T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GEN AĞI ÇIKARIMI İÇİN PROTEOMİK VE GEN İFADE VERİLERİNİN ENTEGRASYONUNDA İLİŞKİ TAHMİNCİLERİN ETKİSİ

CİHAT ERDOĞAN

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
PROF. DR. BANU DİRİ

İSTANBUL, 2018

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

GEN AĞI ÇIKARIMI İÇİN PROTEOMİK VE GEN İFADE VERİLERİNİN ENTEGRASYONUNDA İLİŞKİ TAHMİNCİLERİN ETKİSİ

Cihat ERDOĞAN tarafından hazırlanan tez çalışması 11.06.2018 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Prof.Dr. Banu DİRİ

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Prof.Dr. Banu DİRİ

Yıldız Teknik Üniversitesi

Prof.Dr. Nizamettin AYDIN

Yıldız Teknik Üniversitesi

Doç.Dr. Erdinç UZUN

Namık Kemal Üniversitesi

Dr.Öğr.Üyesi Bülent BOLAT

Yıldız Teknik Üniversitesi

Dr.Öğr.Üyesi Tuğba İLDENİZ

Acıbadem Üniversitesi

ÖNSÖZ

Bu tezin gerçekleştirilmesinde değerli vaktini ayırıp, bana faydalı olabilmek için elinden geleni fazlasıyla sunan, samimiyetini ve güler yüzünü hiç esirgemeyen, gelecekteki mesleki hayatımda verdiği değerli bilgilerden faydalanacağım, değerli danışman hocam Prof. Dr. Banu DİRİ'ye,

Ayrıca, çalışmamda konu, kaynak ve yöntem açısından fikir vererek yol gösteren, kendisine ne zaman danışsam kıymetli vaktini ayıran, bu alandaki derin bilgilerini paylaşan değerli Dr. Öğr. Üyesi Zeyneb KURT'a,

Yine, tez çalışmam boyunca değerli yorum ve görüşleriyle bana destek veren hocalarım Prof. Dr. Nizamettin AYDIN'a ve Doç. Dr. Erdinç UZUN'a,

Çalışmamı değerlendiren ve geliştirmem için önerilerde bulunan tez savunma sınavı jüri üyeleri Dr. Öğr. Üyesi Bülent BOLAT'a ve Dr. Öğr. Üyesi Tuğba İLDENİZ'e,

Yine tüm bu süreçte yardımlarını esirgemeyen Dr. Öğr. Üyesi Selçuk POYRAZ'a,

Son olarak bu çalışma süresince, tüm zorlukları birlikte göğüslediğim ve hayatımın her anında bana inanarak desteklerini esirgemeyen kıymetli eşim Merve ERDOĞAN'a ve çok sevdiğim, sevinç kaynağım canım oğlum Kerem'e sonsuz teşekkürlerimi sunarım.

Haziran, 2018

Cihat ERDOĞAN

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

τ	Kendall tau korelasyon katsayısı
r	Pearson korelasyon katsayısı

KISALTMA LİSTESİ

ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
AUPR	Area under precision-recall curve
AUROC	Area under receiver operating characteristics
BS	B-spline
C3NET	The Conservative Causal Core Network
CD	Copula Dönüşümü
CLR	Context-Likelihood of Relatedness
CUIs	Concept Unique Identifiers
ÇYT	Çekirdek Yoğunluk Tahmincisi
DepEst	An R package of important dependency estimators for gene network inference algorithms
DisGeNET	A comprehensive platform integrating information on human disease-associated genes and variants
EF	Eşit frekans
EG	Eşit genişlik
FDR-q	False Discovery Rate
FKT	Fisher'ın kesinlik testi
FP	Yanlış Pozitif
GAÇ	Gen Ağı Çıkarımı
GAI	Gene Network Inference
GDC	Genomic Data Commons
GTEx	Genotype-Tissue Expression
ICGC	International Cancer Genome Consortium
KB	Karşılıklı Bilgi
KEG	Küresel eşit ağırlık
Kendall KKD	Kendall Tau Korelasyon Katsayı Değeri
MI	Mutual Information
MINET	An open source R/Bioconductor Package for Mutual Information based Network Inference
MM	Miller-Madow
MRNET	Network inference with maximum relevance/minimum redundancy feature selection
MSigDB	Molecular Signature Database

NCBI	National Center for Biotechnology Information
PC	Pathway Commons
Pearson KKD	Pearson korelasyon katsayı değeri
PMID	The number of supporting publications
RELNET	Relevance Network
SG	Schurmann-Grassberger
Shrink	James-Stein Shrinkage
Spearman KKD	Spearman korelasyon katsayı değeri
SynTReN	A generator of synthetic gene expression data for design and analysis of structure learning algorithms
TARGET	Therapeutically Available Research to Generate Effective Treatments
TCGA	The Cancer Genome Atlas
TCPA	The Cancer Proteome Atlas
TOIL	Scalable and Efficient Workflow Engine
TP	Doğru pozitif
UMLS	Unified Medical Language System
WGCNA	Weighted Correlation Network Analysis

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

GEN AĞI ÇIKARIMI İÇİN PROTEOMİK VE GEN İFADE VERİLERİNİN ENTEGRASYONUNDA İLİŞKİ TAHMİNCİLERİN ETKİSİ

Cihat ERDOĞAN

Bilgisayar Mühendisliği Anabilim Dalı

Doktora Tezi

Tez Danışmanı: Prof. Dr. Banu DİRİ

Bu tez çalışmasında gen ağı çıkarım yöntemleri üzerinde önemli etkiye sahip olan ve moleküler etkileşimleri belirlemek için kullanılan ilişki tahmincilerinin, farklı biyolojik veri türlerinin entegrasyonu üzerindeki etkisi incelenmiştir. Tezde incelenen tüm kanser türleri için gen ifade ve proteomik verileri The Cancer Proteome Atlas (TCPA) tarafından sağlanmıştır. Öncelikle korelasyon tabanlı ilişki tahmincilerin etkisi, literatürde sıklıkla kullanılan Gen Ağı Çıkarım (GAÇ) yöntemleri kullanılarak, on altı farklı kanser türüne ait proteomik veriler analiz edilerek incelenmiştir. Ardından, Amerikan Kanser Topluluğu verilerine göre yaygın olarak görülen beş farklı kanser türüne ait proteomik verileri kullanılarak, hastalıkla ilişkili gen-gen/protein-protein etkileşim alt ağlarındaki merkez genler/proteinler tespit edilmeye çalışılmıştır. Bu işlem sırasında literatürde sıklıkla kullanılan karşılıklı bilgi (KB) ve korelasyon tabanlı dokuz ilişki kestirimci karşılaştırılmıştır. İlişki tahmincilerinin performansını ölçmek için altın standart olarak, Hastalık-Gen ilişkileri entegrasyon platformu (DisGeNET) ve Moleküler İmzalar Veritabanı (MSigDB) kullanılmıştır. Oluşturulan ortak ifade ağları ile hastalıkla ilişkili yollar karşılaştırılmış ve ilişki tahmincilerinin performansını değerlendirmek için Fisher'ın kesinlik testi kullanılmıştır. Ağırlıklı korelasyon ağ analizinde (WGCNA) düzenleyici ağların tahmini için kullanılan Spearman ve Pearson korelasyon yaklaşımlarına göre, KB tabanlı ilişki tahmincilerinin başarımının daha yüksek olduğu gözlenmiştir. Korelasyon tabanlı yöntemlerde beş kanser türü için en iyi ortalama başarı oranı %60 iken, KB tabanlı yöntemlerde ortalama başarı oranı James-Stein Shrinkage (Shrink) için %71, Schurmann-Grassberger (SG) için %64'tür. Sonrasında gen ifade ve proteomik verilerinden

çıkarımlanmış ağların entegrasyonu sağlanmıştır. Son olarak her bir kanser türüne göre merkez genler ve çıkarımlanmış alt ağlar, araştırmacıların ve biyologların incelemesi için sunulmuştur.

Anahtar Kelimeler: İlişki tahmincileri, gen ağı çıkarımı, proteomik, kanser

ABSTRACT

THE ASSOCIATION ESTIMATORS' EFFECT ON THE INTEGRATION OF PROTEOMIC AND GENE EXPRESSION DATA FOR GENE NETWORK INFERENCE

Cihat ERDOĞAN

Department of Computer Engineering

PhD. Thesis

Adviser: Prof. Dr. Banu DİRİ

In this thesis, the effects of association estimators, which have a significant influence on gene network inference methods and used to determine molecular interactions, on the integration of different biological data types were examined. Gene expression and proteomic data for all cancer types used in this thesis were provided from The Cancer Proteome Atlas (TCPA). Firstly, the effect of the correlation-based association estimators on the analysis of proteomic data from sixteen different cancer types was examined by using Gene Network Inference (GAI) methods that are frequently used in the literature. Furthermore, attempts were made to detect the hub genes/proteins in the gene-gene/protein-protein interaction subnetworks associated with the disease by using proteomic data from five different cancer types, which are commonly seen according to American Cancer Society data. During this process, the mutual information (MI) and correlation based nine association estimators, which are commonly used in the literature, were compared. The disease-gene association integration platform (DisGeNET) and the Molecular Signature Database (MSigDB) were used as the gold standard for measuring the performance of the association estimators. The disease-associated pathways were compared with the as-generated co-expression networks and the Fisher's exact test was used to assess the association estimators' performance. Based on the Spearman and Pearson correlation approaches used for the estimation of

regulatory networks in the weighted correlation network analysis (WGCNA), the MI-based association estimators' performance was observed to be higher. The best average success rate for five cancer types is 60% for the correlation-based methods, while for the MI-based methods it is 71% for James-Stein Shrinkage (Shrink), and 64% for Schurmann-Grassberger (SG). Integration of the inferred networks was then conducted by using the gene expression and proteomic data. Finally, for each cancer type, hub genes and inferred subnets are presented for the investigations of researchers and biologists.

Keywords: Association estimators, network inference, proteomic, cancer

BÖLÜM 1

GİRİŞ

Biyoinformatikte kanser, otizm, diyabet, alzaymır (alzheimer) gibi karmaşık yapıya sahip hastalıkların oluşumunda hücreler veya dokular üzerinde gözlenen moleküler değişimleri anlamak için çeşitli yaklaşımlar kullanılır. Bu yaklaşımlardan yaygın olarak kullanılan yöntemlerin başında ağ çıkarım algoritmaları gelmektedir. Ağ çıkarım algoritmaları hücre fizyolojisini ve hastalık patogenezini anlamak ve proteinlerin, genlerin genom çapındaki çalışma mekanizmasını tahmin etmek için bir ifade veri setinden protein-protein/gen-gen ilişkilerinin tespitini yapmak için korelasyon, regresyon, Karşılıklı Bilgi (KB) gibi farklı istatistiksel yöntemler ve sınıflandırma, kümeleme gibi çeşitli makine öğrenmesi yöntemleri kullanmaktadır [1], [2].

Bu tez çalışmasında ilk olarak literatürde sıklıkla kullanılan ağ çıkarım algoritmaları ve bu algoritmalarda kullanılan ilişki tahmincileri incelenmiştir. İncelenen ilişki tahmincileri Bölüm 2'de ve incelenen ağ çıkarım yöntemleri de Bölüm 3'te verilmiştir. Korelasyon tabanlı ilişki tahmincilerin proteomik verilerin analizindeki başarımlarını incelemek için literatürde yaygın olarak kullanılan dört GAÇ algoritması kullanılmıştır. Bu incelemenin ardından literatürde sıklıkla kullanılan ağ çıkarım algoritmalarından olan WGCNA'nın en önemli adımlarından biri olan ilişki tahmin adımı yerine korelasyon ve KB tabanlı ilişki tahmincilerinin entegrasyonu sağlanarak, ilişki tahmincilerinin yöntemin başarımı üzerindeki etkisi incelenmiştir. Bunun için literatürdeki çalışmalardan farklı olarak sentetik veya gerçek gen ifade verileri yerine gerçek proteomik verileri (proteomic) kullanılarak analiz işlemleri gerçekleştirilmiştir. Daha önce benzer bir çalışmayı yapan bir araştırmaya literatürde rastlanmamıştır. Sonrasında en iyi sonucu veren yöntem için gen

ifade ve proteomik verileri kullanılarak çıkarımlanan ağların entegrasyonu sağlanmış ve bulunan önemli sonuçlar çalışmamızda verilmiştir.

1.1 Literatür Özeti

Yapılan araştırmalarda, mikro dizi tahlillerinden elde edilen hem sentetik hem de gerçek gen ifade veri setleri kullanılarak ilişki tahmincilerinin gen ağı çıkarım tekniklerine olan etkisi incelenmiştir.

Olsen vd. [3] Pearson korelasyon katsayı değeri (Pearson KKD), Spearman korelasyon katsayı değeri (Spearman KKD), Ampirik (Empirical), Miller-Madow (MM) ve Shrink tahmincilerinin performansını sentetik ve gerçek mikro dizi veri setlerini kullanarak üç farklı ağ çıkarım algoritması üzerinde değerlendirmişlerdir. Farklı özelliklerde 12 sentetik veri kümesini SynTReN [4] simülatörünü kullanarak oluşturmuşlar ve veri setlerindeki gen sayılarını 100, 200, 300 olarak, örnek sayıları da 50, 100, 200, 300 olarak belirlemişlerdir [3]. Ayrıca, bazı veri setlerine gürültülü veri ekleyerek, gürültülü veri kullanmanın analiz sonuçlarına etkisini de incelemişlerdir [3]. Çalışmada kullanılan algoritmalar ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [5], CLR (Context-Likelihood of Relatedness) [6] ve MRNET (Network inference with Maximum Relevance/Minimum redundancy feature selection) [7] ağ çıkarım yöntemleridir. Ayrıca, detayları Bölüm 2.3'te verilen eşit genişlik (EG) ve eşit frekans (EF) ayrıklaştırma yöntemlerinin ilişkilendirme tahmincilerinin performansı üzerindeki etkilerini incelemek için Ampirik, MM ve Shrink kestirimcilerini kullanmışlardır [3]. MRNET ve Spearman KKD ikilisinin gürültüsüz veri kümelerinde CLR ve Pearson KKD ikilisinden daha başarılı olduğunu ve MRNET ve Spearman KKD ikilisinin daha düşük sapmaya (bias) sahip olduğunu belirtmişlerdir. CLR ve Pearson KKD ikilisinin ise gürültülü veride daha düşük varyanta (variant) sahip olduğu için doğruluğu (robust) daha yüksek sonuçlar verdiğini bildirmişlerdir [3]. Yapay veri kümeleri kullanılarak üretilen deney verilerinin analizinde korelasyon temelli ilişki tahmincileri olan Pearson ve Spearman KKD yöntemlerinin en yüksek başarım sonuçlarını verdiğini gözlemlemişler ve bu sebepten dolayı rastgele değişkenler arasındaki ilişkilerin non-lineer olma ve monoton olmama durumlarının göz ardı edilebileceğini öne sürmüşlerdir. Dolayısıyla, başarılı bir ilişki tahmini için lineer ve monoton ilişkileri hesaba katmanın yeterli olacağını iddia

etmişlerdir [3]. Analiz sonuçlarına göre, gürültüsüz veri kümesi için ilişki tahmincileri en başarılıdan başarısıza doğru performans sıralaması MM > Ampirik > Spearman KKD > Pearson KKD > Shrink şeklinde; gürültülü veri kümesi için başarım sıralaması: Spearman KKD > Pearson KKD > MM > Ampirik > Shrink şeklinde olmuştur [3]. Sonuç olarak, Spearman KKD ile MRNET'in ve Pearson KKD ile CLR'ın, sentetik veri kümesinde daha başarılı sonuçlar verdiğini ve gerçek veri kümesinde de önemli sonuçlar elde edildiğini bildirmişlerdir [3]. Araştırmacılar bu çalışmadaki analiz işlemlerini gerçekleştirmek için MINET [8] R paketini kullanmışlardır. Biz de çalışmamızda MINET R paketini proteomik verilerin analizinde kullandık.

