarihinin belirlenmesine kadar geniş bir uygulama alanına sahip olan yapay zeka, kompleks veri kümelerini işleyebilen akıllı ajanların (İng. intelligent agents) çalışıldığı bir bilgisayar bilimi dalıdır. Akıllı ajanlar, çevresini algılayan ve belirli bir amaçta başarı şansını azami seviyeye çıkartmak için gerekli adımları uygulayan araçlar olarak tanımlanmaktadır [56]. Bu alanın asıl amacı, insanlar tarafından gerçekleştirilen görevlerin otomasyonunu sağlamaktır [57]. Yapay zeka uygulamalarının altında sınıflandırılan makine öğrenmesi; tecrübe ederek öğrenen ve analiz yapabilen, öğrendikçe performansını arttıran algoritmalarıdır. Makine öğrenmesi, kaynakların tahsisi ve proses optimizasyonuna yardımcı olarak kullanılmaktadır [56, 58]. Derin öğrenme, makine öğrenmesinden birtakım yöntemleri içinde barındıran bir algoritma sınıfıdır ve makine öğrenmesinin özelleşmiş bir uygulaması olarak tanımlanır. Özellikle görüntü sınıflandırması ve konuşma tanıma gibi uygulamalarda yüksek doğrulukta sonuç veren derin öğrenme teknikleri, günümüzde diller arası çeviriden hesaplamalı kimyaya kadar birçok farklı alanda kullanılmaktadır [59]. Özellikle tahmin gerektiren problemlerdeki performansı nedeniyle derin öğrenme yöntemleri sıklıkla tercih edilir olmuştur. Promotorların tanımlanması, transkripsiyon başlangıç bölgelerinin bulunması, transkripsiyon bağlanma bölgesinin tahmini gibi fonksiyonel genomik uygulamalarında, hücre fenotiplenmesinde, teşhis amacıyla görüntü işlenmesinde derin öğrenme yöntemleri kullanılarak çeşitli çalışmalar yapılmıştır [60]. Benzer şekilde, farklı derin öğrenme yapıları protein-protein etkileşimi ağları, metabolik ağlar, gen düzenleyici ağlar gibi çeşitli biyolojik ağ verilerinin analizlerinde de kullanılmıştır [61].

Makine öğrenmesi alanı, verileri en iyi şekilde işleyen bir algoritmanın geliştirilmesini amaçlamaktadır. Klasik programlama yöntemlerinde algoritma kodlanarak veri kümesiyle birlikte bilgisayara verilir. Bu algoritma sayesinde veri kümesi işlenerek bir çıktı elde edilir. Makine öğrenmesinde ise; verinin alt kümesi ve ilgili çıktısı alınarak bilgisayara verilir. Bu işlem sonucunda bilgisayar tarafından veri kümesindeki

ilişkileri açığa çıkarabilen ve benzer veri kümelerinden çıkarım yapabilme kapasitesine sahip bir algoritma oluşturulur. Şekil 3.1’de bu durum özetlenmiştir [57]. Makine öğrenmesinde; analiz için toplanan verinin karakteristik özellikleri (örneğin; çiçeklerin sınıflandırılmasında taç yaprak uzunluğu, taç yaprak genişliği, renk dağılımı vb.) veya diğer adıyla öznitelikleri (İng. feature) belirlenerek uygun yapıda bir veri kümesi oluşturulur. Bu belirli öznitelikleri içeren veri kümesi kesin referans (İng. ground truth) olarak kullanılır ve eğitim veri kümesi (İng. training data set) bu verilerden oluşturur. Algoritmanın analizi yapılan verinin özelliklerini öğrenmesi ve ilişkiler kurabilmesi için eğitim veri kümesi kullanılır. Eğitim yapıldıktan sonra, algoritmaya girdi olarak beslenen yeni bir verinin analizi yapılabilmekte, algoritma veri hakkında tahminde bulunabilmektedir. Eğitim veri kümesinin sayısının ve büyüklüğünün artması doğruluğu (İng. accuracy) arttırmaktadır [56].



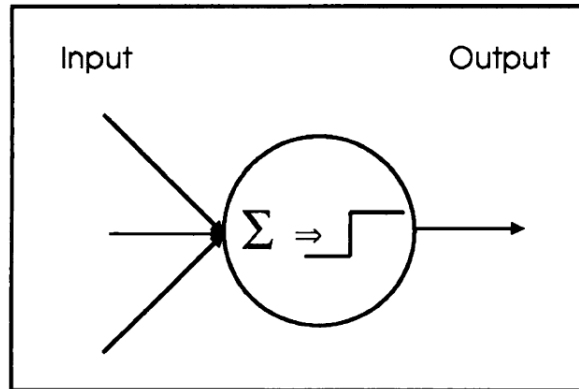
Şekil 3.1 Programlama ve makine öğrenmesi arasındaki farklar. Klasik programlamada (A) algoritma ve veri kümesi kullanılarak çıktı elde edilirken makine öğrenmesinde (B) veri kümesi ve çıktı kullanılarak algoritma elde edilir [57]

Makine öğrenmesinde hem veri için gerekli öznitelik çıkarımının yapılması hem de verinin bilgisayar diline uygun olarak bir çalışma tablosu haline getirilmiş yapılandırılmış veri (İng. structured data) olması gerekmektedir. Büyük, yapılandırılmamış veya çok boyutlu veri kümelerinin işlenmesinde ve özniteliklerin insan müdahalesi olmadan çıkarılmasında derin öğrenme yöntemleri ön plana çıkmaktadır [56]. Makine öğrenmesi yöntemleri, önceden bir ilişki tanımlanmasına ihtiyaç duyulmadan veriden fonksiyonel ilişkilerin öğrenimi için kullanılabilirdiğinden, hesaplamalı biyoloji çerçevesinde biyolojik sürecin temelinde yatan mekanizmalar bilinmese de tahmin modeli çıkarılabildiği için de kullanılmaktadır. [62]. Fakat; geleneksel makine öğrenmesi yöntemleri ham veriyi işlemede kısıtlı kalmıştır. Örüntü tanıma için oluşturulan sistemler, ilgili alan göz önünde bulundurularak ham verinin dönüştürülmesine aracılık eden öznitelik çıkarımının sağlanmasını gerektirmiştir [63]. Derin öğrenme yöntemleri ham veriden doğrudan öğrenebilmekte ve böylece hazır algoritmik yapılar olarak hesaplamalı biyoloji alanında doğrudan tercih edilmektedir [60].

Makine öğrenmesinde algoritmanın performansı özniteliklerin kodlanma biçimine doğrudan bağlı olduğundan sınıflandırma (İng. classification) doğruluğu düşebilmektedir. Bu sorun, derin öğrenmede özniteliklerin algoritmanın içine gömülü bir şekilde kodlanması ile aşılar. Böylece derin sinir ağları (İng. deep neural networks), uygun öznitelikleri keşfederek tahmin doğruluğunu arttırabilir [64]. Makine öğrenmesi yaklaşımlarının aksine ham veriden öznitelik çıkarabilen derin sinir ağları, yeterli düzeyde veri sağlandığı takdirde herhangi bir girdiden uygun çıktıya ulaşabilme potansiyeline sahiptir [60]. Derin öğrenme yöntemleri büyük veri kümelerinin kullanımında daha etkilidir [64] ve verideki yüksek karmaşıklıkta yapıları saptayabilir [61].

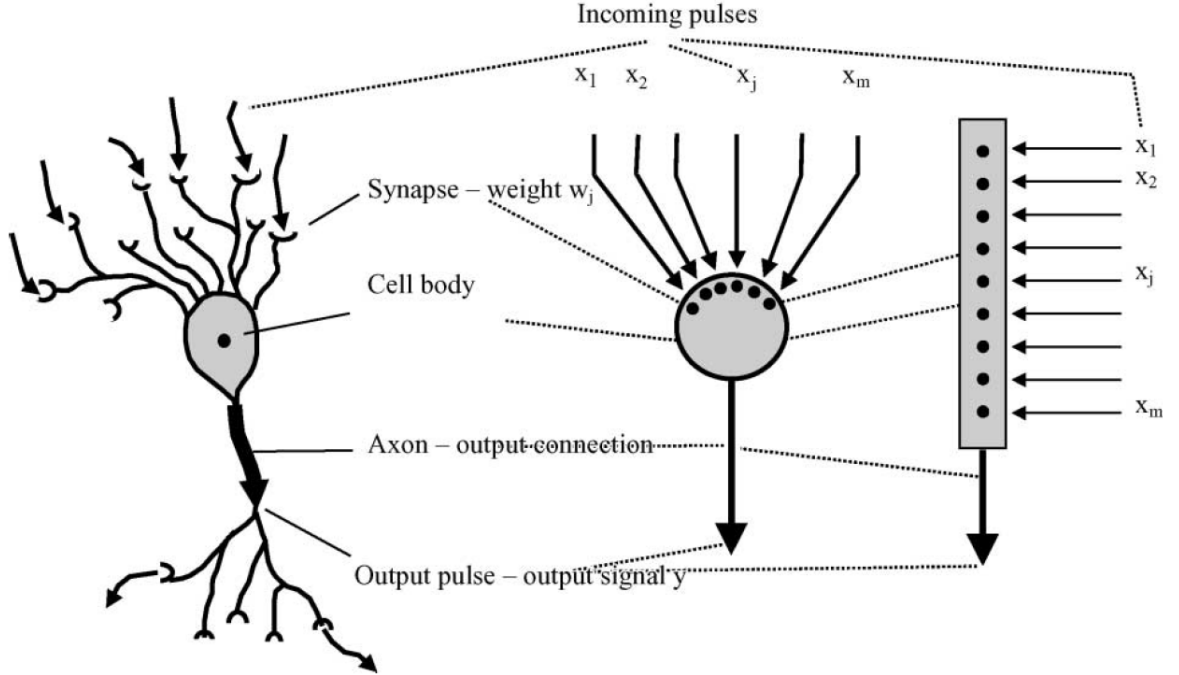
3.1 Yapay Sinir Ağları

Derin öğrenme yaklaşımlarının çıkış noktası; beyindeki sinir hücrelerinin (İng. neuron) bilgiyi nasıl işlediğine dair bir model geliştirilmesi için yapay sinirler üzerine yapılan araştırmalardır [59]. Sinir hücrelerinin dendritler aracılığıyla sinyal alması ve bu sinyali işleyerek aksonlar aracılığıyla iletmesine benzer şekilde girdi ve çıktı bağlantılarına sahip yapay sinir modelleri yapılmıştır [65, 66]. Temel sinir modellenmesine göre; uygun sayıda sinir ve bunlar arasındaki her bir sinaptik bağlantıya atanan ağırlık ile uygun ayarlama yapıldığında, prensipte, her hesaplanabilir fonksiyon hesaplanabilmektedir [66]. Kısaca sinir ağları olarak da adlandırılan yapay sinir ağları [59], sinir adı verilen birimlerin birbirine bağlandığı katmanlardan oluşan matematiksel bir yapıdır [60]. Şekil 3.2'de üç girdi ve bir çıktı sinyali bağlantısı bulunan basit bir sinir birimi görülmektedir. Pratikte bir sinirin binlerce girdi ve çıktısı olabilir. Her sinir, gelen her bilgiyi işleyerek bir çıktı değeri üreten bir birimdir. Bu birimler sinir ağı içerisinde bir matematiksel transfer fonksiyonunu temsil eder [65].



Şekil 3.2 Basit bir yapay sinir birimi. Üç girdi ve bir çıktı bağlantısı bulunmaktadır. Girdi verileri sinirin temsil ettiği transfer fonksiyonu ile işlenir ve bir çıktı değeri elde edilir [65]

Şekil 3.3'te biyolojik sinir ile yapay sinir biriminin karşılaştırılması görülmektedir. Her ne kadar benzerlikler olsa da biyolojik sinir ile yapay sinir arasında bazı temel farklılıklar vardır. Biyolojik sinirin çıktısı sabit genlikte atımlardan oluşan ve frekansı dendritlerden gelen sinyallere göre değişen bir sinyalken yapay sinirin çıktısı belirli bir aralıktaki bir sayıdır. Biyolojik sinirin çıktısı zamanla sürekli olarak değişirken yapay sinirin çıktısı girdiler değiştiği zaman kesikli aralıklarla değişir [65].



Şekil 3.3 Biyolojik sinir ve kolon vektör formuyla birlikte yapay sinir diyagramları [67]

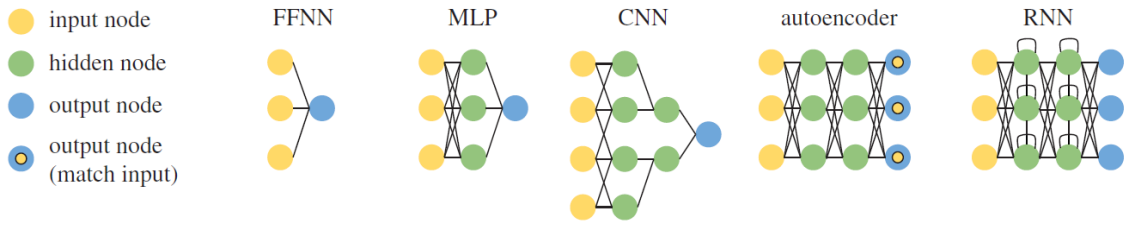
Sinir hücresinin komşu sinir hücrelerinden gelen sinyalleri alması ve işleyerek iletmesi, yapay sinir biriminin transfer fonksiyonunun sonucuna göre, m boyutlu girdi ile orantılı olarak bir y çıktı değeri vermesi ise benzerdir. Herhangi bir girdi sinyali, uygun transfer fonksiyonu kullanılarak, genellikle 0'la 1 arasında değişen sabit maksimum ve minimum sınırları içerisinde bir çıktı değerine dönüştürülür. Şekil 3.3'e göre girdi sinyali (3.1)'deki gibidir [67].

$$X_s = (x_{1s}, x_{2s}, \dots, x_{ms}) \quad (3.1)$$

Sinir birimleri arasındaki bağlantılar ağırlık olarak nitelendirilen bir sinyal kuvvetine sahiptir [65]. Sinirleri temsil eden ağırlık vektörünün genel gösterimi, j siniri için (3.2)'deki gibidir [67].

$$W_j = (w_{j1}, w_{j2}, \dots, w_{ji}, \dots, w_{jm}) \quad (3.2)$$

Sinir ağlarında bilgi girişinin yapıldığı ilk katman girdi katmanını (İng. input layer) ve bilginin dönüşüm sonrası çıkış yaptığı katman çıktı katmanını (İng. output layer) olarak adlandırılır [65]. Bu iki katman arasında sinir birimleri bulunduran katmanlar saklı katman (İng. hidden layer) olarak adlandırılır. Aynı katman içerisindeki sinir birimleri arasında bağlantı yoktur; sinir birimleri komşu katmandaki sinir birimleriyle bağlantı kurar [68]. Bilgi akışı, diyagramlarda genellikle soldan sağa doğrudur. Akışın tek yönlü ve her zaman ileriye doğru olduğu ağlar ileri beslemeli sinir ağı (İng. feed-forward neural network) adlandırılır [65]. Bununla birlikte bazı özel sinir ağı yapılarında gecikme payı bırakarak kendiyle bağlantı kurabilen sinir birimleri bulunabilir. Özyinelemeli sinir ağları (İng. recurrent neural networks) buna örnektir. Şekil 3.4'te bazı sinir ağı yapıları gösterilmiştir. Çok katmanlı algılayıcı (İng. multilayer perceptron), en az bir saklı katman barındırır. Otokodlayıcılar (İng. autoencoders) girdisiyle eşleşen bir çıktı üretmek üzere eğitilen çok katmanlı sinir ağlarıdır. Özyinelemeli sinir ağları ise; sıralı ya da zamana bağlı girdileri hafızada tutan ve ileri beslemeli olmayan özel bir yapıdır [59].



Şekil 3.4 Derin öğrenmede sıklıkla kullanılan farklı sinir ağı yapıları. Sarı düğümler girdi katmanını, yeşil düğümler saklı katmanını, mavi düğümler çıktı katmanını, noktalı düğümler girdiyle eşleşen çıktı katmanını temsil etmektedir. FFNN, ileri beslemeli sinir ağı; MLP, çok katmanlı algılayıcı; CNN, evrişimli sinir ağı; RNN, özyinelemeli sinir ağını belirtmektedir [59]

Sinir ağlarında kullanılan transfer fonksiyonları, sinire beslenen tüm girdilerin ilgili bağlantı ağırlığı değeriyle çarpılmış toplamını veren toplama fonksiyonunu genellikle içermektedir. Toplama fonksiyonunun matematiksel gösterimi (3.3)'teki gibidir. Burada n , j sinir birimine gelen girdi bağlantı sayısını temsil ederken w_{ij} ise j sinir birimi ile önceki katmanda bulunan i sinir birimi arasındaki bağlantının ağırlık değerini belirtmektedir [65].

$$\sum = \sum_{i=1}^n x_i w_{ij} = x_1 w_{1j} + x_2 w_{2j} + \dots + x_n w_{nj} \quad (3.3)$$

Eşik fonksiyonları (İng. thresholding functions), basamak fonksiyonu (İng. step function), hiperbolik tanjant fonksiyonu [65], doğrultulmuş lineer birim fonksiyonu (İng. rectified linear unit function) (ReLU) [62], sigmoid fonksiyonu gibi farklı

transfer fonksiyonları çeşitli sinir ağı yapılarında tercih edilmektedir [68].

Ağırlıklar serbest parametrelerdir ve modelin girdi ve çıktı verilerinden öğrenmesiyle süreç içerisinde elde edilir. Kayıp fonksiyonu (İng. loss function) modelin çıktısının gerçek değerle ne kadar örtüştüğünü ölçer ve öğrenme sağlandıkça kayıp fonksiyonunun değeri azaltılmış olur. Öğrenme süreci boyunca tahmin edilmiş etiketler (İng. labels) gerçek etiketlerle karşılaştırılarak ağırlıkların ve kaybın hesaplanması sağlanır. Geri yayılım (İng. backpropagation) algoritması ile kayıp fonksiyonu her defasında güncellenir. Model eğitimi boyunca farklı parametreler ayarlanarak ağ yapısına uygun parametreler bulunur [62].

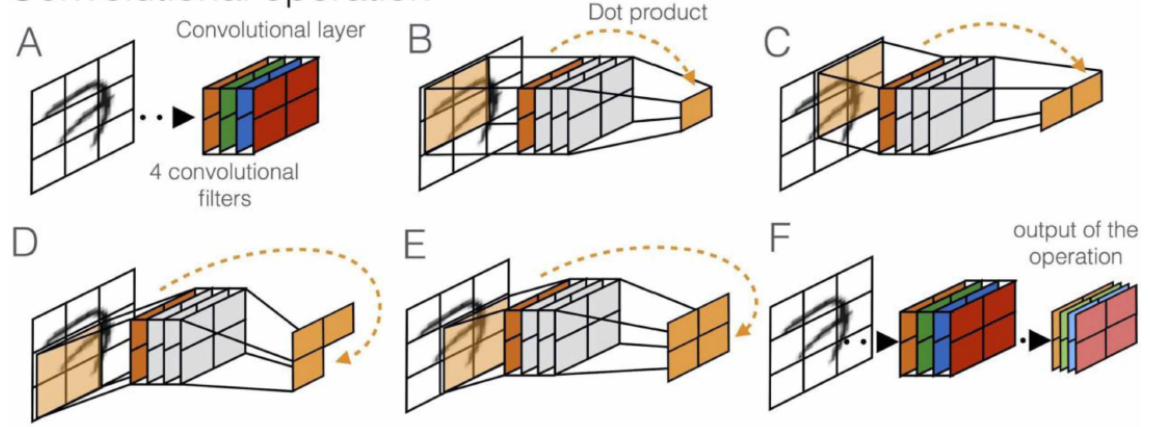
3.1.1 Evrişimli Sinir Ağları

Evrişimli sinir ağları (CNN); adını ilk katmanlarda tam bağlantılı ağlardaki (İng. full connected networks) temel işlem olan iç çarpım yerine evrişimler uygulamasından alan, öznetelikleri ve parametreleri birlikte öğrenebilen sinir ağlarıdır [66]. CNN, iki boyutlu görüntüler gibi çoklu veri tiplerini işleyebilen ve mekansal bilgi analizinde oldukça başarılı olan yapılar olduğundan, nesne tanıma alanındaki geliştirilmelerinden dolayı özellikle biyomedikal görüntülemelerde önemli bir potansiyele sahiptir [69]. Bununla birlikte CNN yapıları düzenleyici elemanlar gibi biyolojik dizilerdeki örüntüleri tanımayı öğrenebilir [70].