Simoes ve Streib [9], MM, Ampirik, Shrink ve SG ilişki tahmincilerini, C3NET (Conservative Causal Core NETwork) [10] ağ çıkarım algoritması ile birlikte üç farklı ayrıklaştırma yöntemiyle (EG, EF, KEG - Küresel eşit genişlik - global equal with), hem tahmincilerin hem de ayrıklaştırma yöntemlerinin ağ çıkarım algoritmalarındaki etkilerini sentetik gen mikro dizi veri setlerini kullanarak incelemişlerdir. Çalışmalarında ayrıklaştırma yöntemlerinin ilişki tahmincilerine nazaran gen ağı çıkarımında daha etkin olduğunu savunmuşlardır. Farklı örnek sayısına sahip 5 sentetik veri kümesini SynTReN [4] simülatörünü kullanarak oluşturmuşlar ve veri setlerindeki gen sayısını 100 olarak, örnek sayılarını da 50, 100, 200, 500 ve 1.000 olarak belirlemişlerdir [9]. Fakat [3]'ten farklı olarak veri setlerine gürültü eklememişlerdir. Analiz sonuçlarına göre, C3NET GAÇ yöntemiyle çıkarımlanan ağlarda EG ve KEG ayrıklaştırma yöntemlerinin EF ayrıklaştırma yöntemine göre daha başarılı olduğunu ve örnek sayısı arttıkça başarım skorlarının da arttığını bildirmişlerdir [9]. Ayrıca, ilişki tahmincilerinin başarım sıralamasının *MM > Ampirik > SG > Shrink* şeklinde olduğunu bildirmişler ve sonuç olarak ta MM tahmincisinin, EG ve KEG ayrıklaştırma yöntemleriyle kullanıldığında diğer kestiricilerden daha iyi performans gösterdiğini bulmuşlardır [9].

Çalışma [3]'te ve [9]'da kullanılan Ampirik ve MM yöntemleri ile KB ilişki tahminini yapılırken deneysel entropi kullanılmış ve yapıları gereği bu işlem öncesinde Bölüm 2.3'te verilen ayrıklaştırma işlemi gerçekleştirilmiştir. Dolayısıyla, veri kümesindeki her bir hücreye ait olasılık dağılımını bulmak için hücrelerdeki veri örneklerinin sayısı kullanılmıştır. MM ilişki tahmincisi Ampirik yöntemden farklı olarak ilişki tahmin skorunu hesaplarken sapmayı da (bias) hesaba katmakta ve bu sayede sapmadan kaynaklanan

hatayı giderebilmektedir. Shrink ve SG ilişki tahmincileri ise, MM ve Amprik yöntemlerden farklı olarak, her bir hücredeki verilerin olasılık dağılımını hesaplar. Sonrasında bu olasılık dağılımları kullanılarak da bireysel ve birleşik entropiden KB tabanlı ilişki tahmin skorları hesaplanmaktadır. Bahsedilen bu dört ilişki tahmincisinin hesaplanmasında kullanılan bağıntılar detaylı bir şekilde Bölüm 2'de verilmiştir.

Daub vd. [11], B-spline (BS) ve Çekirdek Yoğunluk Tahmincisi (ÇYT, Kernel Density Estimator) yöntemlerinin büyük ölçekli gen ifade veri setleri üzerindeki performanslarını araştırmışlardır. Çalışmada iki veri seti kullanılmış olup, bunlardan birincisinde 300 örneğe ait 5.345 gen ifade verisi ve diğerinde 102 örneğe ait 22.608 gen ifade verisi bulunmaktadır. Sonuç olarak, BS tahmincisinin performansının farklı eğri (spline) değerleri için değişkenlik gösterdiğini bulmuşlardır [11]. Başarı sırası şöyledir: BS (spline sırası = 3) > ÇYT > BS (spline sırası = 1). Ayrıca, ÇYT'nin işlem karmaşıklığının BS'ye göre 10^4 kat daha büyük olduğunu belirtmişlerdir [11]. İncelediğimiz ver seti çalışma [11] göre daha küçük olduğundan tezimizde ÇYT yönteminin proteomik veriler üzerindeki başarımı da değerlendirilmiştir.

Kurt vd. [12], 14 farklı ilişki tahmincisini 3 farklı ağ çıkarım algoritması (RELNET [13], ARACNE [5], C3NET [10]) ile değerlendirmek için iki sentetik ve iki gerçek biyolojik veri seti kullanmışlardır. Çalışmalarında SynTReN [4] simülatörünü kullanarak oluşturmuş oldukları sentetik veri setlerinin birincisinde 100 gen ve 100 örnek, ikincisinde 100 gen ve 1.000 örnek bulunmaktadır. Gerçek biyolojik veri setlerinden birincisinde *E.coli* bakterisine ait 1.146 gen ve 524 örnek, ikincisinde *S.cerevisiae* maya hücresine ait 2.760 gen ve 247 örnek bulunmaktadır. Ayrıca, çalışmalarında Copula Dönüşümü (CD) ön işleminin ilişki tahmincilerinin performansı üzerindeki etkilerini de incelemişlerdir. B-spline, Pearson tabanlı Gauss ve Spearman tabanlı Gauss tahmincileri, tüm yöntemlere göre en iyi performans gösteren tahminciler olarak gözlemlemişlerdir. Ayrıca, CT ön işleminin, sentetik veri setleri için tahmincilerin ilişki tahmin performanslarını arttırdığını da vurgulamışlardır [12].

Ish-Horowicz ve Reid [14] çalışmalarında 9 farklı ilişki tahmincisinin başarımını 3 farklı ağ çıkarım algoritması (ARACNE [5], CLR [6], MRNET [8]) üzerinde incelemişlerdir. Analiz işlemleri için bir sentetik ve iki gerçek veri seti kullanmışlardır. GeneNetWeaver [15]

simülatörü ile oluşturdukları sentetik veri setinde 1.643 gen ve 805 örnek, gerçek veri setinin ilkinde *E.coli* bakterisine ait 4.297 gen ve 805 örnek, ikincisinde *S.cerevisia*e maya hücresine ait 5.950 gen ve 593 örnek bulunmaktadır. Analiz sonucunda BS tahmincisinin sürekli olarak gerçek ve sentetik veri kümeleri üzerinde iyi performans gösterdiğini, CLR ağ çıkarım yönteminin de en başarılı sonuçları verdiğini bildirmişlerdir [14].

Bu tezde [3], [9], [11], [12], [14]'de kullanılan 9 farklı ilişki tahmincisinin proteomik verilerin analizindeki başarımları incelenmiştir.

Tek başına gen ifade verilerinin karmaşık yapıdaki biyolojik sistemlerin tespitinde yeterli olmadığı ve farklı biyolojik veri türlerinin entegrasyonu ile bu yetersizliğin giderilmeye çalışıldığı birçok çalışmada değinilmiştir [16], [17]. Onun içindir ki çoklu -omik verilerin bütünleştirici analizi için farklı yöntemler geliştirilerek moleküler sistemlerin etkinliğinin daha eksiksiz ve doğru bir resmi çizilmeye çalışılmıştır [18].

Kong vd. [16], alzaymır hastalığını tetikleyen sinyal yolaklarının (signaling pathway) ve düzenleyici ağların tespiti için gen ifade ve protein-protein ilişkilerinin entegrasyonunu sağlamışlardır. Bunun için [19]'de önerilen tamsayı doğrusal programlama (integer linear programming) yöntemini ve doğrusal regresyon modelini (linear regression model) kullanmışlardır. İlgili çalışmada Ulusal Biyoteknoloji Bilgi Merkezinden alınan (National Center for Biotechnology Information - NCBI) 7 hastaya ait 22.283 gen ifade verisi ile BioGRID'den alınan 12.466 proteine ait 40.323 protein-protein ilişki verisi kullanılmıştır. Sonuç olarak alzaymır hastalığına etki eden 6 adet önemli genin tespitini gerçekleştirmeyi başarmışlardır.

Shi vd. [17], kanser gibi karmaşık yapıya sahip hastalıklara ilişkin alt ağları tespit etmek için gen ifade verileriyle protein-protein ilişki verilerinin entegrasyonunu sağlamışlardır. Entegrasyon işlemi için Markof Rastgele Alanını (Markov Random Field) KB ilişki tahmin yöntemini bir araya getiren bir yöntem önermişlerdir. Yöntemlerinin başarımını değerlendirmek için öncelikle SynTReN simülatörü ile elde ettikleri veri setini kullanmışlardır. Ardından da göğüs kanserine ait gerçek gen ifade verisi ve protein-protein ilişki verisi kullanarak önerdikleri yöntemi test etmişlerdir. Sonuç olarak göğüs kanserinde önemli etkiye sahip merkez genleri başarılı bir şekilde tahmin ettiklerini bildirmişlerdir [17].

Ağ çıkarım yöntemleri, farklı kanser türleri üzerine yapılan çok sayıda araştırmada kullanılmıştır. Şenbabaoğlu vd. [20] TCPA'dan alınan 3467 hasta ve 11 farklı kanser türüne ait proteomik verileri, 13 farklı ağ çıkarım metodu kullanarak analiz etmişlerdir ve sonuç olarak ağ çıkarım yöntemlerinin başarımının kanser türüne göre değiştiğini bildirmişlerdir. Madhamshettiwar vd. [21], 7 farklı ağ çıkarım metodu kullanarak yumurtalık kanserinin gelişimindeki biyolojik mekanizmaları araştırmışlardır. WGCNA [22] en sık kullanılan gen ağı çıkarımı ve kümeleme yaklaşımlarından biridir ve verilen bir hastalıkla ilgili yüksek oranda korelasyona sahip gen/protein modülleri (küme, alt ağ) ve yüksek bağlana bilirliğe sahip merkez genleri/proteinleri bulmak için kullanılır. WGCNA, beyin kanseri [23], maya hücre döngüsü [24], fare genetiği [25], [26], diyabet [27] ve diğer pek çok hastalık ve karmaşık özelliklerden gelen gen/protein ekspresyon verilerini analiz etmek için kullanılmıştır.

Bu tez çalışmasında, TCPA [28] tarafından üretilen farklı kanser türlerinin proteomik ve gen ifade verileri aracılığıyla, biyoinformatik çalışmalarda belirli bir hastalıkla ilgili moleküler yapıları saptamak için sıkça kullanılan ağ çıkarım algoritmalarındaki ilişkilendirme tahmincilerinin etkileri değerlendirilmiştir.

1.2 Tezin Amacı

Biyolojik sistemlerin karmaşıklığı, teknolojik limitleri, çok sayıda biyolojik değişken ve nispeten düşük sayıda biyolojik numune, çoklu omik veri kümelerinin analizini önemli bir problem haline getirmiştir.

Çalışmamızda gen ifade verileri kullanılarak kestirilen gen ağlarıyla, proteomik verileri kullanılarak kestirilen protein ağlarının entegrasyonunda ilişki tahmincilerin etkisinin araştırılması yapılacak olup, daha doğru sonuçların elde edilebildiği moleküler ağların kestirilmesinin sağlanması amaçlanmaktadır. Bu çalışmada, kanser gibi hastalıklı hücrelerden elde edilen mikro dizin gen ve proteomik veri kümeleri kullanılarak, kansere sebep olabilecek moleküler etkileşimlerin daha doğru tespit edilebildiği ağ kestirimlerinin yapılmasına katkı sağlanacaktır. Ayrıca, ilişki tahmincilerin farklı veri türlerinin analizi üzerindeki etkisi de araştırılacaktır. Bu sayede doğruluğu daha yüksek biyolojik ağlar üretilerek, genlerin veya proteinlerin hastalıklar üzerindeki etkisini

araştıran biyologların, hastalıkla ilişkili olma durumu daha yüksek olan genlere veya proteinlere odaklanmasını sağlamayı amaçlamaktayız.

Bu tez çalışmasında genlerden üretilen genomik verileri kullanılarak oluşturulan ağlarla, proteinlerden üretilen proteomik verileri kullanılarak oluşturulan ağların entegrasyonunu sağlamak için bir yöntem önerilmiştir. Geliştirilen yöntemle, ilişki tahmincilerin ağ çıkarımına olan etkisinden faydalanılarak, biyolojik ve istatistiksel olarak hastalıkla ilişkisi yüksek biyolojik ağ oluşturulmaya çalışılmıştır. Önerilen yöntem Bölüm 5'te açıklanmıştır.

1.3 Hipotez

Düzenleyici ağlar (regulatory networks) diğer bir deyişle GAÇ yöntemleri insan hastalıklarının patogenezinin genom çapında anlaşılmasında önemli ilerleme sağladığı için literatürde sıkça kullanılmaktadır. Fakat yapılan bazı çalışmalarda korelasyon temelli ağ çıkarım yöntemlerinin kanser, otizm, diyabet, alzaymır gibi karmaşık yapıdaki moleküler etkileşimlerin yer aldığı hastalıkların oluşumundaki lineer olmayan ilişkilerin kestiriminde yeterince başarılı olmadığı belirtilmiştir [17], [29]. Bu sorunun üstesinden gelmek için de KB tabanlı yöntemlerin kullanıldığı birçok GAÇ yöntemi geliştirilmiştir [5], [6], [7], [10]. Bu algoritmaların detayları Bölüm 3'te verilmiştir. Ancak [5], [6], [7] ,[10] ve [14]'de ağ, alt ağlara (birimlere-modüllere) ayrılarak hastalıkla ilişkili bu alt ağlara odaklanan bir yöntem önerilmemiştir. Biz tezimizde, WGCNA [22] yönteminde genler/proteinler arasındaki ilişki tahmininde varsayılan olarak kullanılan korelasyon tabanlı yöntemler yerine, lineer olmayan ilişkilerin tespitini de sağlamak amacıyla, KB tabanlı ilişki tahmincileri yönteme entegre ederek yöntemin başarısının iyileştirilmesini ve hastalıkla ilişkili modüllerin tespitinin yapılmasını hedeflemekteyiz.