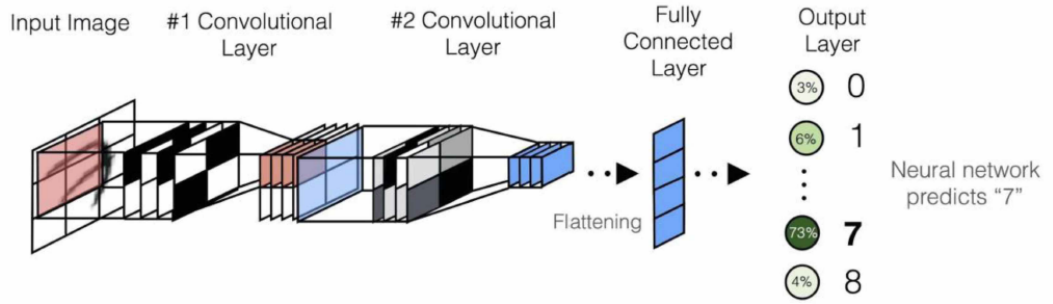
Şekil 3.5'te CNN algoritmasının temel yapısı ve işleyişi tasvir edilmiştir. CNN yapısında bir ya da birden fazla evrişim katmanı içerir ve bu katmanların her birini aynı ağırlık değeriyle kullanarak, örneğin görüntü verisi gibi bir verinin farklı bölgelerini filtreleyebilir. Böylece verideki konumundan bağımsız olarak aynı örüntünün eşlenmesi sağlanır. Bu durum öğrenilmesi gereken parametre sayısını da azaltmaktadır. Evrişimli katmanları takip eden derin ve tam bağlantılı katmanlar ile çıktının üretimi mümkün kılınır [60].

Herhangi bir CNN algoritmasının kullanımı için izlenen temel adımlar; evrişim, lineer olmama (İng. nonlinearity) ve ortaklamadır (İng. pooling) [66]. Şekil 3.6'da bu adımlar temel katmanlar olarak gösterilmiştir. Öznetelik haritası (İng. feature map) adı da verilen toplam ağırlıklı yerel gruplar her evrişim katmanında filtreler sayesinde hesaplanarak verinin korelasyon gösteren alt bölgelerinin eldesi sağlanır. Evrişim katmanları ile filtreler tüm veri kümesi boyunca tekrarlı olarak uygulanmış olur. Böylece öğrenilmesi gereken parametre sayısı azaltılır ve öğrenme verimi artar. Öznetelik haritalarının lineer olmama ile ilgili özellikleri, modelin lineer olmama basamağında ilgili katmanlar tarafından artırılır. Her ortaklama katmanında öznetelik haritalarındaki çakışmayan bölgeler için, alt örnekleme maksimum ya da ortalama

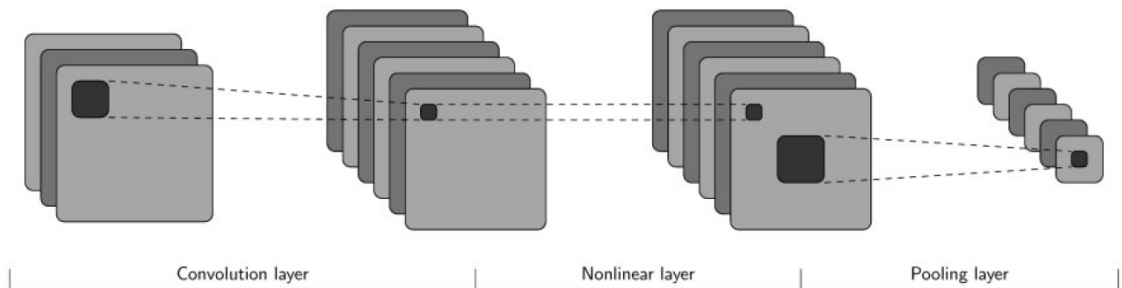
Convolutional operation



Convolutional network



Şekil 3.5 Evrişimli sinir ağlarının yapısı ve işleyişi. A ve F boyunca farklı renklerde filtreler görülmektedir. Bu filtreler özel evrişimli sinirlerdir ve girdi verisinin farklı bölümlerine aynı ağırlık değerleriyle filtreleme yaparlar. Şekilde B-F arasında filtreleme işlemi tasvir edilmiştir. Alt tarafta gösterilen evrişimli ağ modelinde görüldüğü üzere; CNN'ler bir ya da birden fazla evrişim katmanına sahiptir ve bu katmanların ardından çıktıyı üreten tam bağlantılı derin katmanlar bulunmaktadır [60]



Şekil 3.6 Evrişimli sinir ağlarını oluşturan temel yapılar [69]

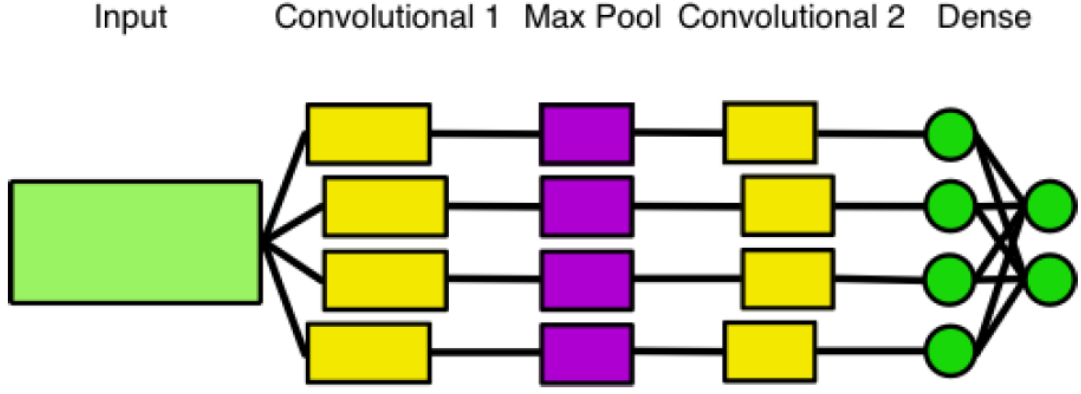
değerler için yapılır. Böylece anlamsal olarak benzer öznitelikler bir şekilde farklılık göstermiş olsalar dahi kullanılabilir hale gelir ve daha kompleks özniteliklerin çıkarımı için yerel öznitelikler kümelenebilir [69].

3.2 CNN ile Transkripsiyon Faktörü-Hedef Gen Çiftlerinin Bulunması

CNN algoritması, çok boyutlu diziler (İng. array) gibi veri tiplerini işleyebildiğinden gen ifadesinin analizinde sıklıkla kullanılmaktadır [2]. Örneğin; gen ifadesi verisinin kategorik veriye dönüştürülerek hücre döngüsünü düzenleyen genlerin bulunduğu bir çalışmada CNN algoritması tercih edilmiştir [71].

Biyolojik dizilerdeki eşsiz öznitelikleri bulmadaki başarısı nedeniyle CNN'ler gen düzenleyici ağların modellenmesinde sıklıkla tercih edilmektedir [52]. Bu amaçla yapılan bir çalışmada tek hücre gen ifadesi verisi kullanılarak eş zamanlı gen ifadesi için CNN modeli geliştirilmiştir ve bu model transkripsiyon faktörü-hedef genlerin tahmininde kullanılmıştır. Çalışmada her bir çift için gen ifadesi verileri görüntüye dönüştürülmüş ve CNN kullanılarak gen düzenleyici ağların çıkarımı için önerilmiştir [72]. Yine tek hücre RNA-Seq verisinin kullanıldığı bir diğer çalışmada üç boyutlu eş ifade matris analizi önerilmiş ve üçlü genlerin üç boyutlu eş ifade matrisleri ile üç boyutlu CNN algoritması kullanılarak gen düzenleyici ağların çıkarımı üzerinde durulmuştur [73]. *Drosophila* göz gelişimi için gen düzenleyici ağların çıkarımını hedefleyen bir başka çalışma ise; *Drosophila* embriyolarının *in situ* hibridizasyon görüntüleri ile gen ifadesinin mekansal örüntülerini içeren görüntü verilerini işleyen CNN tabanlı bir derin öğrenme modeli kullanmıştır [74].

Farklı gen ifadesi veri türlerinin kullanıldığı ve gen düzenleyici ağların çıkarıldığı bu çalışmaların yanı sıra, gen ifadesi verisinin mikrodizi yönteminden elde edildiği bir diğer çalışmada, CNN modeli önerilerek *Arabidopsis thaliana* için transkripsiyon faktörü-hedef gen çiftlerinin tahmini için kullanılmıştır. AtRegNet veri tabanındaki bilinen çiftler ile bu çiftlerden rastgele şekilde oluşturulan sahte çiftler ve farklı koşullardaki abiyotik strese maruz bırakılan *Arabidopsis thaliana*'lardan elde edilen AtGenExpress veri setindeki gen ifadesi verileri işlenmiş, bu verilerin girdi verisi olarak kullanılmasıyla bir CNN modeli geliştirilmiştir. 3.7'de kullanılan ve sonradan geliştirilen CNN modelinin temel tasarımı gösterilmiştir. Gen ifadesi verilerini içeren matrisler hazırlanarak önce ilk katman olan evrişim katmanına, ardından maksimum ortaklama katmanına, sonrasında ikinci evrişim katmanına ve son olarak yoğunluk ağ katmanına beslenmiştir. Optimize edilen CNN modeli ile %95 doğruluk değeri elde edilmiştir [75].



Şekil 3.7 Çalışmada kullanılan CNN modelinin temel şematik gösterimi [75]

Geliştirilen CNN modeli bu tez çalışmasına adapte edilerek bu çalışmada; TRRUST (Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining) [76] veri tabanından elde edilen insan transkripsiyon faktörü-hedef gen çiftleri ve Genotype Tissue Expression Project (GTEx) [77] ile The Cancer Genome Atlas (TCGA) [78] projelerinden elde edilen normal örnekler ve tümör örneklerini içeren RNA-Seq verileri kullanılmış ve bu veriler işlenerek CNN yaklaşımıyla insan transkripsiyon faktörü-hedef gen çiftlerinin tahminini sağlayan bir modelin geliştirilebilmesi araştırılmıştır.

Bu çalışmada temel olarak insan transkripsiyon faktörü-hedef gen çiftleri ve RNA-Seq'ten elde edilen gen ifadesi verileri kullanılmış ve bu verilerin evrişimli sinir ağı (CNN) algoritmasında girdi olarak kullanılmasıyla insan transkripsiyon faktörü-hedef gen çiftlerinin tahmini için bir model oluşturulması amaçlanmıştır.

TRRUST¹ , 20 milyon Pubmed makalesinden cümle tabanlı metin madenciliği (İng. sentence-based text mining) kullanıldıktan sonra 11237 tanesinden elle düzenlenerek çıkarılmış transkripsiyon faktörü-hedef etkileşimlerini, insan ve fare transkripsiyonel düzenleyici ağlarını içeren bir veri tabanıdır. TRRUST'ın güncel versiyonu insan için 800 transkripsiyon faktörünün toplamda 8444 hedef düzenleyici ilişkisini, fare için 828 transkripsiyon faktörünün 6552 hedef düzenleyici ilişkisini barındırmaktadır. Bununla birlikte, aktivasyon ve baskılanma olarak toplamda 8972 etkileşim için düzenleyici ilişkinin şeklini de sağlamaktadır [76]. Bu çalışmada kullanılan insan transkripsiyon faktörü-hedef çiftleri TRRUST v2'den çekilmiştir.

Açık kaynak kodlu iş akışı yazılımı olan Toil [79] aracılığıyla TCGA [78] ve GTEx [77] çalışmalarına ait örneklerin DESeq2 normalizasyonundan sonra log2 dönüşümü yapılan gen ifadesi verileri elde edilmiştir. BioMart Grch 37 versiyonundaki² protein kodlayan genlerin çekilmesi ve her iki veri kümesiyle birleştirilip filtrelenmesi ile bu veriden toplamda elde edilen gen sayısı 19744 olmuştur. TCGA'dan elde edilen tümör örneklerinin sayısı 9182, GTEx'ten elde edilen normal örneklerin sayısı 7775'tir.

TRRUST veri tabanından alınan transkripsiyon faktörü-hedef çiftleri ile BioMart Grch 37 versiyonundaki protein kodlayan genleri içeren insan veri kümesi birleştirilerek yalnızca protein kodlayan genlerin alınması sağlanmıştır. GTEx ve TCGA projelerinden çekilen veri ile TRRUST veri tabanından gelen transkripsiyon faktörü-hedef gen çiftlerinin ne kadarının tüm genler içerisinde bulunduğu bakılmıştır. Kendi kendini baskılayan ya da aktive eden transkripsiyon faktörleri filtrelenmiştir. Bu yolla

¹<https://www.grnpedia.org/rrust/>

²<http://grch37.ensembl.org>

toplamda 8312 transkripsiyon faktörü-hedef gen çifti elde edilmiştir. Elde edilen veri kümesi Tablo 4.1’de gösterilmiştir.

Tablo 4.1 Bilinen transkripsiyon faktörü-hedef gen çiftleri veri kümesi. İlk 10 çift örnek olarak gösterilmiştir

TF	Hedef
ENSG00000275700	ENSG00000087088
ENSG00000275700	ENSG00000124762
ENSG00000275700	ENSG00000142515
ENSG00000275700	ENSG00000136997
ENSG00000275700	ENSG00000141510
ENSG00000097007	ENSG00000087088
ENSG00000097007	ENSG00000171791
ENSG00000097007	ENSG00000113916
ENSG00000097007	ENSG00000118971
ENSG00000097007	ENSG00000124762

Yalnızca protein kodlanan genlerin çıkarıldığı, GTEx verilerinden hazırlanan veri kümesi 19744 gen için 7775 normal örnekten alınan gen ifadesi verilerini içermektedir. Kolaylık olması açısından veri kümesinin ilk 10 satırı ve ilk 2 sütunu Tablo 4.2’de gösterilmiştir. Benzer şekilde, TCGA verilerinden hazırlanan veri kümesi 19744 gen için 9182 tümör örneğinden alınan gen ifadesi verilerini içermektedir.

Tablo 4.2 GTEx projesinden oluşturulan ve içerisinde her gen için normal örneklerden elde edilen gen ifadesi seviyesini içeren veri kümesi. Elverişli olması için bir kısmı gösterilmiştir

	GTEX-UTHO-1226-SM-3GAEE	GTEX-146FH-1726-SM-5QGQ2
ENSG00000167578	8.0504	9.2276
ENSG00000078237	6.6236	7.9018
ENSG00000146083	10.1114	10.5426
ENSG00000158486	3.9405	2.8614
ENSG00000198242	13.2168	13.2903
ENSG00000134108	11.4848	11.6365
ENSG00000172137	6.2410	6.5051
ENSG00000276644	7.5066	5.4537
ENSG00000094963	13.6543	11.2910
ENSG00000182141	7.8674	6.8952

Çalışmada kullanılan CNN’in eğitimi sınıflandırma tabanlı olduğu için [2, 75] kullanılacak veride bilinen transkripsiyon faktörü-hedef gen çiftlerinin yanı sıra

ikinci bir etikete daha ihtiyaç duyulmaktadır. Bu nedenle, rastgele transkripsiyon faktörü-hedef gen çiftleri seçilerek 8312 çiftten oluşan yeni bir veri kümesi hazırlanmıştır. Pozitif veri olarak da nitelenen bilinen transkripsiyon faktörü-hedef gen çiftleri TRUE, negatif veri olarak nitelenen rastgele oluşturulan transkripsiyon faktörü-hedef gen çiftleri FALSE olarak etiketlenmiştir. Negatif ve pozitif çiftler hem TCGA hem GTExin filtrelenmiş veri kümesiyle birleştirilmiş, ardından toplamda 16624 çifti barındıran TCGA ve GTEx için birer veri kümesi oluşturulmuştur. Eğitimde kullanılmak üzere oluşturulan bu veri kümeleri TCGA için 16624 x 2 x 9182 x 1 boyutlu, GTEx için 16624 x 2 x 7775 x 1 boyutlu dizileri içermektedir. Her iki veri kümesi de yaklaşık olarak %80 eğitim verisi, %10 test verisi ve %10 validasyon verisi olmak üzere bölümlenmiştir. Eğitim için kullanılan CNN modelinin oluşturulması için gerekli kodlar Tablo 4.3'te gösterilmiştir.

Tablo 4.3 Model ve kullanılan parametrelerin kodları

```

1 model <- keras_model_sequential() %>%
2   layer_conv_2d(filters = 32, kernel_size = c(2,5),
3     activation = "relu", input_shape = c(2,num_samples,1),
4     padding = "same") %>%
5   layer_max_pooling_2d(pool_size = c(2,2)) %>%
6   layer_conv_2d(filters = 64, kernel_size = c(2,5),
7     activation = "relu", padding = "same") %>%
8   layer_flatten() %>%
9   layer_dense(units = 64, activation = "relu") %>%
10  layer_dense(units = 1, activation = "sigmoid")
11
12 summary(model)
13
14 model %>% compile(
15   optimizer = "rmsprop",
16   loss = "binary_crossentropy",
17   metrics = c("accuracy")
18 )
19
20 history <- model %>% fit(
21   x_train,
22   y_train,
23   epochs = 15,
24   batch_size = 64,
25   validation_data = list(x_val, y_val)
26 )

```

GTEX için kullanılan CNN'in yapısı ve model parametreleri Tablo 4.4'te görülmektedir.

Tablo 4.4 GTEX verileriyle eğitilen CNN'in yapısı ve model parametreleri

```

1 Model: "sequential_1"
2 -----
3 Layer (type)          Output Shape          Param #
4 =====
5 conv2d_3 (Conv2D)      (None, 2, 7775, 32)    352
6 -----
7 max_pooling2d_1        (None, 1, 3887, 32)    0
8 -----
9 conv2d_2 (Conv2D)      (None, 1, 3887, 64)    20544
10 -----
11 flatten_1 (Flatten)    (None, 248768)         0
12 -----
13 dense_3 (Dense)        (None, 64)             15921216
14 -----
15 dense_2 (Dense)        (None, 1)              65
16 =====
17 Total params: 15,942,177
18 Trainable params: 15,942,177
19 Non-trainable params: 0
20 -----

```

TCGA için kullanılan CNN'in yapısı ve model parametreleri Tablo 4.5'te görülmektedir.