Son yıllarda biyoinformatik alanında yapılan araştırmalarda biyolojik sistemlerin nasıl çalıştığını daha doğru tespit etmek için farklı veri türlerinin entegrasyonu üzerinde çalışmalar yapılmıştır [16], [17], [18]. Bu çalışmalarda kullanılan çoğu biyolojik veri türü, genomik (DNA dizilimi, genom haritalama, vb.), proteomik (protein yapısı ve işlevleri), transkriptomik (ifade verileri, mRNA'ları içeren RNA seti) ve metabolomikler (hücresel işlemler sonucu metabolitlerin incelenmesi) gibi "-omik" sınıflandırmalarından biridir.

Bu çalışmada da, doğruluğu daha yüksek düzenleyici ağlar oluşturmak için gen ifade ve proteomik verileri (genomik ve proteomik) kullanılarak oluşturulan ağların entegrasyonu üzerine bir yaklaşım sunulacaktır. Bu veri türlerinden oluşan ağların entegrasyonu sağlanarak gen düzenleyici ağların farklı koşullar ve hastalık halleri altında biyolojik sistemlerin daha iyi açıklanabileceğinin gösterimi sağlanmaya çalışılacaktır.

Ayrıca, gen verilerinden elde edilen mRNA'ların ifade düzeyleriyle, bu mRNA'lardan kodlanan proteinlerin miktarları arasında istatistiksel olarak zayıf bir doğrusal ilişkinin bulunduğu bildirilmiştir [30]. Yapılan biyolojik çalışmalarda hücrelerin fonksiyonlarının genler ve mRNA'lardan ziyade proteinler tarafından düzenlendiği belirtilmiştir [31]. Dolayısıyla sadece gen ifade verisi kullanılarak çıkarımlanan ağların, biyolojik etkileşimlerin belirlenmesinde yeterli olmayacağı açıktır. Bu sebepten, biyolojik olarak daha anlamlı ağların oluşturulması için tezimizde proteomik verileri de kullanılmıştır.

BÖLÜM 2

İLİŞKİ TAHMİNCİLERİ

Bu bölümde hücrelerdeki moleküller arasındaki etkileşimin büyük bir bölümünü keşfetmemizi sağlayan ilişki tahmincileri ve bu tahmincilerin açıklamalarında sıklıkla belirtilen entropi ve KB gibi çeşitli bilgiler açıklanacaktır.

2.1 İlişki Tahmincilerinin Kullanımı

Gen çiftleri arasındaki ilişki, gen ağı çıkarımı işlemi öncesinde verimli ve doğru olarak tahmin edilmelidir. Eğer bu adım doğru bir şekilde yerine getirilmezse hangi gen ağı çıkarım yönteminin kullanıldığı fark etmeksizin sonuçta yöntemin başarım oranı düşük olacaktır. Bu sebepten gen ağı çıkarımının başarısını arttırmak için farklı ilişki tahmincileri geliştirilmiş ve kullanılmıştır. Tez çalışmamızda incelediğimiz ilişki tahmincilerini, kullandıkları yöntemlere bakarak korelasyon tabanlı ve karşılıklı bilgi tabanlı olmak üzere iki sınıfa ayırabiliriz. Bu yöntemlerden korelasyon tabanlı olanlar herhangi bir ön işlem gerektirmeksizin veri setine uygulanabilmektedir. Karşılıklı bilgi temelli yaklaşımlar entropi temelli ve doğrudan karşılıklı bilginin hesaplanmasına dayalı olabilmektedir. Eğer entropi temelli bir ilişki tahminci analiz için seçilmişse, ilişki tahmincisi kullanılmadan önce ayrıklaştırma işleminin gerçekleştirilmesi gerekmektedir. Çalışmamızda incelediğimiz ilişki tahmincilerinin kullandıkları yöntemlere göre sınıflandırılması Çizelge 2.1'de verilmiştir. Ayrıca, KB tabanlı ilişki tahmincilerinde kullandığımız ayrıklaştırma yöntemleri hakkında genel bir bilgi Bölüm 2.3'te verilmiştir.

Çizelge 2. 1 İlişki tahmincilerin kullandıkları yöntemlere göre sınıflandırılmaları

Korelasyon Tabanlı ilişki tahminciler	**KB Tabanlı (Hücre olasılıkları kullanan "Entropi tabanlı") ilişki tahmincileri**	
	Her bir hücrede bir tek örnek içeren İlişki Tahmincileri	**Her bir hücrede çok sayıda örnek içeren kestirimciler**
Pearson KKD [3]	Çekirdek Yoğunluğu İlişki Tahmincisi (ÇYT) [32]	Ampirik (Emprical, Maximum-Likelihood, Naive) [33]
Spearman KKD [3]		Miller-Madow (MM) [34]
Kendall Tau KKD [35]		James-Stein Shrinkage (Shrink) [36]
		B-Spline (BS) [11]
		Schurmann-Grassberger (SG) [37]

2.2 Entropi ve Karşılıklı Bilgi

Entropi rastgele bir değişken hakkındaki belirsizlik metriğidir. Bilgisayar bilimleri alanında genellikle sıkıştırma uygulamalarında kullanılmakta olup, haberleşme (metin veya resim sıkıştırma, sensör konumlarını analiz etme), doğal dil işleme, sinyal analizi (segmentasyon, algılama, görüntü kaydı ve doku sınıflaması), istatistiksel öğrenme, kimya, fizik vb. gibi çeşitli uygulama alanlarında da kullanılmaktadır.

Entropisi ölçülecek rastgele değişkenin öngörülemezliği arttıkça entropisi de artar. Ancak, öngörülemezlik azaldıkça yani rastgele değişkenin durumu daha bilinir oldukça entropi azalır. Dolayısıyla bir rastgele değişkenin durumu kesin olarak biliniyorsa bu değişkenin entropisi sıfır olur. Rasgele bir X değişkeni için Shannon entropisi Bağıntı (2.1)'deki gibi tanımlanır:

$$H(X) = -\sum_{x_i \in X} P(X = x_i) \log_2(P(X = x_i)) \quad (2.1)$$

Daha önce belirtildiği gibi, eğer entropi temelli bir ilişki kestirimci kullanılacaksa, rastgele değişkenlerin entropisini elde edebilmek için veri kümesinin ayrıklaştırılarak hücrelere bölünmesi gerekmektedir. Ayrıklaştırma işleminin sonucunda her bir *i* hücresinin olasılık

dağılımı $P(X=x_i)$ olmak üzere Bağıntı (2.1) kullanılarak entropi elde edilir. Ayrıklaştırma işlemi ve kullanılabilecek teknikler Bölüm 2.3'de verilmiştir.

İki rastgele değişken arasındaki KB değeri, bu iki değişkenin birbiriyle olan bağımlılığını, ilişkisini ve etkileşimini ölçer. H(X) ve H(Y) bireysel entropi, H(X,Y) birleşik entropi olmak üzere KB değeri (2.2)'den elde edilebilir:

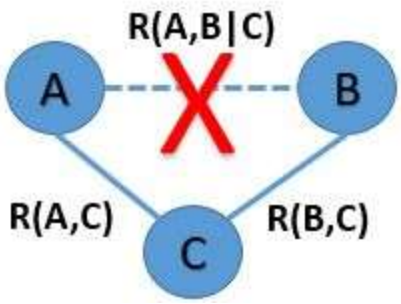
$$KB(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$= -\sum_{x_i \in X} \sum_{y_j \in Y} P\big(X = x_i, Y = y_j\big)\, log \frac{P(X=x_i, Y=y_j)}{P(X=x_i)\, P(Y=y_j)} \quad (2.2)$$

$H(X,Y)$ birleşik entropisi $P(X=x_i, Y=y_j)$ birleşik olasılık dağılımı kullanılarak elde edilir:

$$H(X,Y) = -\sum_{x_i \in X} \sum_{y_j \in Y} P\big(X = x_i, Y = y_j\big) \log\big(P(X = x_i, Y = y_j)\big) \quad (2.3)$$

İki rastgele değişken arasındaki ilişki değeri korelasyon ile elde edilirken sistemdeki diğer değişken veya değişkenlerin bu iki değişkenin korelasyonu üzerinde bazı etkileri olabilir. Bu etkiler, gerçekte birbiriyle doğrudan bir ilişkisi olmayan bu iki değişkenin arasında bir bağlantı olduğu yanılgısına sebep olabilir. Böyle bir durumda bu iki değişken arasındaki dolaylı bağlantıların tespit edilerek elenmesi gerekmektedir. Bu durumu bir örnek ile açıklayacak olursak: Şekil 2.1'de verildiği gibi, A ile C ve B ile C değişkenleri arasında doğrudan bir bağlantı olduğunu, ancak A ve B arasında doğrudan bir bağlantı olmadığını kabul edelim. A ve B arasında doğrudan bir bağlantı bulunmamasına rağmen ikisi de C ile ilişkide oldukları için bu iki değişken arasında yüksek bir korelasyon skoru elde etmek olasıdır. Bu durumda, bu korelasyonun C'den kaynaklandığı ve A ile B arasında doğrudan bir ilişki bulunmadığı tespit edilerek bu bağlantı kopartılmalıdır. Bu amaçla literatürde farklı yöntemler [38], [39] önerilmiştir. Bu yöntemler, A ile B'nin C ile ilişkide bulunmayan kısımlarının ilişki değerlerini hesaplar ve hesaplanan değer belli bir eşik değerinin altında ise A ile B arasında doğrudan bir bağlantı olmadığı kabul edilir.



Şekil 2. 1 A ve B değişkenleri C değişkeni üzerinden dolaylı ilişkidedir

2.3 Ayrıklaştırma Yöntemleri

KB skoru tahmin eden kestirimcilerin bir kısmı bu skoru doğrudan elde edebilirken, bir kısmı bu skoru bireysel ve birleşik entropiden bulur. Entropi tabanlı KB skoru kestirimcileri Pearson, Spearman vb. gibi korelasyon tabanlı kestirimcilerden ve diğer KB skoru kestirimcilerinden farklı olarak veri kümesinin ayrıklaştırılmasını veya hücrelere bölünmesini gerektirir. Ayrıklaştırma işleminden sonra Bağıntı (2.1)'deki gibi her bir hücrenin olasılık dağılımından entropi; entropiden de Bağıntı (2.2)'deki gibi KB skoru elde edilir. Bu tezde iki temel ayrıklaştırma tekniğini kullanılmıştır. Bunlar eşit genişlik ve eşit frekans ayrıklaştırma yöntemleridir ve Bölüm 2.3.1 ve 2.3.2'de sırasıyla verilmiştir.

2.3.1 Eşit Genişlik (Equal Width) Ayrıklaştırma Yöntemi

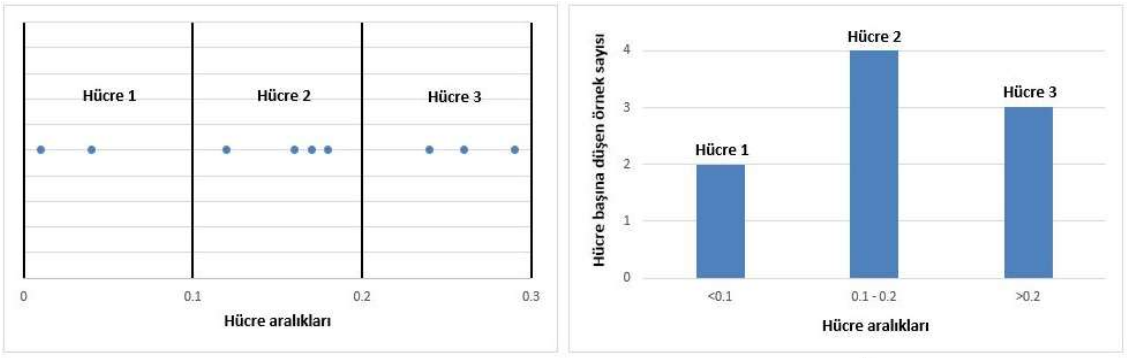
Veri kümesi bu teknik ile ayrıklaştırılırken her bir hücrenin eşit genişlikte (EG) olacak şekilde bölütlenmesi sağlanır. Böylece her bir hücreye düşen örnek sayısı değişken olmaktadır. Bir A kümesinin örnek değerleri (0,01; 0,04; 0,12; 0,16; 0,17; 0,18; 0,24; 0,26; 0,29) olsun. G: genişlik, N: örnek sayısı, EB: kümenin en büyük elemanı, EK: kümenin en küçük elemanı olmak üzere bir kümenin EG ayrıklaştırma yöntemine göre kaç hücreye bölünmesi gerektiğini hesaplamak için Bağıntı 2.4 kullanılır. Kümenin kaç hücreye ayrılması gerektiği hakkında kesin bir yaklaşım olmamakla birlikte literatürde genellikle örnek sayısının karekökü alınarak hesaplanmaktadır [3], [9].

$$G \cong \frac{EB-EK}{\sqrt{N}} \cong \frac{0,29-0,01}{\sqrt{9}} \cong 0,1 \qquad (2.4)$$

A kümesi eşit genişlik ayrıklaştırma yöntemine göre ayrıklaştırıldığında her bir hücrenin genişliği "0,1" birim olmuş ve Hücre 1'e iki, Hücre 2'ye dört ve son olarak Hücre 3'e de

üç örnek düşmüştür. Ayrıklaştırma işlemi sonucu oluşan hücrelerin genişliğinin ve hücre başına düşen örnek sayısının grafiksel gösterimi Şekil 2.2'de verilmiştir. Şekil 2.2 (a)'da ve (b)'de x ekseni hücre genişliğini bildirmekte olup, y ekseninde örneklerin daha net görülebilmesi için tüm değerler aynıdır.

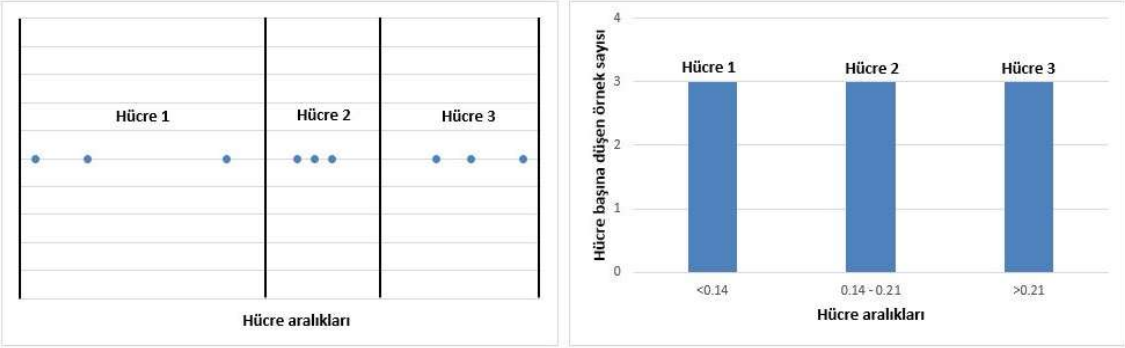
Eşit genişlik ayrıklaştırma yöntemi farklı veri kümeleri için birbirinden bağımsız şekilde uygulanır. Örneğin X ve Y genlerine ait ifade verileri için hücre aralıkları ayrı ayrı belirlenir. Küresel eşit genişlik (global equal with) ayrıklaştırma yönteminde ise X ve Y genlerine ait ifade verileri için aynı hücre aralığı kullanılır [9].