Tablo 4.5 TCGA verileriyle eğitilen CNN'in yapısı ve model parametreleri

```

1 -----
2 Layer (type)          Output Shape          Param #
3 =====
4 conv2d_1 (Conv2D)      (None, 2, 9182, 32)    352
5 -----
6 max_pooling2d          (None, 1, 4591, 32)    0
7 -----
8 conv2d (Conv2D)        (None, 1, 4591, 64)    20544
9 -----
10 flatten (Flatten)      (None, 293824)         0
11 -----
12 dense_1 (Dense)        (None, 64)             18804800
13 -----
14 dense (Dense)          (None, 1)              65
15 =====
16 Total params: 18,825,761
17 Trainable params: 18,825,761
18 Non-trainable params: 0
19 -----

```

TCGA ya da GTEx veri kümesiyle eğitilmiş olan her iki CNN modeli ile tümördeki transkripsiyon faktörü-hedef gen çiftleri ve normal transkripsiyon faktörü-hedef gen çiftlerinin bulunması için tahmin yapılmıştır. Bu amaçla toplam 781 transkripsiyon faktörü, veri kümesindeki kalan her bir gen ile eşleştirilerek her iki CNN modelinden de çift olma olasılığı ve çift olma tahmini için değerler elde edilmiştir. Eşik değeri olarak 0,5 değeri kullanılmış ve olasılık olarak 0,5'in üzerinde değer alan gen çiftleri ilgili CNN modeli tarafından tahmin olarak 1 değeri alarak çift olarak nitelendirilmiş, eşik değeri ve altına tekabül eden olasılık değerlerini alan gen çiftleri ise tahmin olarak 0 değeri alarak çift değildir şeklinde nitelendirilmiştir.



5 SONUÇ VE ÖNERİLER

Hücre içi hayati önem teşkil eden çeşitli biyolojik etkinliklerde görev alan genler arasındaki işlevsel ilişkinin ortaya çıkarılması veya sağlıklı ve hasta bireyler arasındaki fenotip farkına neden olan faktörlerin belirlenmesi için, yüksek çıktılı veri kümelerinden elde edilen anlamlı bilginin uygun yöntemlerle işlenmesi ve genler arası etkileşimin ortaya çıkarılması gerekmektedir [6]. Her bir gen çifti için deney düzeneği oluşturularak genler arasındaki ilişkinin deneysel olarak tespit edilmesi hem zor hem de elverişsiz bir işlem olduğundan, mikrodizi ve RNA-Seq teknolojilerinden elde edilen gen ifadesi verilerinin nicel analizi yapılmakta ve çeşitli algoritmalar yardımıyla biyolojik ağların çıkarımı yapılmaktadır [47]. Bu çalışmada, tümör örneklerinden ve normal örneklerden elde edilen insan RNA-Seq verileri kullanılarak transkripsiyon faktörü-hedef gen çiftlerinin belirlenmesi için iki ayrı CNN modelinin geliştirilmesi hedeflenmiştir. Bu amaçla, MacLean [75] tarafından çalışması yapılan ve *Arabidopsis thaliana*'ya ait transkriptom verileri üzerinde uygulanan CNN yaklaşımı kullanılmıştır.

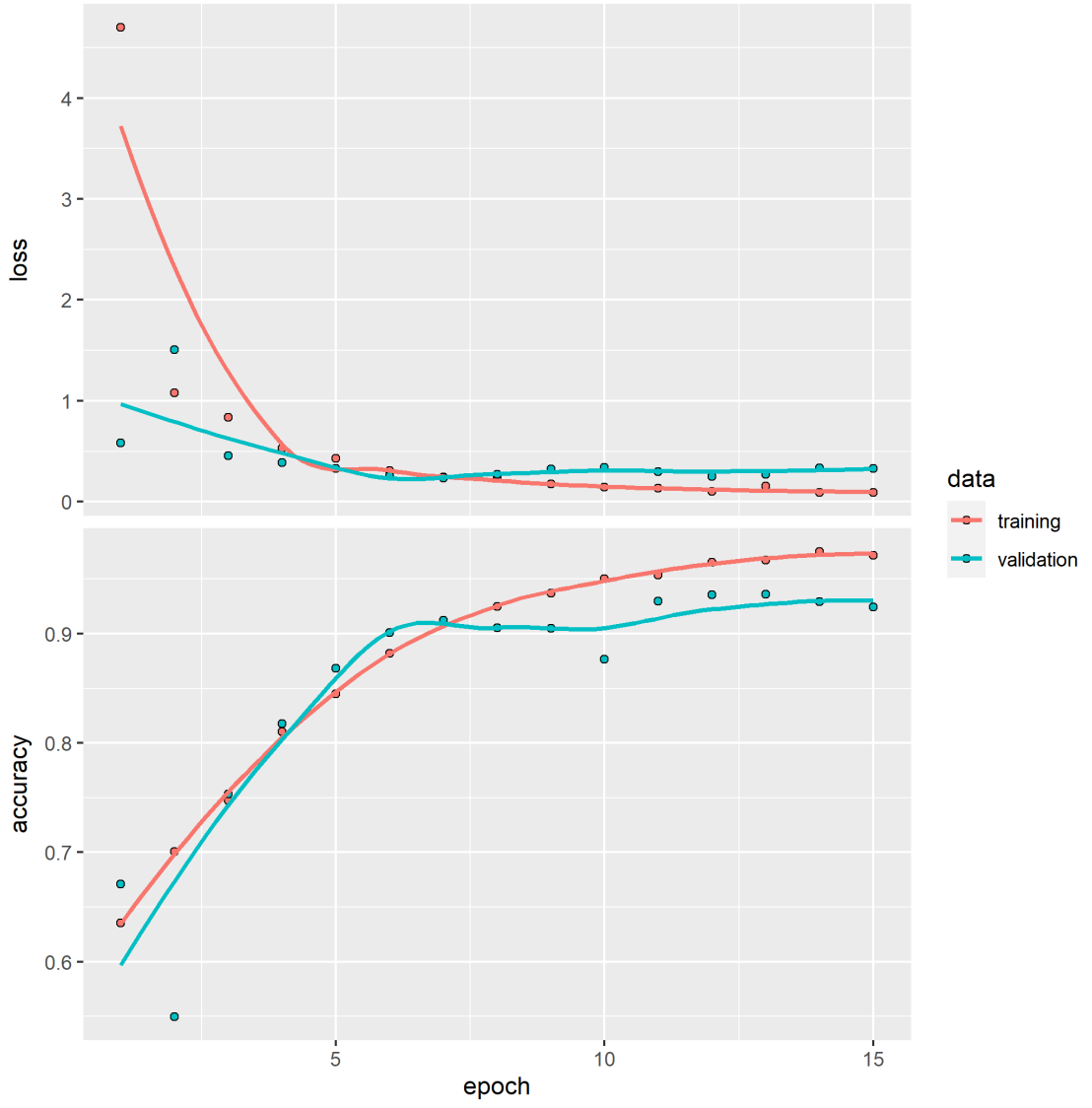
Tablo 5.1'de TRRUST, GTEx ve TCGA veri tabanlarından çekilen ve ön işleme yapıldıktan sonra elde edilen veriler özet bir şekilde sunulmuştur.

Tablo 5.1 Çalışmada kullanılan verilerin özet tablosu

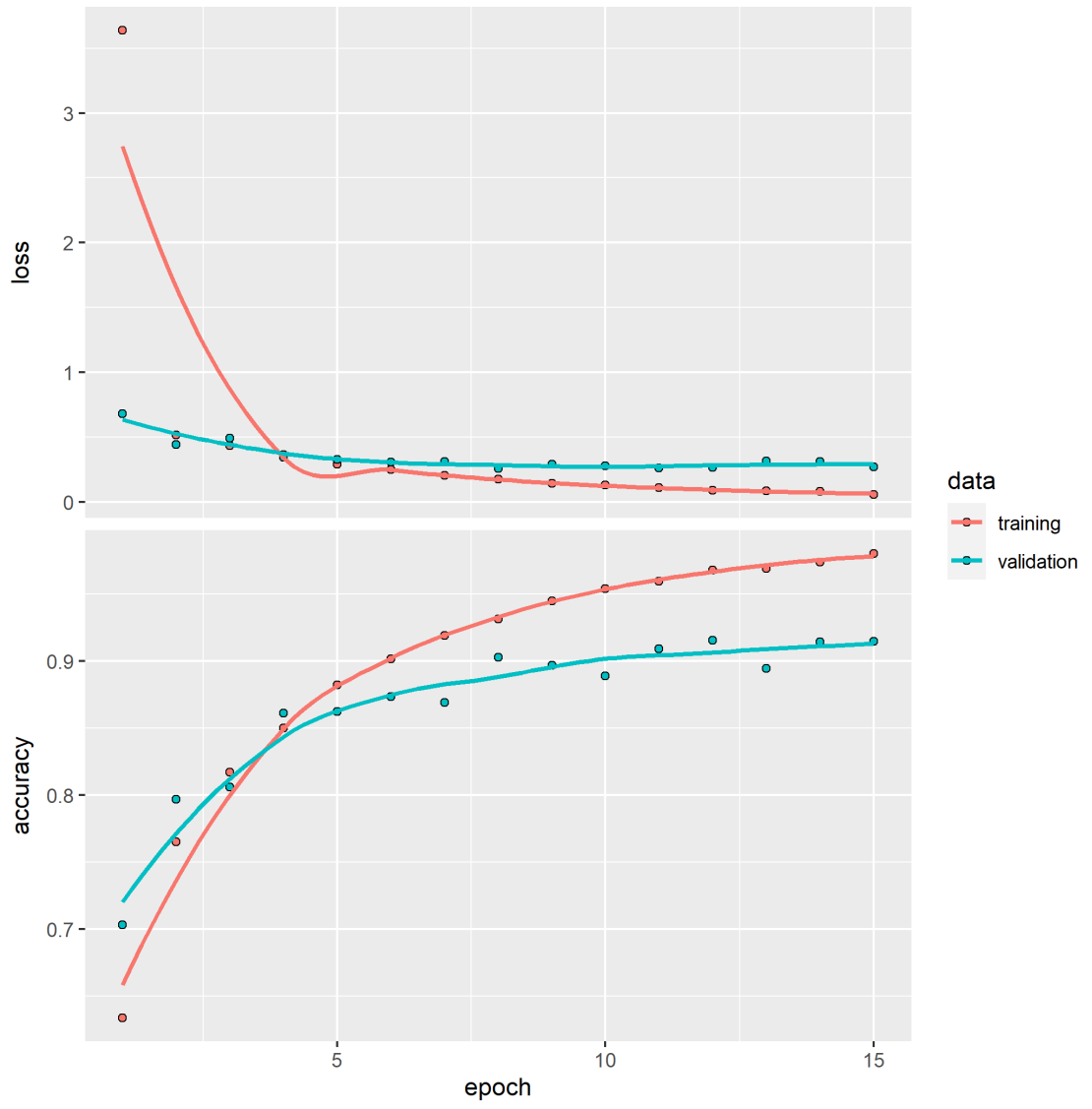
Veri Türü	Veri Niceliği
Transkripsiyon Faktörü	781
Transkripsiyon Faktörü-Hedef Gen Çifti	8312
Bilinen ve Sahte Gen Çiftleri	16624
Toplam Gen	19744
Normal Örnek	7775
Tümör Örneği	9182
GTEx Dizisi Boyutu	[16624, 2, 7775, 1]
TCGA Dizisi Boyutu	[16624, 2, 9182, 1]

GTEEx ve TCGA projelerinden çekilen ve BioMart aracılığıyla yalnızca protein kodlayan genleri içerecek şekilde filtrelenen toplam 19744 gen ve gen ifadesi verisi, TRRUST veri tabanından çekilen ve filtrelenen veri ile birlikte işlenerek toplamda 8312 adet transkripsiyon faktörü-hedef gen çifti ve sonrasında yine 8312 adet rastgele seçilerek oluşturulan sahte transkripsiyon faktörü-hedef gen çifti içeren büyük bir veri kümesi oluşturulmuştur. Çift olup olmamasına negatif ya da pozitif olarak etiketlenen veriler GTEEx için $16624 \times 2 \times 7775 \times 1$ boyutlu, TCGA için ise $16624 \times 2 \times 9182 \times 1$ boyutlu olacak şekilde iki farklı dizide toplanmıştır. Oluşturulan iki dizi de %80 eğitim verisi, %10 test verisi ve %10 validasyon verisi olacak şekilde bölünmüştür. CNN'e beslenen dizi, temelde bireysel olarak 16624×2 boyutlu uzun matrislerden oluşmaktadır. TCGA için 9182 tümör örneği sayısından, GTEEx için 7775 normal örnek sayısından gelen verilerdir. Kullanılan veri, art arda eklenmiş, iki boyutlu görüntü verileri gibi düşünülebilir. Böyle bir veri türü, CNN ile modelleme için elverişli bir yapıdadır.

Bu çalışmada kullanılan derin öğrenme algoritması, ardışık evrişim, ortaklama, düzleştirme ve derin ağ katmanlarından oluşan bir CNN modelidir. Tümör örneklerinin girdi verisi olarak kullanıldığı CNN'in eğitimi sonucu ortaya çıkan eğitim grafiği hem eğitim hem validasyon için kayıp ve doğruluk değerlerini göstermekte ve Şekil 5.1'de görülmektedir. Tümör örnekleri için geliştirilen CNN modeli, 15 döngüden sonra %97 doğruluğa ulaşırken %92 validasyon doğruluğu göstermiştir. Normal örneklerin girdi verisi olarak kullanıldığı CNN'in eğitimi sonucu ortaya çıkan eğitim grafiği hem eğitim hem validasyon için kayıp ve doğruluk değerlerini göstermekte ve Şekil 5.2'de görülmektedir. Normal örnekler için geliştirilen CNN modeli, 15 döngüden sonra %97 doğruluğa ulaşırken %91 validasyon doğruluğu göstermiştir. Test verisi kullanıldığında; tümör örnekleriyle eğitilen CNN için %92 doğruluk değeri elde edilirken normal örneklerle eğitilen CNN için %90 doğruluk değeri elde edilmiştir. Böylece aşırı öğrenme olmadan insan transkripsiyon faktörü-hedef gen çiftlerinin tahmininde kullanılabilecek iki farklı CNN modeli geliştirilmiştir.



Şekil 5.1 TCGA verileri kullanılarak eğitilen modelin eğitim ve validasyon verileri için gösterdiği eğitim grafiği



Şekil 5.2 GTEx verileri kullanılarak eğitilen modelin eğitim ve validasyon verileri için gösterdiği eğitim grafiği

Geliştirilen her iki CNN modeli de toplamda 19744 gen ve ilgili örnekten gelen gen ifadesi verisini içeren filtrelenmiş veri kümeleri ile test edilerek her bir gen için transkripsiyon faktörü-hedef gen çiftlerinin tahmin değerleri elde edilmiştir. Eşik değeri olarak 0,5 kullanılmıştır. 0,5 üzerinde olasılık değeri alan gen çiftleri, CNN modeli tarafından 1 değeri alarak çift olarak nitelendirilmektedir. Tablo 5.2’de GTEx verisi kullanılarak eğitilen CNN modelinin tek bir gen için yaptığı tahmin ve olasılık değerleri görülmektedir.

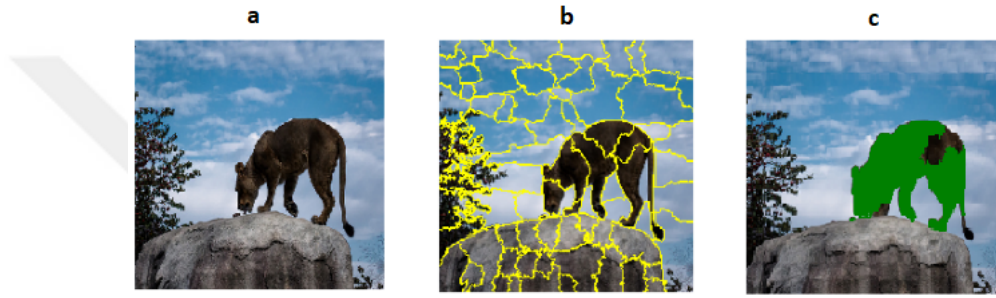
Tablo 5.2 CNN kullanılarak bir gen için elde edilen çift tahminleri. İlk 6 satırı gösterilmiştir

Transkripsiyon Faktörü	Hedef Gen	Olasılık	Tahmin
ENSG00000001167	ENSG00000167578	0.698	1
ENSG00000001167	ENSG00000078237	0.895	1
ENSG00000001167	ENSG00000146083	0.918	1
ENSG00000001167	ENSG00000158486	0.0820	0
ENSG00000001167	ENSG00000198242	0.863	1
ENSG00000001167	ENSG00000134108	0.914	1

Yapılan bu çalışma ile oluşturulan CNN modelleri, tümör örnekleri ve normal örnekler için ayrı ayrı olmak üzere; iki gen için çifttir ya da değildir şeklinde tahmin yapmaktadır. Gelecek çalışmalarda uygun verilerin elde edilmesi ve işlenmesi koşuluyla, çift olan ve birbirini düzenleyen genlerin aktivasyon ya da baskılama işlevlerinin tahmini için CNN modellerinin geliştirilmesi üzerine çalışılabilir. TRRUST veri tabanından alınan ve bu çalışmada kullanılan gen çifti verilerinden kendi kendini baskılayan ya da aktive eden transkripsiyon faktörleri filtrelendiğinden, ileriki çalışmalar bu transkripsiyon faktörlerini de içerecek biçimde genişletilebilir.

Derin öğrenmeyi de kapsayan makine öğrenmesi yöntemleri oldukça karmaşık ve insan tarafından algılanması güç matematiksel modeller şeklinde çalıştığından, bu algoritmaların yorumlanabilirliğini sağlamak için çeşitli yöntemler geliştirilmiştir. Kara kutu olarak da adlandırılan bu gibi algoritmaların nasıl çalıştığını açıklamak için geliştirilen yöntemlerden biri de LIME (Local Interpretable Model-Agnostic Explanation) yöntemidir [80, 81]. LIME algoritması, görüntülerde nesne tanıma uygulamalarında açıklayıcı algoritma olarak kullanılabilir. LIME algoritması görüntüdeki önem arz eden süperpikselleri ortaya çıkararak, incelenen modelin neye göre tahminde bulunduğunu gösterebilme kapasitesine sahiptir [81]. Genel olarak, görüntü sınıflandırma problemleri için geliştirilen teknikler çıktının girdi katmanına belirli fonksiyonlarla geri yayılımını içermekte, bazı teknikler ise modele özgü olarak tasarlanmaktadır. Bu teknikler her bir pikselin önem ölçüsüne göre

değer aldığı bir ısı haritası oluşturur. LIME algoritması ise; görüntüyü süperpiksellere bölümler ve lineer model kullanarak her bir süperpikselin önem ölçüsüne göre değerini belirler. LIME algoritması geri yayılım tekniği kullanmadığından ve modele özgü olmadığından farklı makine öğrenmesi algoritmalarında da kullanılabilir [82]. Şekil 5.3'te LIME algoritmasının bir görüntü işleme modeline uygulanması ve modelin tahminini hangi faktörlere göre oluşturduğunun belirlenmesi için bölümlenme ve süperpiksel oluşturulması görülmektedir. Görüntü verisi önce bölümlenerek süperpiksellere ayrılmış, ardından LIME algoritmasının en önemli 5 pozitif etken olarak belirlediği katsayılar görüntüde vurgulanmıştır. Böylece model hakkında varsayımda bulunulmadan uygulanan modelin çalışma şekline ve tahminini hangi etkenlere göre yaptığına dair çıkarımlar yapılabilir [83].