Şekil 2. 2 (a) Eşit genişlik ayrıklaştırma yöntemi sonucu oluşan hücrelerin genişliği ve örnek dağılımları (b) Hücre başına düşen eleman sayıları

2.3.2 Eşit Frekans (Equal Frequence) Ayrıklaştırma Yöntemi

Veri kümesi bu teknik ile ayrıştırılırken her bir hücreye eşit sayıda örnek düşecek şekilde hücre genişlikleri belirlenir. Bu yöntemde hücre başına düşen eleman sayılarının eşit olması hedeflenirken hücre aralıkları farklılaşır. Bölüm 2.3.1'de verilen A kümesi bu yöntem ile ayrıklaştırıldığında her bir hücreye düşen eleman sayısı üç olmuştur. a_i, hücre aralıklarındaki örnek değerlerini belirtmek üzere; Hücre 1 için, $0 < a_i < 0{,}14$ aralığında; Hücre 2 için, $0{,}14 \leq a_i \leq 0{,}21$ aralığında; son olarak Hücre 3 için de $0{,}21 < a_i < 0{,}3$ aralığında oluşmuştur. Ayrıklaştırma işlemi sonucu oluşan hücrelerin genişliğinin ve hücre başına düşen örnek sayısının grafiksel gösterimi Şekil 2.3'te verilmiştir. Şekil 2.3 (a)'da ve (b)'de x ekseni hücre genişliğini bildirmekte olup, y ekseninde örneklerin daha net görülebilmesi için tüm değerler aynıdır.



Şekil 2. 3 (a) Eşit frekans ayrıklaştırma yöntemi sonucu oluşan hücrelerin genişliği ve örnek dağılımları (b) Hücre başına düşen eleman sayıları

2.4 Pearson Korelasyon Katsayı Değeri (Pearson KKD)

Karl Pearson [40] tarafından önerilen bu yöntem iki rastgele değişken arasında doğrusal bir ilişkinin istatistiksel olarak değerini ölçmek için kullanılan korelasyon tabanlı bir yaklaşımdır. Pearson KKD, biyoinformatik alanındaki birçok çalışmada [14], [38], [41], [42], [43] kullanılmıştır ve istatistik çalışmalarında *r* ile gösterilmektedir.

A ve B bağımsız değişkenlerinin kovaryansını, kov(A,B); A'nın standart sapmasını, S_a; B'nin standart sapmasını, S_b; belirtmek üzere Pearson KKD:

$$r = \frac{kov(A,B)}{S_a\, S_b} \tag{2.5}$$

bağıntı (2.5) ile hesaplanır.

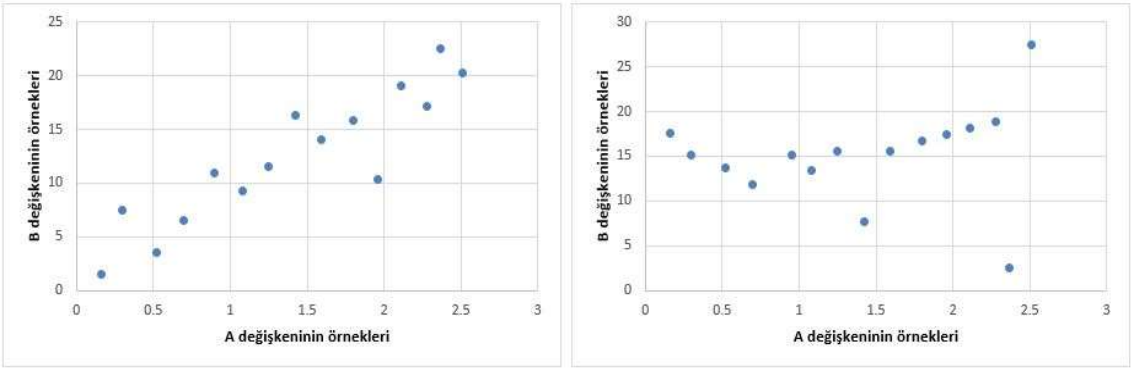
A ve B bağımsız değişkenlerinin örnek sayısı: N, A'nın örnekleri: a_i ve B'nin örnekleri: b_i olmak üzere A ve B arasındaki Pearson KKD:

$$r = \frac{N\sum_{i=1}^{N} a_i b_i - \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i}{\sqrt{N\sum_{i=1}^{N} a_i^2 - \left(\sum_{i=1}^{N} a_i\right)^2}\sqrt{N\sum_{i=1}^{N} b_i^2 - \left(\sum_{i=1}^{N} b_i\right)^2}} \tag{2.6}$$

bağıntı (2.6) ile hesaplanır.

KKD her zaman [-1, 1] aralığında değer alır. Eğer iki değişken arasında KKD katsayısı 0'dan büyükse, değişkenler arasında pozitif bir doğrusal ilişki vardır ve KKD 1'e yaklaştıkça bu pozitif doğrusallık artar. Eğer iki değişken arasında KKD katsayısı 0'dan küçükse, değişkenler arasında negatif bir doğrusal ilişki vardır ve KKD -1'e yaklaştıkça bu negatif doğrusallık artar. Eğer KKD 0 ise bağımsız değişkenler arasında doğrusal bir ilişki yoktur.

KKD'nin 0 olması bu değişkenlerin birbirinden kesinlikle bağımsız olması anlamına gelmez. Aralarında doğrusal olmayan bir ilişki de bulunabilir. PKD sadece doğrusal ilişkileri ölçebilen bir yöntem olduğu için doğrusal olmayan ilişkilerin skorlarını tahmin edemez. Bu durumla ilgili bir örnek Şekil 2.4'te verilmiştir. Şekil 2.4 (a)'da verilen örnekte iki rastgele değişken arasında doğrusala yakın bir ilişki olduğu görülmektedir. Pearson KKD sonucunda da ilişki değeri 0,911 olarak elde edilmiştir. Ancak, Şekil 2.4 (b)'deki dağılıma bakıldığında değişkenler arasındaki ilişkinin doğrusallıktan uzaklaştığı görülmektedir, Pearson KKD ile ilişkinin değeri 0,163 olarak elde edilmiştir.



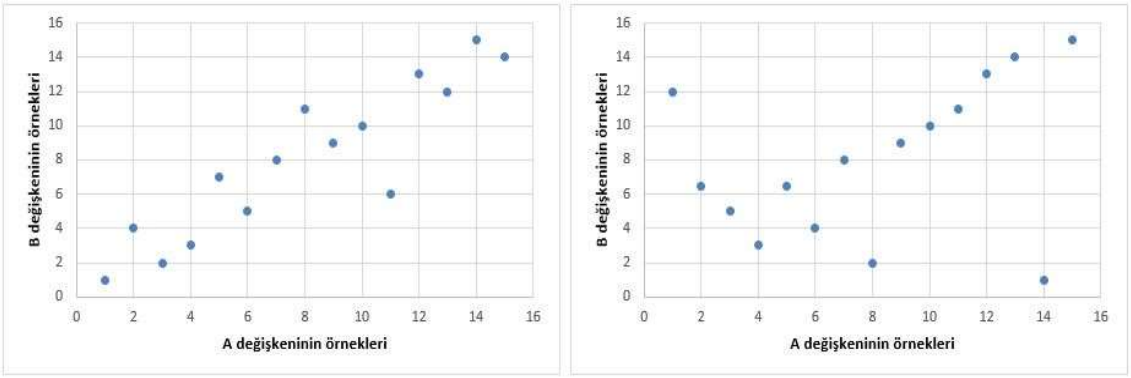
Şekil 2. 4 (a) Bağımsız değişkenler arasındaki ilişki doğrusallığa yakın iken (Pearson KKD: 0,911); (b) Bağımsız değişkenler arasındaki ilişki doğrusallıktan uzak iken (Pearson KKD: 0,163);

2.5 Spearman Korelasyon Katsayı Değeri (Spearman KKD)

Biyoinformatik alanında yapılan birçok çalışmada [3], [14], [38], [44], [45], [46] kullanılmış olan Spearman KKD, Pearson KKD'nin sıralamaya dayalı özel bir versiyonudur. Spearman KKD elde edilirken öncelikle veri kümesindeki orijinal değerler küçükten büyüğe sıralanır ve ardından veri değeri yerine sahip olduğu sıra numarası kullanılır. Spearman KKD yönteminde Bağıntı (2.6)'da ki a_i ve b_i veri örneklerinin gerçek değerleri yerine bu örneklerin sıralanması sonucundaki sıra bilgisi kullanılarak bağımsız değişkenler arasındaki ilişki değeri hesaplanır.

Şekil 2.4 (a-b)'de verilen örnek değerler için sıra değerleri hesaplanmış ve bu değerlerin grafiksel gösterimi Şekil 2.5 (a-b)'de sırasıyla verilmiştir. Şekil 2.5 (a)'da ki gibi bağımsız değişkenler arasındaki ilişki doğrusallığa yakın iken Spearman KKD 0,91 çıkarak Pearson KKD'ye çok yakın olmuştur. Fakat Şekil 2.5 (b)'de ki gibi bağımsız değişkenler arasındaki

ilişki doğrusallıktan uzak iken Spearman KKD 0,356 çıkarak Pearson KKD'den daha iyi sonuç vermiştir.



a b

Şekil 2. 5 (a) Bağımsız değişkenler arasındaki ilişki doğrusallığa yakın iken (Pearson KKD: 0.911, Spearman KKD: 0,91); (b) Bağımsız değişkenler arasındaki ilişki doğrusallıktan uzak iken (Pearson KKD: 0,163, Spearman KKD: 0,356);

Spearman KKD ile bağımsız iki değişken arasındaki ilişkinin monoton olması koşuluyla doğrusal ilişkiler ile birlikte doğrusal olmayan ilişkiler için de ilişki değeri başarılı olarak tespit edilebilir. Fakat tekdüze olmayan ilişkiler için, değişkenler arasında doğrusallık olsun olmasın, ilişki değeri Spearman KKD ile etkili bir şekilde yapılamayabilir. Şekil 2.5 (b)'deki veriler için hesaplanmış değerler bunu açıkça göstermektedir.

2.6 Kendall Tau Korelasyon Katsayı Değeri (Kendall KKD)

Kendall KKD biyoinformatik alanında yapılan çalışmalarda [47], [48], [49] kullanılan ilişki tahmincilerden biridir ve Spearman KKD gibi verilerin sıra değerlerine göre hesaplanır. Sıra değerleri belirlenen A ve B bağımsız değişkenleri için her bir çift değerin, hem A değişkeninde, hem de B değişkeninde artıp azalmasına göre uyumluluk (concordant) ve uyumsuzluk (discordance) sayıları belirlenir. A ve B bağımsız değişkenlerinin örnek sayısı: N; A'nın örnekleri: a_i; B'nin örnekleri: b_i olmak üzere örneklerin uyumlu ya da uyumsuz oldukları şu şekilde belirlenir:

$$Sonuç \begin{cases} Eğer\ (a_i > a_j\ ve\ b_i > b_j)\ veya\ (a_i < a_j\ ve\ b_i < b_j)\ ; uyumlu\ +1 \\ Eğer\ (a_i > a_j\ ve\ b_i < b_j)\ veya\ (a_i < a_j\ ve\ b_i > b_j)\ ; uyumsuz\ -1 \\ Eğer\ a_i = a_j\ veya\ b_i = b_j); nötr\ 0 \end{cases} \quad (2.7)$$

Bu işlem sonucunda bulunan toplam uyumlu sayısı C, toplam uyumsuz sayısı D olmak üzere Kendall KKD (τ) Bağıntı 2.8'e göre hesaplanır:

$$\tau = \frac{C-D}{\frac{1}{2}(N-1)} \quad (2.8)$$

Literatürde Kendall KKD ve Spearman KDD'nin çoğu zaman yakın ilişki değerleri ürettiği belirtilmiş olmakla birlikte Kendall KDD'nin uyumlu ve uyumsuz çiftlere göre bir hesaplama yapması, istatistiksel olarak daha iyi özelliklere sahip olabilmesini sağlamaktadır [47], [49].

2.7 Ampirik (Empirical) İlişki Tahmincisi

Ampirik ilişki tahmincisi KB tabanlı yöntemlerin temelini oluşturur ve daha önce de belirtildiği gibi entropinin hesaplanmasını gerektirir. Entropi hesabı için de veri kümesinin ayrıklaştırılması gerekmektedir ve ayrıklaştırma işlemleri ile ilgili gerekli açıklamalar Bölüm 2.3'te detaylı bir şekilde verilmiştir.

Bazı çalışmalarda en büyük benzerlik (Maximum Likelihood) [12], [14] olarak da geçen ampirik ilişki tahmincisinde ayrıklaştırma işlemi sonucunda her bir hücredeki eleman sayısına bakılır ve her bir hücre için bireysel frekanslar elde edilir. Ayrıklaştırılan iki ayrı rastgele değişkenin veri örneklerinin birleşik frekansı da her bir hücre için elde edilir.

Ampirik kestirimde deneysel entropi, gözlemlerin olasılık dağılımından elde edilir. Tek bir rastgele değişken için örnek sayısını, X; hücre sayısı, b; *i*'inci hücredeki örnek sayısı, x_i olmak üzere Bağıntı 2.9'dan deneysel entropi olan H_{Deney} elde edilir.

$$H_{Deney} = -\sum_{i=1}^{b} \left(\frac{x_i}{X}\right) log \left(\frac{x_i}{X}\right) \quad (2.9)$$

H_{Deney}, ayrık bir rastgele değişken için Ampirik entropi tahminini vermektedir. Bu yöntemde ayrıklaştırma işleminde kullanılan hücre sayısının (b'nin) artmasının hücre olasılıklarının seyrek-örneklenmesine (undersampling) sebep olacağı için tahmin edilen H_{Deney} değeri gerçek entropi değerinden uzaklaşacaktır [3], [33], [36]. Bu durumda Ampirik ilişki tahmincisinin asimptotik sapmasının (Asymptotic bias) Bağıntı 2.10'daki değere ulaşmasına sebep olur.

$$Sapma\left(H_{Deney}\right) = -\frac{b-1}{2X} \quad (2.10)$$

2.8 Miller-Madow (MM) İlişki Tahmincisi

Literatürde biyoinformatik alanında yapılan farklı çalışmalarda da [14], [50], [51] kullanılmış olan MM ilişki tahmincisi Ampirik ilişki tahmincisindeki tahmin sapmasından kaynaklanan olumsuzluktan kaçınmak için geliştirilmiş bir yöntemdir. Başka bir deyişle MM ilişki tahmincisi Bağıntı 2.10'da verilen sapmayı hesaba katarak daha başarılı tahminlerde bulunabilmektedir ve literatürde birçok çalışmada başarılı bir şekilde kullanılmıştır.

Her bir ilişki tahmincisinin amacı hem sapmayı hem de varyansı en küçük seviyede tutabilmektir. Ancak, sapma ve varyans arasında ters orantı bulunduğu için biri artarken diğeri azalmaktadır. Bir kestirimcinin karmaşıklığı arttıkça sapma azalır, ancak varyans çok artar. MM ise Bağıntı 2.11'i kullanarak karmaşıklığı ve varyansı artırmadan sapmayı azaltabilmektedir.

$$H_{MM} = H_{Den} \quad + \frac{b-1}{2X} \tag{2.11}$$

2.9 James-Stein Shrinkage (Shrink) İlişki Tahmincisi

Biyoinformatik alanında yapılan farklı çalışmalarda [3], [14] incelenmiş olan ve Hausser ve Strimmer [36] tarafından uyarlanan Shrink ilişki tahmincisi aşırı uyumluluktan (overfitting) kaçınmak için ampirik kestirimin konveks toplamını ve hedef dağılımı kullanmaktadır. x_i geninin aykırı dağılımı için Shrinkage kestirimi bağıntı 2.12 kullanılarak bulunur.

$$p_{shrink}(x_i) = \lambda_M t_i + (1 - \lambda_M) p(x_i) \tag{2.12}$$

Burada t_i hedef dağılımı belirtir ve $\lambda_M \in [0,1]$ olmak üzere marginal srinkage yoğunluğunu ifade eder. Hedef dağılım, n hücre sayısı olmak üzere, entropiyi en üst düzeye çıkarmak için "1/n" olarak seçilir. Aykırı Srinkage yoğunluğu Bağıntı 2.13 ile bulunur.

$$\lambda_M = \frac{1-\sum_i p(x_i)}{(n-1)\ \sum_i (t_i - p(x_i))^2} \tag{2.13}$$

x_i ve y_j genleri için birleşik Shrinkage dağılımı, λ_J birleşik Shrinkage yoğunluğunu belirtmek üzere bağıntı 2.14 kullanılarak bulunur.

$$p_{shr} \quad (x_i, y_j) = \lambda_J t_{ij} + (1 - \lambda_J) p(x_i, y_j) \tag{2.14}$$

Yine hedef dağılım, n hücre sayısı olmak üzere, "1/n" olarak seçilir. Birleşik Shrinkage yoğunluğu $\lambda_J \in [0;1]$'dir ve Bağıntı 2.15 ile bulunur.

$$\lambda_J = \frac{1-\sum_i \sum_j p(x_i,y_j)^2}{(n-1)\ \sum_i \sum_j (t_{ij}-p(x_i,y_j))^2} \quad (2.15)$$

Shrinkage bireysel entropi tahmini için Bağıntı 2.16 ve birleşik entropi tahmini Bağıntı 2.17 kullanılarak hesaplanır.

$$H_{Shri} \quad (x) = -\sum_i p_s(x_i)\ log\ p_s(x_i) \quad (2.16)$$

$$H_{Shrink}(x,y) = -\sum_i \sum_j p_s(x_i,y_j)\ log\ p_s(x_i,y_j) \quad (2.17)$$

2.10 Schurmann-Grassberger (SG) İlişki Tahmincisi

Literatürde biyoinformatik alanında yapılan birçok çalışmada [8], [9], [10], [17], [36] kullanılmış olan SG ilişki tahmincisi Schurmann ve Grassberger [37] tarafından uygulanmış olan entropi temelli bir kestirimcidir. SG ilişki tahmincisi ayrık rasgele değişkenin entropisini tahmin etmek için Dirichlet dağılımını kullanır. Dirichlet dağılımı, beta dağılımının çok değişkenli genelleştirilmesidir. Ayrıca, Bayes istatistiklerinde multinominal dağılımdan önceki eşleniktir. Dirichlet dağılımı, olasılık dağılımlarının ortalama θ_k dağılımı ile tanımlanır ve Bağıntı 2.18'de bir Dirichlet dağılım yoğunluğu verilmiştir.

$$f(X;\theta) = \frac{\Pi_{k\in X}\Gamma(\theta_k)}{\Gamma(\sum_{k\in X}\theta_k)} \prod_{k\in X} X_k^{\theta_{k-1}} \quad (2.18)$$

Her bir *k* hücresi için ortalama θ_k olasılığı Schurmann-Grassberger parametresi $\frac{1}{b}$ kullanılarak Bağıntı 2.19 ile tahmin edilir.

$$\theta_k = \frac{n_k+\frac{1}{b}}{N+1} \quad (2.19)$$

Bağıntı 2.19'da n_k, *k*'ncı hücrede bulunan örnek sayısını; *N,* toplam örnek sayısını; *b* ise toplam hücre sayısını belirtmektedir. Genel olarak *N* toplam örnek sayısına bir eklenir.

Entropi Bağıntı 2.20 ile elde edilir.

$$H_{SG} = -\sum_{k=1}^{b} \theta_k \log \theta_k \quad (2.20)$$

2.11 B-spline (BS) İlişki Tahmincisi

Daub vd. [8] tarafından önerilen B-spline yöntemi Ampirik ilişki tahmincisinin bir modifikasyonudur ve örnekleri birden fazla hücreye yerleştirerek dağılımı düzeltmeyi (smoothing) hedefler. Bu düzeltme B-spline polinomları ile yapılır.

B-spline, her bir polinom "temel fonksiyonlar" (basis functions) denilen doğrusal bir kombinasyon olmak üzere, parçalı bir polinom fonksiyonudur. *k'ncı* dereceden bir B-spline'da bu temel fonksiyonların derecesi *k-1* olan polinomlardır. Parçalı polinomlar *t* olarak belirtilen bir düğüm vektöründe (knock vector), hücre sayısı: m, düğüm sayısı: *m+k* olmak üzere *k'ıncı* dereceden bir B-Spline ile birleştirilir. Düğüm vektörleri Bağıntı 2.19 ile tespit edilir:

$$t_i = \begin{cases} 0 & eğer \quad i < k \\ i-k+1 & eğer \quad k \le i \le m-1 \\ m-k-1 & eğer \quad i > m-1 \end{cases} \tag{2.19}$$

Düğüm vektörü azalmaz ve B-spline temel fonksiyonlarını tamamen belirler. *k'ıncı* dereceden bir B-spline için, temel fonksiyonlar *k* ardışık düğüm arasında tanımlanır. Temel fonksiyonlar *k* = 0'dan istenilen düzlem sırasına ulaşana kadar tekrarlanacak şekilde belirlenir. *z* ∈ [0; m-k+1] için, düğüm vektörünün aralığı, m temel fonksiyonlar olmak üzere (kutu başına bir temel fonksiyon) Cox-de Boor özyineleme formülü tarafından Bağıntı 2.20 ve Bağıntı 2.21 ile belirlenir:

$$B_{i,1}z = \begin{cases} 1 & eğer \ \ z \in [t_i, t_{i+1}) \\ 0 & değilse \end{cases} \tag{2.20}$$

$$B_{i,k}(z) = B_{i,k-1}(z)\frac{z-t_i}{t_{i+k-1}-t_i} + B_{i+1,k-1}(z)\frac{t_{i+k}-z}{t_{i+k}-t_{i+1}} \tag{2.21}$$

k = 1 için temel fonksiyonlar, bir bölme içinde 1, diğer bölmelerde 0 olan adım işlevleridir, bu nedenle herhangi bir örnek yalnızca konumunu kaplayan bölmeye yerleştirilir. Aslında, *k*. dereceden B-spline her bir örneği *k*. kutulara yerleştirir, yani k = 1 olan bir B-spline tahmincisinin ampirik, diğer adıyla maksimum olabilirlik (Maximum Likelihood) tahmincisine eşdeğer olduğu anlamına gelir. Bunun nedeni, *z*'nin her bir değerinde tanımlanan *k* temel fonksiyonlarının olması ve her temel fonksiyonun tek bir bölmeye tekabül etmesidir. Herhangi bir *z*'deki temel fonksiyonların toplamı 1 olduğundan,

bunları her bir noktayı *z*-değerine dayanarak *k* hücrelerine yerleştirmek için kullanabiliriz. Bu, her hangi bir *X* geni için X_{max} en büyük, X_{min} en küçük ifade değerini belirtmek üzere Bağıntı 2.22 ile tek bir genin ifade değerlerini kullanarak düğüm vektörünün aralığına dönüştürmemizi gerektirir.

$$z = (m - k + 1)\frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.22}$$

Tek gen için B-spline değerinin değerlendirilmesi verilen bir $A^X \in R^{N \times m}$ matrisindeki *ij*'. eleman *i*'. örneğin *j*'. hücredeki ağırlığını belirtmek üzere bu genin ayrık olasılığı Bağıntı 2.23 ile belirlenir.

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} A_{ij}^{x}, \qquad j = 1, \dots, m \tag{2.23}$$

Başka bir y geni için de gerekli hesaplamalar yapıldıktan sonra birleşik olasılık dağılımı Bağıntı 2.24 ile belirlenir.

$$p(x, y) = \frac{1}{N} A^x A^y \tag{2.24}$$

Sonrasında Bağıntı 2.3 kullanılarak birleşik KB hesaplanır.

Çalışmamızda B-spline ilişki tahmincisinin başarımının değerlendirilmesi esnasında *k* eğri derecesi için 1, 2 ve 3 değerleri seçilmiştir.

2.12 Çekirdek Yoğunluğu İlişki Tahmincisi (ÇYT)

Moon vd. [32] tarafından geliştirilen ÇYT (Kernel Density Estimator) tahmincisi olasılık dağılımlarını tahmin ederken dikdörtgen histogram hücreleri yerine çekirdek fonksiyonları kullanır. Biyoinformatik alanında farklı çalışmalarda [52], [53], [54] kullanılmış olan ÇYT mikro dizin veri kümeleriyle kullanılması, ilk olarak ARACNE GAÇ algoritmasının KB tahmini aşamasında önerilmiştir ve çalışmada Gaussian çekirdek tahmincisi kullanılarak gen ifade verileri için KB'yi tahmin etmişlerdir [5]. Gen ifade verilerinin olasılık yoğunluk fonksiyonu (probability density function), her bir gen çifti için ve her bir gen için ayrı ayrı ÇYT ile tahmin edilebilir. Her bir gen için bireysel olasılık yoğunluğu ve gen çiftleri için birleşik olasılık yoğunluğu hesaplandıktan sonra, gen çiftleri arasındaki KB ilişki değeri Bağıntı 2.28 ile elde edilebilir.

ÇYT ile f(x) fonksiyonunu tahmin ederken, bu yoğunluk fonksiyonundan seçilen G_1, G_2, …, G_N örnekleri kullanılır. ÇYT'nin tanımı bağıntı 2.25 ile verilmiştir:

$$f_h(x) = \frac{1}{N} \sum_{i=1}^{N} K_h(x - x_i) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) \quad (2.25)$$

Bağıntı 2.25'te $K_h(.)$: ölçeklenmiş çekirdek fonksiyonunu, $K(.)$: çekirdek fonksiyonunu, N: örnek adedini ve h: bant genişliğini belirtmektedir. Bant genişliği, pencere genişliği veya düzleştirme parametresi olarak da bilinmektedir. ARACNE'de kullanıldığı gibi yapılan çalışmalarda daha çok Gauss fonksiyonu çekirdek fonksiyon olarak seçilmiştir [52], [53], [54]. Çekirdek fonksiyonun bant genişliğini belirten h parametresi $h>0$ şartını sağlamalıdır.

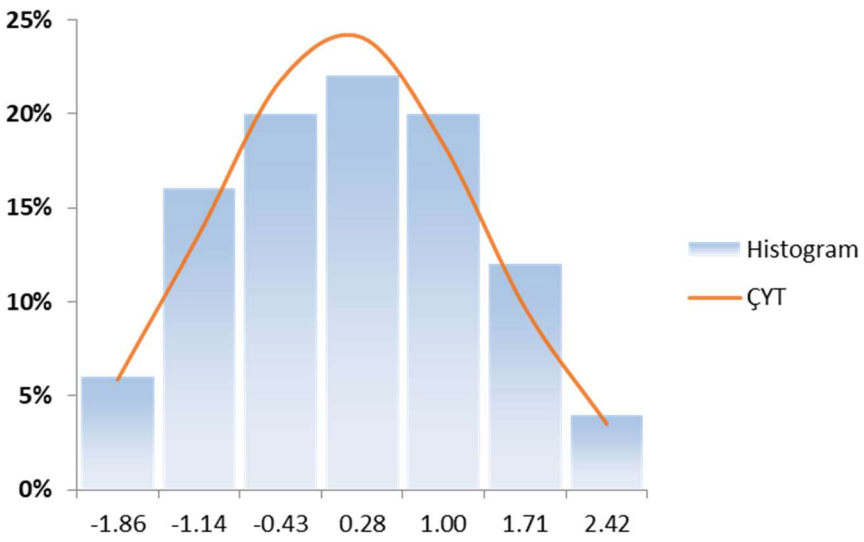
ÇYT ilişki tahmincisi ile rastgele değişkenlere ait olasılık dağılımı histogram temelli yöntemlere göre daha hızlı bir şekilde gerçek olasılık dağılımı değerlerine ulaşabilmektedir. Şekil 2.6'da -2.21 ile 2.78 aralığında rasgele alınan 50 değer için ÇYT ve Histogram dağılımına ait grafiksel gösterim verilmiştir. Şekil 2.6'da histogram dağılımı için örnekler eşit genişlik ayrıklaştırma yöntemiyle 7 (7 değeri 50 olan örnek sayısının karekökü alınarak hesaplanmıştır.) hücreye bölünmüştür. Y ekseni her bir hücrenin frekansını, X ekseni ise her bir hücre için en büyük ve en küçük örnek değerinin ortalamasını göstermektedir. ÇYT için ise çekirdek fonksiyonu olarak standart sapması h olan bir Gauss fonksiyonu kullanılmıştır. Şekil 2.6'dan da görüleceği üzere, ÇYT ile tahmin edilen olasılık yoğunluğu, histogram yöntemi ile elde edilen tahmine göre daha yumuşak geçişlere sahiptir ve bu sebepten ÇYT'ler olasılık dağılımının tahmininde daha yaygın kullanılır.