Şekil 5.3 LIME algoritmasının görüntü verisinin işlendiği tahmin modelinde uygulaması. (a) Bu örnekte görüntü verisinin aslan içerip içermediğini tahmin eden algoritmanın LIME ile açıklanması hedeflenmektedir. (b) Açıklaması yapılan görüntü bölünerek süperpiksellere ayrılmıştır. (c) LIME algoritması varsayılan parametrelerle çalıştırıldığında pozitif etki veren en önemli ilk 5 katsayı şekilde vurgulanmıştır [83]

Bu çalışmada geliştirilen modellerin özellikle sağlık ile ilgili çalışmalarda kullanılabileceği göz önünde bulundurulduğunda, gelecek çalışmalarda, CNN modellerinin gen ifadesi verilerini temsil eden hangi süperpiksellere, dolayısıyla hangi etkenlere göre tahmin yaptığı LIME algoritması ile analiz edilebilir ve modellerin tahmin mekanizması açıklanabilir.

- [1] M. Mahmud, M. S. Kaiser, A. Hussain, S. Vassanelli, "Applications of deep learning and reinforcement learning to biological data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2063–2079, 2018.
- [2] C. Gunavathi, K. Sivasubramanian, P. Keerthika, C. Paramasivam, "A review on convolutional neural network based deep learning methods in gene expression data for disease diagnosis," *Materials Today: Proceedings*, vol. 45, pp. 2282–2285, 2021. DOI: 10.1016/j.matpr.2020.10.263.
- [3] M. K. Elbashir, M. Ezz, M. Mohammed, S. S. Saloum, "Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data," *IEEE Access*, vol. 7, pp. 185 338–185 348, 2019. DOI: 10.1109/access.2019.2960722.
- [4] B. Lyu, A. Haque, "Deep learning based tumor type classification using gene expression data," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM, 2018, pp. 89–96. DOI: 10.1145/3233547.3233588.
- [5] H. Choi, K. J. Na, "A risk stratification model for lung cancer based on gene coexpression network and deep learning," *BioMed Research International*, vol. 2018, pp. 1–11, 2018. DOI: 10.1155/2018/2914280.
- [6] K. Kishan, R. Li, F. Cui, Q. Yu, A. R. Haake, "GNE: A deep learning framework for gene network inference by aggregating biological information," *BMC Systems Biology*, vol. 13, no. S2, pp. 1–14, 2019. DOI: 10.1186/s12918-019-0694-y.
- [7] S. Biswas, S. Acharyya, "Neural model of gene regulatory network: A survey on supportive meta-heuristics," *Theory in Biosciences*, vol. 135, no. 1, pp. 1–19, 2016. DOI: 10.1007/s12064-016-0224-z.
- [8] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, Z. Zakaria, "A review on the computational approaches for gene regulatory network construction," *Computers in Biology and Medicine*, vol. 48, pp. 55–65, 2014. DOI: 10.1016/j.combiomed.2014.02.011.
- [9] B. Schmidt, A. Hildebrandt, "Deep learning in next-generation sequencing," *Drug Discovery Today*, vol. 26, no. 1, pp. 173–180, 2021. DOI: 10.1016/j.drudis.2020.10.002.
- [10] M. L. Bonet, A. Palou, "Regulation of gene expression," in *Principles of Nutrigenetics and Nutrigenomics*, Elsevier, 2020, pp. 17–25. DOI: 10.1016/b978-0-12-804572-5.00003-3.
- [11] D. P. Clark, N. J. Pazdernik, M. R. McGehee, "Analysis of gene expression," in *Molecular Biology*, Elsevier, 2019, pp. 654–690. DOI: 10.1016/b978-0-12-813288-3.00021-5.

- [12] M. E. Futschik, M. Morkel, R. Schäfer, C. Sers, “The human transcriptome: Implications for understanding, diagnosing, and treating human disease,” in *Essential Concepts in Molecular Pathology*, Elsevier, 2020, pp. 113–138. DOI: 10.1016/b978-0-12-813257-9.00007-3.
- [13] E. D. Au, M. H. Farkas, “RNA sequencing and transcriptome analysis,” in *Genetics and Genomics of Eye Disease*, Elsevier, 2020, pp. 41–53. DOI: 10.1016/b978-0-12-816222-4.00004-6.
- [14] J.-H. Hung, Z. Weng, “Analysis of microarray and rna-seq expression profiling data,” *Cold Spring Harbor Protocols*, vol. 2017, no. 3, pdb-top093104, 2017. DOI: 10.1101/pdb.top093104.
- [15] J. De Las Rivas, S. Aibar, B. Roson, “Gene expression analysis and profiling of microarrays data and RNA-sequencing data,” in *Comprehensive Analytical Chemistry*, vol. 63, Elsevier, 2014, pp. 355–384.
- [16] R. L. Stears, T. Martinsky, M. Schena, “Trends in microarray analysis,” *Nature Medicine*, vol. 9, no. 1, pp. 140–145, 2003. DOI: 10.1038/nm0103-140.
- [17] S. Behjati, P. S. Tarpey, “What is next generation sequencing?” *Archives of Disease in Childhood-Education and Practice*, vol. 98, no. 6, pp. 236–238, 2013. DOI: 10.1136/archdischild-2013-304340.
- [18] S. Verma, R. K. Gazara, “Next-generation sequencing: An expedition from workstation to clinical applications,” in *Translational Bioinformatics in Healthcare and Medicine*, Elsevier, 2021, pp. 29–47. DOI: 10.1016/b978-0-323-89824-9.00003-3.
- [19] J. B. Wolf, “Principles of transcriptome analysis and gene expression quantification: An RNA-seq tutorial,” *Molecular Ecology Resources*, vol. 13, no. 4, pp. 559–572, 2013. DOI: 10.1111/1755-0998.12109.
- [20] K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter, F. Stahl, “Transcriptome analysis using next-generation sequencing,” *Current Opinion in Biotechnology*, vol. 24, no. 1, pp. 22–30, 2013. DOI: 10.1016/j.copbio.2012.09.004.
- [21] C. Børsting, N. Morling, “Next generation sequencing and its applications in forensic genetics,” *Forensic Science International: Genetics*, vol. 18, pp. 78–89, 2015. DOI: 10.1016/j.fsigen.2015.02.002.
- [22] C. Di Resta, S. Galbiati, P. Carrera, M. Ferrari, “Next-generation sequencing approach for the diagnosis of human diseases: Open challenges and new opportunities,” *EJIFCC*, vol. 29, no. 1, p. 4, 2018.
- [23] R. Hrdlickova, M. Toloue, B. Tian, “RNA-seq methods for transcriptome analysis,” *Wiley Interdisciplinary Reviews: RNA*, vol. 8, no. 1, e1364, 2017. DOI: 10.1002/wrna.1364.
- [24] R. Stark, M. Grzelak, J. Hadfield, “RNA sequencing: The teenage years,” *Nature Reviews Genetics*, vol. 20, no. 11, pp. 631–656, 2019. DOI: 10.1038/s41576-019-0150-2.
- [25] F. Ozsolak, P. M. Milos, “RNA sequencing: Advances, challenges and opportunities,” *Nature Reviews Genetics*, vol. 12, no. 2, pp. 87–98, 2011. DOI: 10.1038/nrg2934.

- [26] D. S. Latchman, "Transcription factors: An overview," *The International Journal of Biochemistry & Cell Biology*, vol. 29, no. 12, pp. 1305–1312, 1997.
- [27] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, "A census of human transcription factors: Function, expression and evolution," *Nature Reviews Genetics*, vol. 10, no. 4, pp. 252–263, 2009.
- [28] H. Hu, Q. Zhang, F-F. Hu, C.-J. Liu, A.-Y. Guo, "A comprehensive survey for human transcription factors on expression, regulation, interaction, phenotype and cancer survival," *Briefings in Bioinformatics*, vol. 22, no. 5, 2021. DOI: <https://doi.org/10.1093/bib/bbab002>.
- [29] S. A. Lambert *et al.*, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018. DOI: 10.1016/j.cell.2018.01.029.
- [30] S. B. Carroll, "Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution," *Cell*, vol. 134, no. 1, pp. 25–36, 2008.
- [31] M. T. Weirauch, T. R. Hughes, "Conserved expression without conserved regulatory sequence: The more things change, the more they stay the same," *Trends in Genetics*, vol. 26, no. 2, pp. 66–74, 2010.
- [32] E. Hernández-Lemus, H. Tovar, "Networks of transcription factors," in *Genome Plasticity in Health and Disease*, Elsevier, 2020, pp. 137–155. DOI: 10.1016/b978-0-12-817819-5.00009-7.
- [33] D. H. Gonzalez, "Introduction to transcription factor structure and function," in *Plant Transcription Factors*, Elsevier, 2016, pp. 3–11. DOI: 10.1016/b978-0-12-800854-6.00001-4.
- [34] V. Matys *et al.*, "TRANSFAC: Transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003. DOI: 10.1093/nar/gkg108.
- [35] A. Khan *et al.*, "JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework," *Nucleic Acids Research*, vol. 46, no. D1, pp. D260–D266, 2018. DOI: 10.1093/nar/gkx1126.
- [36] M. A. Hume, L. A. Barrera, S. S. Gisselbrecht, M. L. Bulyk, "UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein–dna interactions," *Nucleic Acids Research*, vol. 43, no. D1, pp. D117–D122, 2015.
- [37] N. A. Kolchanov *et al.*, "Transcription regulatory regions database (TRRD): Its status in 2002," *Nucleic Acids Research*, vol. 30, no. 1, pp. 312–317, 2002.
- [38] D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, Y. Gilad, "The functional consequences of variation in transcription factor binding," *PLoS Genetics*, vol. 10, no. 3, e1004226, 2014.
- [39] C. J. Banks, A. Joshi, T. Michoel, "Functional transcription factor target discovery via compendia of binding and expression profiles," *Scientific Reports*, vol. 6, no. 1, pp. 1–10, 2016.
- [40] M. Maienschein-Cline, J. Zhou, K. P. White, R. Sciammas, A. R. Dinner, "Discovering transcription factor regulatory targets using gene expression and binding data," *Bioinformatics*, vol. 28, no. 2, pp. 206–213, 2012.