N adet rastgele değer içeren bir x ifadesi için çekirdek fonksiyonu olarak standart sapması h olan bir Gauss fonksiyonunda rastgele değişkenlerin bireysel olasılık yoğunluk fonksiyonlarının tahmini Bağıntı 2.26 ile hesaplanır.

$$f_h(x) = \frac{1}{N} \frac{1}{\sqrt{2\pi}h} \sum_i exp\left(-\frac{(x-x_i)^2}{2h^2}\right) \quad (2.26)$$

Aynı sayıda (N) rastgele değer içeren x ve y değişkenlerinin birleşik olasılık yoğunluk tahmini Bağıntı 2.27 ile hesaplanır.

$$f_h(x,y) = \frac{1}{N} \frac{1}{\sqrt{2\pi}h^2} \sum_i exp\left(-\frac{(x-x_i)^2 + (y-y_i)^2}{2h^2}\right) \quad (2.27)$$



Şekil 2. 6 Histogram ve ÇYT örnek grafiği

ÇYT ile *f(x), f(y)* ve *f(x,y)* fonksiyonları hesaplandıktan sonra KB ilişki tahmin değeri Bağıntı 2.28 ile bulunur.

$$KB(x,y) = \frac{1}{N}\sum_i \log\left(\frac{f(x_i,y_i)}{f(x_i)f(y_i)}\right) \tag{2.28}$$

Bu tez çalışmasında ÇYT'de çekirdek fonksiyon olarak Gauss fonksiyonu kullanıldığından bant genişliği olan *h* parametresi için Gauss fonksiyonun standart sapması kullanılmıştır.

BÖLÜM 3

GEN AĞI ÇIKARIMI

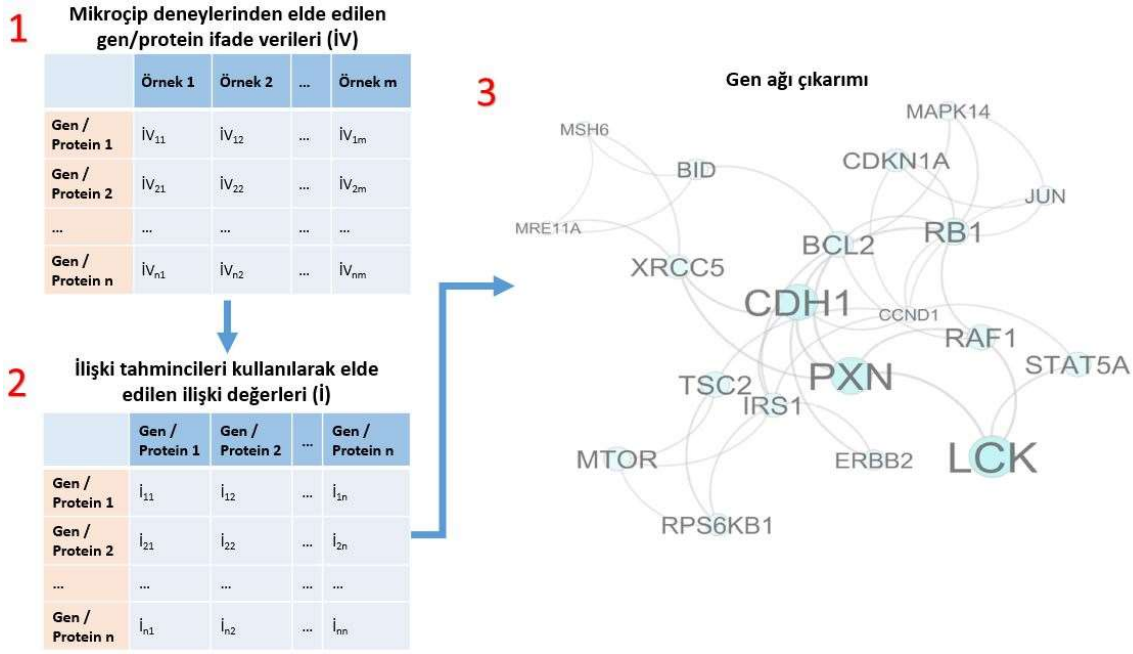
Gen/Protein ağı çıkarım algoritmaları biyoinformatik alanında önemli bir rol oynamaktadırlar ve genom çapında genler/proteinler gibi gen ürünlerindeki ilişkileri göstermek için literatürde yaygın bir şekilde kullanılmaktadırlar. Literatürde, GAÇ teknikleri; hastalıkla ilgili genlerin işlevlerinin bulunması, düzenleyici ve regüle genlerin tespiti, ilaç hedefleri ve ilgili hastalık için biyolojik belirteçler vb. gibi farklı amaçlar için yaygın olarak kullanılmaktadır.

Literatürde ağ çıkarımı için farklı yöntemler geliştirilmiş olup, bu yöntemlerin genel olarak nasıl çalıştığı üç adımda özetlenebilir.

- İlk olarak ağ çıkarımı yapılacak olan gen ifade veya proteomik veriler bir matris şeklinde girdi olarak alınır. Matriste bulunan her bir satır bir geni veya bir proteini belirtirken, her bir sütun bir örneği belirtir. Her bir hücre de, bir gene veya proteine ait ifade verisini gösterir. Gen ifade veya proteomik veriler farklı mikroçip deneylerinden elde edilir.
- İkinci adımda ise girdi olarak alınan gen ifade veya proteomik verileri kullanılarak bu genler veya proteinler arasındaki ilişki değerleri farklı ilişki tahmincileri kullanılarak hesaplanır. Literatürde bu adım genellikle ilişki tahmini (association estimation) olarak adlandırılır ve ilişki tahmini için genellikle korelasyon veya entropi tabanlı karşılıklı bilgi yöntemleri kullanılmaktadır. Bu adım sonucunda gen veya protein sayısına göre ilişki değerlerinin yer aldığı bir kare matris elde edilir.
- Son adımda ise bir önceki adımdan elde edilen ilişki değerlerine ait kare matrisi kullanılarak (Genler/proteinler arasındaki İlişki değerleri belirlenen bir eşik

değerinin altındaysa genellikle bu ilişkiler ağ çıkarımından önce elenir.) ağ çıkarım işlemi yapılır.

Ayrıca, ağ çıkarım yöntemlerinin çalışma mantığı Şekil 3.1'de şematik olarak verilmiştir.



Şekil 3. 1 Ağ Çıkarım Adımları

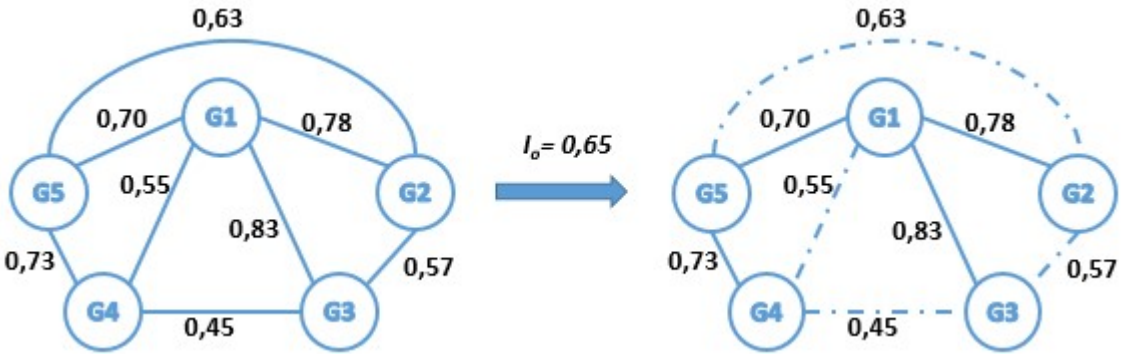
Ağ çıkarımı için farklı algoritmalar geliştirilmiş olup, bu algoritmalardan literatürde sıklıkla kullanılanlar bu bölümde ayrıntılı olarak anlatılmıştır.

3.1 RELNET (RELevance NETwork)

Literatürde GAÇ için geliştirilen KB tabanlı ilk algoritma RELNET'tir ve geliştirilen diğer KB tabanlı GAÇ algoritmalarının temelini oluşturmaktadır [13]. RELNET algoritmasında ilişki tahmin işleminden sonra basit bir işlemle istatistiksel olarak önemsiz olduğu düşünülen ilişki değerlerine sahip aday bağlantılar elenir. Bunun için genler arasındaki ilişki değeri belirli bir eşik değerini temsil eden I_0'*ın* altındaysa ilgili genler arasındaki bağlantı kopartılır, değilse de ağda bırakılır. Bu işlemin bir temsili Şekil 3.2'de verilmiştir.

Şekil 3.2'de beş genden oluşan küçük bir gen ağı için RELNET'in önemsiz genleri elemesine örnek verilmiştir. Eşik değeri $I_0 = 0{,}65$ olarak seçilmiş ve bu değerin altındaki skorlar ağdan elenerek sonuçta gen ağı elde edilmiştir. Eşik değeri permütasyon testi gibi

çeşitli istatistiksel testler ile belirlenebilir. Permütasyon testinin nasıl yapıldığı ile ilgili detaylı bir bilgi Bölüm 3.7'de verilmiştir.



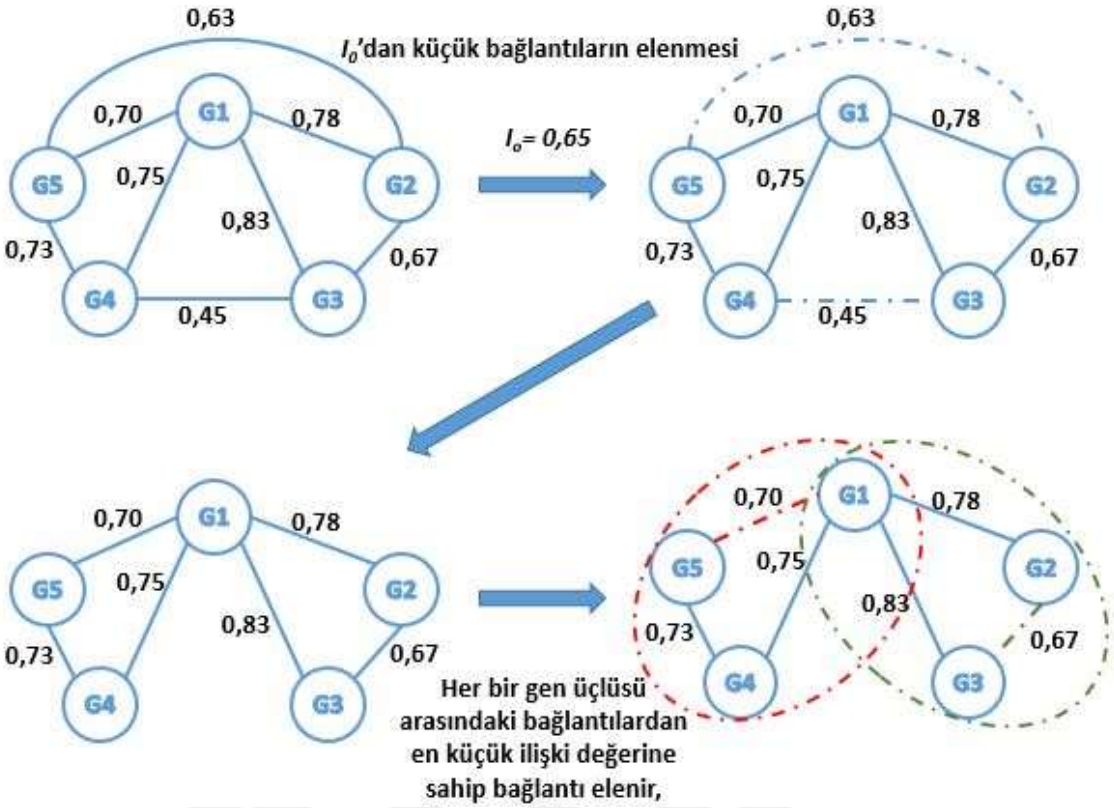
Şekil 3. 2 RELNET (I0=0,65 için) önemsiz bağlantıların elenmesi

3.2 ARACNE (The motivation of the Algorithm for the Reconstruction of Accurate Cellular Networks)

Literatürde birçok çalışmada kullanılmış olan ARACNE [5] GAÇ algoritmasında, girdi olarak verilen gen ifade/proteomik verileri kullanılarak ilişki tahmin değerlerinin bulunduğu ilişki matrisi elde edilir ve aday bağlantılar arasından istatistiksel olarak önemsiz olanların elenmesi için sonraki adımlara geçilir.

Öncelikle ARACNE'de RELNET'te uygulandığı üzere permütasyon testi ile belirlenen bir I_0 eşik değerinin altındaki skorlar elenir. Sonraki adımında Veri İşleme Eşitsizliği (VİE, Data Processing Inequality) işlemi uygulanarak potansiyel olarak dolaylı (indirect) olduğu düşünülen bağlantıların elenmesi sağlanır. Dolaylı bağlantının nasıl oluştuğu ve tespit edilerek elenmesi gerektiğinin nedeni ayrıntılı olarak Bölüm 2.2'de anlatılmıştır.

VİE işlemi, A ve B genlerine doğrudan bağlı olmayıp, C geni üzerinden dolaylı olarak bağlı olmaları durumunda $KB(A,B) \leq \min[KB(A,C),KB(C,B)]$ koşulunu sağlamaları gerektiğini varsayar. Bunun için de I_0 eşik değerinden büyük olan her gen üçlüsünü inceler. Gen üçlülerinin ikili kombinasyonları arasından en küçük ilişki değerinin dolaylı bağlantıdan kaynaklandığını varsayar ve bu ilişki değerine sahip bağlantıyı ağdan eler. Son olarak da geriye kalan bağlantıları kullanarak gen ağını oluşturur. ARACNE GAÇ algoritmasının işlem adımları beş genli küçük bir ağ için Şekil 3.3'te verilmiştir.