- [41] G. R. Lee, C. G. Spilianakis, R. A. Flavell, "Hypersensitive site 7 of the th 2 locus control region is essential for expressing th 2 cytokine genes and for long-range intrachromosomal interactions," *Nature Immunology*, vol. 6, no. 1, pp. 42–48, 2005.
- [42] S. R. Kulkarni, K. Vandepoele, "Inference of plant gene regulatory networks using data-driven methods: A practical overview," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1863, no. 6, p. 194447, 2020. DOI: 10.1016/j.bbagrm.2019.194447.
- [43] M. N. McCall, "Estimation of gene regulatory networks," *Postdoc journal: a journal of postdoctoral research and postdoctoral affairs*, vol. 1, no. 1, p. 60, 2013.
- [44] V. A. Huynh-Thu, G. Sanguinetti, "Gene regulatory network inference: An introductory survey," in *Gene Regulatory Networks*, Springer, 2019, pp. 1–23.
- [45] F. Emmert-Streib, M. Dehmer, B. Haibe-Kains, "Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks," *Frontiers in Cell and Developmental Biology*, vol. 2, p. 38, 2014. DOI: 10.3389/fcell.2014.00038.
- [46] K. Mochida, S. Koda, K. Inoue, R. Nishii, "Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets," *Frontiers in Plant Science*, vol. 9, p. 1770, 2018. DOI: 10.3389/fpls.2018.01770.
- [47] M. M. Saint-Antoine, A. Singh, "Network inference in systems biology: Recent developments, challenges, and applications," *Current Opinion in Biotechnology*, vol. 63, pp. 89–98, 2020. DOI: 10.1016/j.copbio.2019.12.002.
- [48] F. M. Delgado, F. Gómez-Vela, "Computational methods for gene regulatory networks reconstruction and analysis: A review," *Artificial Intelligence in Medicine*, vol. 95, pp. 133–145, 2019. DOI: 10.1016/j.artmed.2018.10.006.
- [49] W. Yan, W. Xue, J. Chen, G. Hu, "Biological networks for cancer candidate biomarkers discovery," *Cancer Informatics*, vol. 15, CIN-S39458, 2016. DOI: 10.4137/cin.s39458.
- [50] S. Barbosa, B. Niebel, S. Wolf, K. Mauch, R. Takors, "A guide to gene regulatory network inference for obtaining predictive solutions: Underlying assumptions and fundamental biological and data constraints," *Biosystems*, vol. 174, pp. 37–48, 2018. DOI: 10.1016/j.biosystems.2018.10.008.
- [51] M. Banf, S. Y. Rhee, "Computational inference of gene regulatory networks: Approaches, limitations and opportunities," *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, vol. 1860, no. 1, pp. 41–52, 2017. DOI: 10.1016/j.bbagrm.2016.09.003.
- [52] M. S. Kwon, B. T. Lee, S. Y. Lee, H. U. Kim, "Modeling regulatory networks using machine learning for systems metabolic engineering," *Current Opinion in Biotechnology*, vol. 65, pp. 163–170, 2020. DOI: 10.1016/j.copbio.2020.02.014.

- [53] D. Mercatelli, L. Scalambra, L. Triboli, F. Ray, F. M. Giorgi, “Gene regulatory network inference resources: A practical overview,” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1863, no. 6, p. 194430, 2020. DOI: 10.1016/j.bbagrm.2019.194430.
- [54] A. C. Babbie, M. P. Stumpf, T. Thorne, “Gene regulatory network inference,” in *Systems Medicine*, Elsevier, 2021, pp. 86–95. DOI: 10.1016/b978-0-12-801238-3.11346-7.
- [55] E. Clough, T. Barrett, “The gene expression omnibus database,” in *Statistical Genomics*, Springer, 2016, pp. 93–110.
- [56] S. A. Bini, “Artificial intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care?” *The Journal of Arthroplasty*, vol. 33, no. 8, pp. 2358–2361, 2018. DOI: 10.1016/j.arth.2018.02.067.
- [57] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, J. P. Campbell, “Introduction to machine learning, neural networks, and deep learning,” *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 14–14, 2020.
- [58] J. M. Helm *et al.*, “Machine learning and artificial intelligence: Definitions, applications, and future directions,” *Current Reviews in Musculoskeletal Medicine*, vol. 13, no. 1, pp. 69–76, 2020. DOI: 10.1007/s12178-020-09600-8.
- [59] T. Ching *et al.*, “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.
- [60] W. Jones, K. Alasoo, D. Fishman, L. Parts, “Computational biology: Deep learning,” *Emerging Topics in Life Sciences*, vol. 1, no. 3, pp. 257–274, 2017.
- [61] G. Muzio, L. O’Bray, K. Borgwardt, “Biological network analysis with deep learning,” *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1515–1530, 2021.
- [62] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, “Deep learning for computational biology,” *Molecular Systems Biology*, vol. 12, no. 7, p. 878, 2016.
- [63] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [64] G. Eraslan, Ž. Avsec, J. Gagneur, F. J. Theis, “Deep learning: New computational modelling techniques for genomics,” *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019.
- [65] C. H. Wu, J. W. McLarty, “Neural network basics,” in *Methods in Computational Biology and Biochemistry*, Elsevier, 2000, pp. 19–27. DOI: 10.1016/s1874-5938(00)80003-1.
- [66] S. Theodoridis, “Neural networks and deep learning,” in *Machine Learning*, Elsevier, 2020, pp. 901–1038. DOI: 10.1016/b978-0-12-818803-3.00030-1.
- [67] J. Zupan, “Basics of artificial neural networks,” in *Data Handling in Science and Technology*, Elsevier, 2003, pp. 199–229. DOI: 10.1016/s0922-3487(03)23007-0.

- [68] M. Kubat, “Artificial neural networks,” in *An Introduction to Machine Learning*, Springer, 2017, pp. 91–111. DOI: 10.1007/978-3-319-63913-0_5.
- [69] S. Min, B. Lee, S. Yoon, “Deep learning in bioinformatics,” *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 851–869, 2016. DOI: 10.1093/bib/bbw068.
- [70] M. Wainberg, D. Merico, A. Delong, B. J. Frey, “Deep learning in biomedicine,” *Nature Biotechnology*, vol. 36, no. 9, pp. 829–838, 2018. DOI: 10.1038/nbt.4233.
- [71] C. Liu, P. Cui, T. Huang, “Identification of cell cycle-regulated genes by convolutional neural network,” *Combinatorial Chemistry & High Throughput Screening*, vol. 20, no. 7, pp. 603–611, 2017. DOI: 10.2174/1386207320666170417144937.
- [72] Y. Yuan, Z. Bar-Joseph, “Deep learning for inferring gene relationships from single-cell expression data,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 52, pp. 27 151–27 158, 2019. DOI: 10.1073/pnas.1911536116.
- [73] Y. Fan, X. Ma, “Gene regulatory network inference using 3D convolutional neural network,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 99–106.
- [74] Y. Yang, Q. Fang, H.-B. Shen, “Predicting gene regulatory interactions based on spatial gene expression data and deep learning,” *PLOS Computational Biology*, vol. 15, no. 9, e1007324, 2019. DOI: 10.1371/journal.pcbi.1007324.
- [75] D. MacLean, “A convolutional neural network for predicting transcriptional regulators of genes in arabidopsis transcriptome data reveals classification based on positive regulatory interactions,” *bioRxiv*, Apr. 2019. DOI: 10.1101/618926.
- [76] H. Han *et al.*, “TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D380–D386, 2018. DOI: 10.1093/nar/gkx1013.
- [77] G. Consortium *et al.*, “The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans,” *Science*, vol. 348, no. 6235, pp. 648–660, 2015. DOI: 10.1126/science.1262110.
- [78] J. N. Weinstein *et al.*, “Cancer genome atlas research network,” *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.
- [79] J. Vivian *et al.*, “Toil enables reproducible, open source, big biomedical data analyses,” *Nature Biotechnology*, vol. 35, no. 4, pp. 314–316, 2017. DOI: 10.1038/nbt.3772.
- [80] D. Mardaoui, D. Garreau, “An analysis of LIME for text data,” in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, PMLR, vol. 130, 2021, pp. 3493–3501.
- [81] D. Garreau, U. Luxburg, “Explaining the explainer: A first theoretical analysis of LIME,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, PMLR, vol. 108, 2020, pp. 1287–1296.

- [82] I. Palatnik de Sousa, M. Maria Bernardes Rebuzzi Vellasco, E. Costa da Silva, “Local interpretable model-agnostic explanations for classification of lymph node metastases,” *Sensors*, vol. 19, no. 13, p. 2969, 2019. DOI: 10 . 3390 / s19132969.
- [83] D. Garreau, D. Mardaoui, “What does LIME really see in images?” In *Proceedings of the 38th International Conference on Machine Learning*, M. Meila, T. Zhang, Eds., ser. Proceedings of Machine Learning Research, PMLR, vol. 139, 2021, pp. 3620–3629.



Konferans Bildirisi

1. G. Çelen, A. Yılmaz, "Prediction of regulatory network interactions with CNN model using human RNA-Seq data" The 14th International Symposium on Health Informatics and Bioinformatics, 2021, p. 69.