Şekil 3. 3 ARACNE GAÇ algoritmasında önemsiz ve dolaylı bağlantıların elenmesi

3.3 CLR (The Context Likelihood of Relatedness network method)

Temelde RELNET algoritmasını baz alan CLR [6] algoritması her bir gen çifti için KB'yi hesaplar ve bu KB değerlerinin ampirik dağılımıyla ilgili bir skor elde eder. Öncelikle X_i ve X_j genleri için KB(x_i, x_j) değeri yerine, k:1'den n'e kadar olmak üzere KB(x_i, x_k) değerlerinin μ_i ortalamasını ve σ_i standart sapmasını belirtmek üzere z_i skoru bağıntı 3.1 kullanılarak belirlenir.

$$z_i = Max(0, \frac{(KB(x_i, x_j) - \mu_i)}{\sigma_i}) \quad (3.1)$$

Benzer şekilde z_j değeri de belirlendikten sonra CLR algoritması tarafından genler arasındaki ilişki skoru Bağıntı 3.2 kullanılarak hesaplanır. Belirlenen z skorları kullanılarak gen ağı oluşturulur.

$$z_{ij} = \sqrt{{z_i}^2 + {z_j}^2} \quad (3.2)$$

3.4 MRNET (The Minimum Redundancy NETworks)

MRNET [7], makine öğrenmesinde özellik seçimi için kullanılan MRMR (azami uygunluk/minimum artıklık, maximum relevance/minimum redundancy) yöntemine dayalı bir gen ağı çıkarımı yöntemidir. "Azami uygunluk", hedef değişkenle (genle) yüksek karşılıklı bilgi içeren özellikleri (genleri) seçmektir. "Minimum artıklık", bu özelliklerin aralarındaki karşılıklı bilgilerin olabildiğince düşük olacağı şekilde seçilmesidir. Olsen vd. [3], MRMR'yi hedef bir *i* geni ve özellikleri de geride kalan diğer genler olmak üzere geliştirdikleri ağ çıkarım algoritmasında uygulamışlardır. Algoritmanın adımları şunlardır:

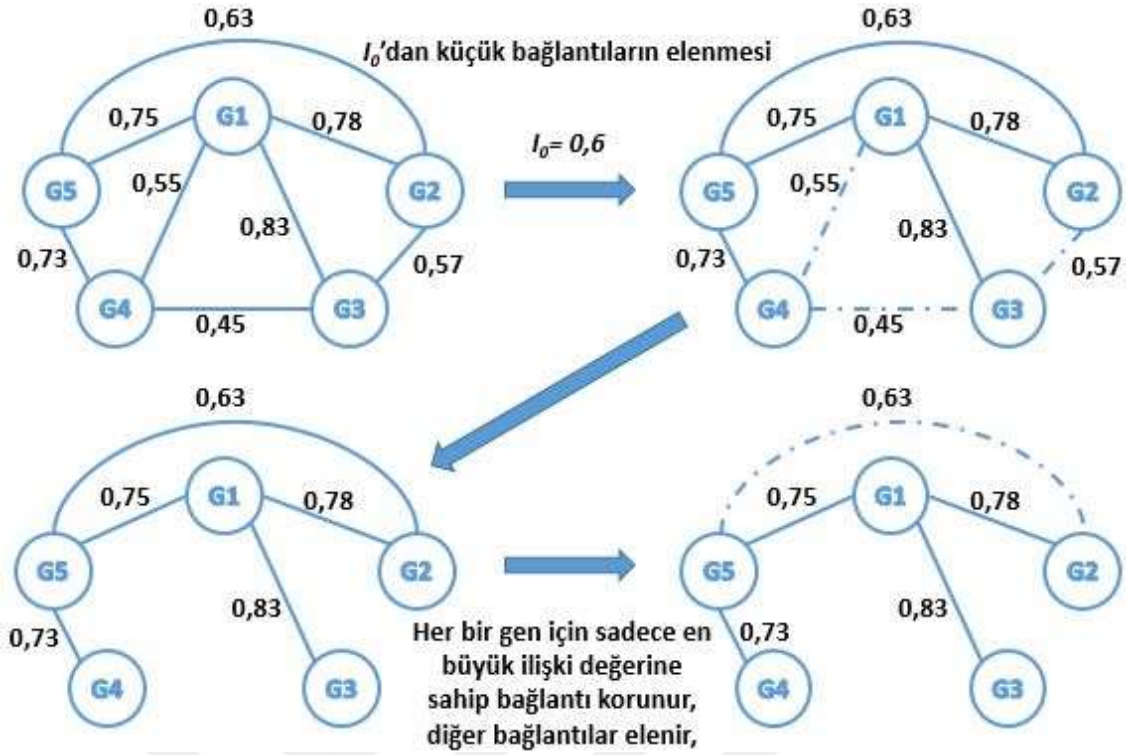
- Hedef *y* geni ile en yüksek KB'ye sahip olan x_i geni seçilir.
- Sonraki x_j genini seçmek için X_s: diğer kalan genleri; u_j: azami uygunluk ve r_j: minimum artıklık değerini belirtmek üzere Bağıntı 3.3'teki s_j değerini maksimum yapan gen seçilir.

$$s_j = u_j - r_j = KB(x_i, x_j) - \frac{1}{|X_s|} \sum_{x_k \in X_s} KB(x_j, x_k) \tag{3.3}$$

- Tüm hedef genler için işlem tekrarlanır.

3.5 C3NET (The Conservative Causal Core NETwork)

C3NET [10] algoritması ARACNE algoritmasına benzeyen bir algoritma olup, öncelikle girdi olarak verilen gen ifade/proteomik verilerini kullanarak ilişki matrisini hesaplar. Ardından istatistiksel olarak anlamsız olan bağlantıların elenmesi için, belirlenen bir I_0 değerinin altındaki bağlantılar elenir. Ardından ARACNE GAÇ algoritmasında kullanılan VİE yerine farklı bir yaklaşım uygular. Bu yaklaşımda potansiyel olarak dolaylı olduğu düşünülen bağlantıların elenmesi için her bir genin tüm olası ilişkileri incelenir. Her bir gen için sadece en büyük ilişki değerine sahip bağlantı korunur, diğer bağlantılar elenir. Burada dikkat edilmesi gereken nokta ağın simetrik ve yönsüz yapısının korunması gerektiğidir. C3NET'in işlem adımları beş genli küçük bir ağ için Şekil 3.4'te verilmiştir.



Şekil 3. 4 C3NET GAÇ algoritmasında önemsiz ve dolaylı bağlantıların elenmesi

Şekil 3.4'teki örnekte de görüldüğü üzere ilk olarak 0,6 olarak alınan I_0 eşik değerinin altındaki bağlantılar elenmiştir. Sonrasında da her bir gen için sadece en büyük ilişki değerine sahip bağlantı korunmuş ve diğerleri elenmiştir. Örneğin G1 geni için en büyük ilişki değeri G3 geniyle olan bağlantıya aittir (0,83); G1 geni için diğer bağlantılar elenmiştir. G2 geni için en büyük ilişki değeri G1 geni ile olan bağlantıya aittir (0,78); G2 geni için diğer bağlantılar elenmiştir. Ağdaki simetrik yapının korunması için G1-G2 genlerinin arasındaki bağlantı iki yönlü olarak yeniden kurulmuştur. G3 geni için en büyük ilişki değeri daha önce bulunan G1 geni ile olan bağlantıdır (0,83). G4 geni için en büyük ilişki değeri G5 geniyle olan bağlantısıdır (0,73). G5 geni için en büyük ilişki değeri G1 geni ile olan bağlantıdır ve G5 geni için diğer bağlantılar elenmiştir. Ağdaki simetrik yapının korunması için G4-G5 genlerinin arasındaki bağlantı iki yönlü olarak yeniden kurulmuştur.

Görüldüğü üzere C3NET GAÇ algoritmasının diğer yöntemlere karşı en büyük artısı, daha çok eleme yaparak daha seyrek ancak daha güvenilir gen ağı elde edebilmesidir. C3NET ile bulunan bağlantıların gerçek biyolojik ağda bulunma olasılığı, diğer GAÇ algoritmalarıyla bulunan bağlantılara göre daha yüksektir. Buna karşın çok fazla eleme

işlemi yaptığı için biyolojik olarak anlamlı ilişkileri de kopartma olasılığının yüksek olması algoritmanın en zayıf yönüdür.

3.6 WGCNA - Ağırlıklı Korelasyon Ağ Analizi

Literatürde birçok araştırmada kullanılmış olan WGCNA [22] ağ çıkarım yöntemi hedef hastalıkla ilgili yüksek korelasyona sahip gen gruplarını (modül, alt-ağ) ve bu gen gruplarındaki yüksek bağlanabilirlik değerlerine sahip merkez (hub) genleri tespit etmek için kullanılmaktadır. Algoritmanın temel adımları aşağıda verilmiştir.

- İlk olarak, gen ifade/proteomik verileri kullanılarak veri kümesindeki genlerin/proteinlerin arasındaki ilişki değerlerini içeren ilişki matrisi (adjacency matrix – Genel olarak Pearson KKD ve Spearman KKD ile hesaplanır.) oluşturulur.

  Verilen bir *X* gen ifade matrisi göz önüne alındığında; *kor*, korelasyonu belirtmek üzere, *i* ve *j* genleri arasındaki ortak ifade benzerliği b_{ij}, Bağıntı 3.4'le tanımlanabilir.

  $$b_{ij} = |kor(i,j)| \tag{3.4}$$

  WGCNA'da *i* ve *j* genleri arasındaki benzerlik (adjaceny) değeri (A_{ij}) bir *β* (*β ≥ 1 olmak üzere*) katsayısı kullanılarak Bağıntı 3.5 ile belirlenir.

  $$A_{ij} = b_{ij}^{\beta} \tag{3.5}$$

  β katsayısı kullanılarak güçlü bağıntı değerleri daha da güçlendirilmiş ve zayıf bağlantılar daha da zayıflatılmış olur.

- Birbiriyle yüksek ilişkili gen/protein gruplarını tespit etmek için literatürde farklı çalışmalarda [55], [56] başarılı bir şekilde uygulanmış olan topolojik örtüşme ölçütü (topological overlap measure - TOM) [57] kullanılarak ilişki matrisi elde edilir. TOM kullanılmasının sebebi sadece korelasyon skorlarına bakılarak oluşturulacak gen ağının ağ topolojisini oluşturmada yetersiz olduğunun düşünülmesidir. Bu yetersizliğin giderilmesi için korelasyon değerlerinin ağ topolojisi seviyesine yansıtılması TOM ile sağlanılmaktadır. TOM'a dayalı ilişki matrisi A_{TOM}, orijinal ilişki matrisi $A^{Orijinal}$'i topolojik örtüşme matrisine eşlemek için Bağıntı 3.6'yı kullanır.

$$A_{TOM}\left(A^{Orjinal}\right) = \frac{\sum_{l\neq j,i} A_{il}^{Orjinal} A_{lj}^{Orjinal} + A_{il}^{Orjinal}}{\min\left(\sum_{l\neq i} A_{il}^{Orjinal}, \sum_{l\neq j} A_{jl}^{Orjinal}\right) - A_{ij}^{Orjinal} + 1} \quad (3.6)$$

- Elde edilen benzerlik matrisi kullanılarak birbiriyle ilişkili protein kümelerini belirlemek için hiyerarşik kümeleme yöntemiyle kümeleme ağacı (clustering tree) oluşturulur (hclust).
- Oluşturulan bu ağaç üzerinde dinamik ağaç kesme yöntemi ile anlamlı olabilecek modüller tespit edilmeye çalışılır (cutreeDynamic).
- Son olarak her bir modüldeki yüksek bağlı merkez protein/gen bulunur (chooseOneHubInEachModule).

3.7 Eşik Değeri *(I_0)* Belirlemede Permütasyon Testi Kullanımı

İstatistiksel bir test tekniği olan Permütasyon testi literatürde I_0 eşik değerini belirlemek için başta RELNET olmak üzere ARACNE ve C3NET GAÇ algoritmalarında da kullanılmıştır. Permütasyon testi istatistiksel olarak önemli olan bağlantıların korunmasını ve önemsiz olan bağlantıların elenmesi için bir eşik değerinin belirlenmesinde kullanılmaktadır.

Permüstasyon testinin nasıl yapıldığını kısaca açıklamak gerekirse: Belirli bir adım sayısı boyunca gen ifade değerlerinin tutulduğu matris rastgele karıştırılarak yeni veri kümeleri elde edilir. Bu veri kümelerinin ilişki skorları bulunup bir vektöre yerleştirilir. Belirlenen istatistiksel bir α önem seviyesine göre bu vektörün belli bir noktasındaki değer ((1-α)L indisindeki değer) I_0 eşik değeri olarak atanır.

Buna göre permütasyon testinde yapılan işlemleri özetleyecek olursak; Belirlenmiş bir adım sayısınca gen ifade veya proteomik verilerine ait örneklerin bulunduğu veri seti karıştırılır ve yeni veri setleri elde edilir. Bu veri setlerinin ilişki değerleri korelasyon veya KB kullanılarak hesaplanarak bir *İ* vektörüne yerleştirilir. Daha önceden belirlenmiş olan bir *s* önem seviyesine (significance level) göre bu vektörün belirli bir noktasındaki değer I_0 eşik değeri olarak atanır.

Bu işlemin kaba kodu aşağıda verilmiştir.

1. İlişki değerlerini tutmak için boş bir *İ* vektörü tanımlanır.
2. For (a = 1'den bellirli_bir_adim_sayısına_kadar)

a. Gen ifade veya proteomik verilerine ait örneklerin bulunduğu veri setindeki değerler rastgele bir şekilde karıştırılır ve yeni veri seti elde edilir.

b. Yeni veri setinin korelasyon veya KB değerleri hesaplanır.

c. Hesaplanan bu yeni değerler *İ* vektörüne eklenir.

3. For döngüsü biter

4. *İ* vektörü küçükten büyüğe doğru sıralanır.

5. Bir *s* önem seviyesi değeri belirlenir.

6. *İ* vektörünün boyutu *B* olmak üzere; bu vektörün *(1-s)B* 'nci elemanı I_0 eşik değeri olarak atanır.

BÖLÜM 4

GEN AĞI ÇIKARIM ALGORİTMALARINDA KORELASYON TABANLI İLİŞKİ TAHMİNCİLERİN ANALİZİ

Bu bölümde Bölüm 3'te bilgilerini verdiğimiz GAÇ yöntemlerinden ARACNE, MRNET, CLR ve C3NET GAÇ algoritmaları kullanılarak korelasyon tabanlı (Pearson KKD, Spearman KKD ve Kendall KKD) ilişki tahmincilerin proteomik veriler üzerindeki analiz başarımı incelenmiştir.

4.1 Kullanılan Veri Seti ve Analiz İşlemleri

Veri seti olarak TCPA [28] tarafından ters faz protein dizileme (Reverse phase protein array – RPPA [58]) yöntemi kullanılarak oluşturulmuş olan 16 kanser türüne ait proteomik verileri kullanılmıştır ve kullanılan veriler 3'ncü seviyeye aittir. Kullanılan veri setinin özellikleri (Kanser türü, kısaltması, örnek sayısı) Çizelge 4.1'de verilmiştir. Veri setlerinde her bir hastaya ait 190 proteinin ifade verisi yer almaktadır.

Her bir hastalık türündeki örneklerde ortak olan bu 190 proteinin literatürde yer alan protein-protein ilişkilerini tespit etmek için Şenbabaoğlu vd.'nin [20] yapmış oldukları çalışmada olduğu gibi Pathway Commons (PC) [59] kullanılmış ve 177 proteinin birbiriyle etkileşim içinde olduğu görülmüştür. Literatürdeki ilişkileri bulmak için PC [59]'den indirilen BioPAX [60] dosyası içinde yer alan protein-protein ilişkileri Babur vd. [61] tarafından geliştirilen Java uygulaması kullanılarak tespit edilmiştir. Etkileşim içinde olduğu bulunan 177 proteinin PC'deki toplamdaki etkileşim sayısı 1.870 olarak bulunmuştur. Çalışmada bu bulgular ile ortak bilgi tahmin edicilerin başarımları değerlendirilmiştir. Ek olarak, PC'den elde etmiş olduğumuz 19.000'e yakın proteine ait

bir buçuk milyondan fazla protein-protein ilişkisinin başka araştırmalarda da kullanılabilmesi için bir web uygulaması geliştirilmiş olup, Ek A-4'te detayları verilmiştir.

PC, biyolojik-yol (biological-pathway) ve etkileşim verilerini (protein-protein, gen-gen vb.) toplamayı ve yaymayı amaçlamaktadır. Veriler ortak veritabanlarından (Reactome, BioGRID, BIND, MSigDB vb. 25 farklı veritabanı) toplanır ve BioPAX [60] standardında temsil edilir. Pathway Commons BioPAX standartlarında çeşitli biyolojik kavramların ayrıntılı bir temsilini sunabilmektedir. Bu kavramlar: biyokimyasal reaksiyonlar, gen düzenleyici ağlar, genetik etkileşimler, taşıma ve kataliz olayları, proteinlerin etkileşim içinde olduğu DNA, RNA, küçük moleküller ve komplekslerdir.

Çizelge 4. 1 TCPA'dan alınan veriler için tümör tipleri, çalışmada kullanılan kısaltmalar ve her bir tümör tipi için örnek sayısı

Kanser Türü	**Kısaltma**	**Örnek Sayısı**
Mesane Kanseri (Bladder urothelial carcinoma)	BLCA	127
Göğüs Kanseri (Breast invasive carcinoma)	BRCA	747
Kolon Kanseri (Colon adenocarcinoma)	COAD	334
Beyin Kanseri (Glioblastoma multiforme)	GBM	215
Baş ve boyun Kanseri (Head and neck squamous cell carcinoma)	HNSC	212
Böbrek Kanseri (Kidney clear cell carcinoma)	KIRC	454
* Beyin Kanseri (Brain Lower Grade Glioma)	LGG	260
Akciğer Kanseri (Lung adenocarcinoma)	LUAD	237
Akciğer Hücre Kanseri (Lung squamous cell carcinoma)	LUSC	195
Yumurtalık Kanseri (Ovarian serous cystadenocarcinoma)	OVCA	412
*Prostat Kanseri (Prostate adenocarcinoma)	PRAD	164
Rektum (Kalın bağırsak) Kanseri (Rectum adenocarcinoma)	READ	130
* Deri Kanseri (Skin Cutaneous Melanoma)	SKCM	207
*Mide Kanseri (Stomach adenocarcinoma)	STAD	299
*Tiroid Kanseri (Thyroid carcinoma)	THCA	375
Rahim Kanseri (Uterine corpus endometrioid carcinoma)	UCEC	404
	Toplam	4.772

* ile işaretlenen kanser türlerinde hastalara ait sadece 189 proteomik verisi vardır. (ARIDIA proteinini içermezler)

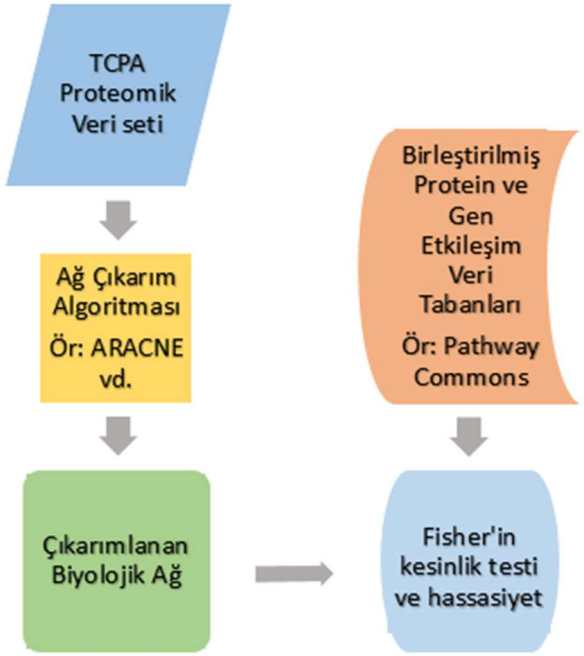
Literatürde sentetik veriler kullanılarak yapılan araştırmalarda GAÇ yöntemlerinin başarımlılarının değerlendirmeleri, genellikle alıcı işletim karakteristikleri altındaki alan (AUROC- area under receiver operating characteristics), F-skor veya kesinlik-hassasiyet eğrisi altındaki alan (AUPR- area under precision-recall curve) gibi metrikler kullanılarak gerçekleştirilir. Bununla birlikte, bizim yaklaşımımızda, literatürden elde edilen etkileşim

veritabanları sentetik verilerdeki gibi tam olarak doğru ilişkileri veremeyeceği için istenileni karşılamaya bilir. Bu da çok sayıda hatalı onaylanmış (False Positive) veriye neden olabilir ve bu nedenle F-skor, AUROC ve AUPR bizim çalışmamız için uygun ölçüt değildir. Bu yüzden geleneksel başarım ölçme kriterlerinden kesinlik (precision) skoru çalışmamızı değerlendirmek için daha uygun bir metriktir. Kesinlik skoru, literatürden çıkarılmış tüm etkileşimlere göre doğru çıkarılan etkileşimlerin oranını verir.

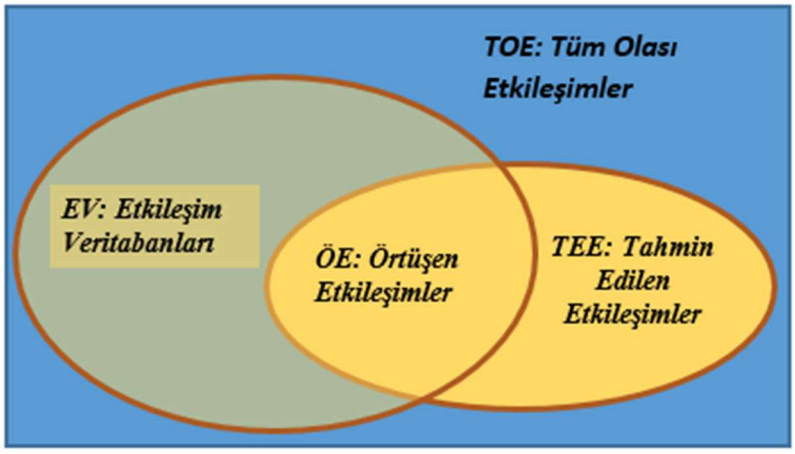
GAÇ algoritmalarının çıkarım performanslarını değerlendirmek için önerilen karşılaştırma işlemi Şekil 4.1'de gösterilmiştir. GAÇ algoritmalarıyla TCPA'dan alınan veri setleri kullanılarak çıkarımlanan protein ağının doğruluğunu tespit etmek için birleştirilmiş protein ve gen etkileşim veri tabanlarından alınan protein-protein ilişkileri kullanılmıştır. PC kullanılarak elde edilen etkileşimlerin tamamen farklı biyolojik terimlerin etkileşimlerini gösterebileceğini, ancak protein çiftleri arasında anlamlı bir korelasyon değerine sahip olan ve diğer biyolojik durumlardan elde edilmiş, deneysel olarak ispatlanmış bir etkileşimi tahmin etmenin, bu etkileşimi doğru pozitif (TP) olarak kabul etmek için güçlü bir varsayım sağladığını kabul ederek GAÇ algoritmalarının performans hesaplanmasında kullanılmıştır. Nitekim diğer biyolojik terimlerin sözde doğrulanmış olan bu etkileşimlerinin, söz konusu veri setinin biyolojik teriminde de etkileşime girip girmediğini ve bunun ilgi konusu biyolojiye yeni bakış açıları kazandırabileceğini doğrulamak ve kanıtlamak adına deney yapmak için en güçlü aday olarak ortaya çıkmaktadır.

GAÇ yöntemlerinin performansını karşılaştırmak için literatürde yaygın olarak kullanılan istatistiksel bir yöntem olan Fisher'ın kesinlik testi (FKT) [62] ve kesinlik (precision) değerleri kullanılmıştır. FKT, Fury ve vd.'nin [63] yaptığı çalışmada anlatıldığı üzere örtüşme olasılığının değerini (p-değeri) hipergeometrik dağılım ile hesaplar. Hesaplanan p-değeri belirlenen bir eşik değerinden küçükse çıkarımlanan ağdaki proteinler arasındaki ilişki istatistiksel olarak rastgelelikten uzak ve biyolojik olarak anlamlı kabul edilir. FKT'nin GAÇ yöntemlerinin performanslarını karşılaştırmak için kullanımını açıklamak amacıyla Şekil 4.2 ve Çizelge 4.2 verilmiştir. Şekil 4.2'de görülebileceği gibi, verilen iki ağ göz önünde bulundurulduğunda (Etkileşim Veritabanları ve Tahmin Edilen Etkileşimler) veri kümesindeki Tüm Olası Etkileşimler içinde Örtüşen Etkileşimlerin

istatistiksel olarak anlamlı olup olmadığını belirlemek için çalışma [63]'tan uyarlanmış olan hipergeometrik test ile Bağıntı 4.1 kullanılarak p-değeri hesaplanır.



Şekil 4. 1 GAÇ algoritmalarının performans değerlendirme işleminin blok diyagramı



Şekil 4. 2 Örtüşme analizi

Şekil 4. 2'de *TOE,* tüm olası protein-protein etkileşimlerini (15.576 adet); *EV,* birleştirilmiş veritabanlarından elde edilen etkileşimlerden bizim incelediğimiz proteinlere ait toplam etkileşim sayısını (1.870 adet); *TEE,* GAÇ yöntemleriyle tahmin edilen etkileşimleri; *ÖE*, *EV* ile *TTE* arasındaki örtüşen etkileşimleri belirtmektedir.

$$P(\ddot{O}E) = \frac{\binom{TEE}{\ddot{O}E}\binom{TOE-TEE}{EV-\ddot{O}E}}{\binom{TOE}{EV}} = \frac{TEE!\quad (TOE-TEE)!\quad EV!\quad (TOE-\quad)!}{\ddot{O}E!\quad (TEE-\ddot{O}E)!\quad (EV-\ddot{O}E)!\quad (TOE-TEE-EV+\ddot{O}E)!\quad TOE!} \tag{4.1}$$

Bağıntı 4.1'de *n* ve *k* tam sayı olmak üzere $\binom{n}{k}$ binom açılımını belirmektedir.

Çizelge 4. 2 FKT parametreleri

	Literatürde yer alan	Literatürde yer almayan
GAÇ ile belirlenmiş	*ÖE*	*TEE - ÖE*
GAÇ ile belirlenememiş	*EV - ÖE*	*TOE – TEE – EV + ÖE*

Çizelge 4.2, Şekil 4.1'de gösterilen bölgeleri özetler ve FKT'de bir girdi olarak kullanılır.

Değerlendirme sürecinde kullanılan bir diğer değerlendirme metriği kesinlik (precision) skorudur ve Bağıntı 4.2 ile belirtilmiştir.

$$Kesinlik = \frac{Gerçek\ Pozitif\ (TP)}{Gerçek\ Pozitif\ (TP) + Yanlış\ Pozitif (FP)} \quad (4.2)$$

Kesinlik skoru, doğru çıkarılan etkileşimlerin literatürden çıkarılmış tüm etkileşimlere göre oranını verir.

Test işlemleri için MINET [8], C3NET [10] R paketleri kullanılmıştır. CLR ve MRNET algoritmaları herhangi parametre almamaktadır. ARACNE ve C3NET algoritmalarının başarımını incelerken almış oldukları parametreler için farklı değerler test edilmiştir. Çizelge 4.3'te incelemiş olduğumuz GAÇ algoritmalarında kullanılan parametre değerleri ile bu parametrelerin en uygun değerleri ve algoritmaların karmaşıklık (complexity) değerleri verilmiştir.

Çizelge 4. 3 GAÇ algoritmalarının parametreleri ve en uygun parametre değerleri

GAÇ Algoritması	Parametre açıklaması	Test edilen değerler	En uygun değer	Karmaşıklık
ARACNE [5]	Bir ilişki elenirken kullanılan eşik değerini belirten sayısal değer. (eps)	eps=(0,0001; 0,001; 0,01; 0,1)	eps=0,0001	$O(N^3)$
CLR [6]	-	-	-	$O(N^2)$
MRNET [7]	-	-	-	$O(N^2)$
C3NET [10]	Örnekleme dağılımını için verileri yeniden örneklemede yineleme sayısı. (itnum) İstatistiksel anlamlılık eşiği (alpha)	itnum=[10; 100; 1000] alpha=(0,5; 0,1; 0,05; 0,01; 0,005; 0,001)	itnum=10; alpha=0,1	$O(N^2)$

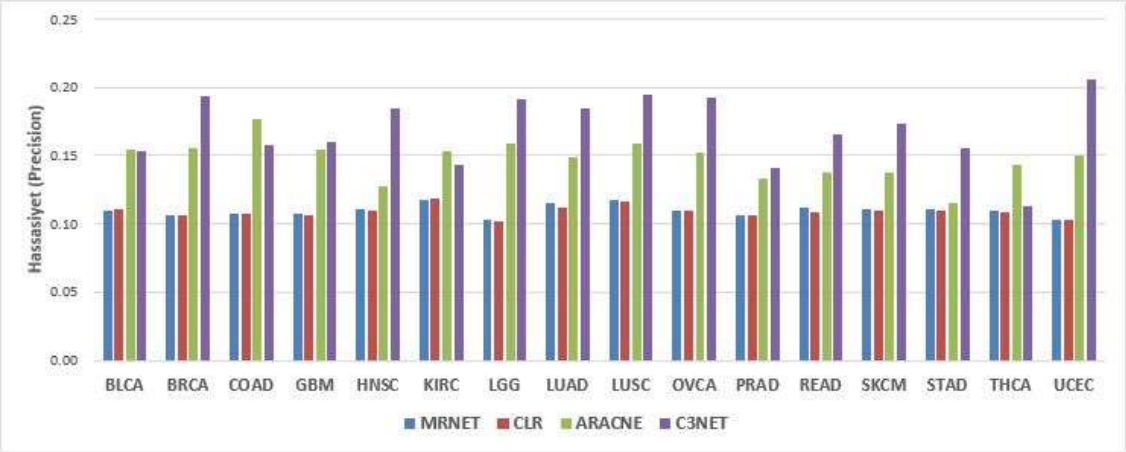
4.2 Analiz Sonuçları

Bu bölümde dört farklı GAÇ algoritmasının başarımı TCPA'dan alınan proteomik veri setleri kullanılarak incelenmiştir. Ayrıca, proteinler arası ilişkiyi tahmin etmek için üç

farklı korelasyon yöntemi kullanılmış ve bu yöntemlerin de GAÇ yöntemlerinin başarımı üzerindeki etkileri incelenmiştir. GAÇ yöntemlerinin başarımları Bölüm 4.1'de ayrıntıları verilmiş olan Kesinlik skorları ve FKT ile hesaplanan p-değeri kullanılarak değerlendirilmiştir. Alt bölümlerde korelasyon tabanlı ilişki çıkarım yöntemlerine göre GAÇ algoritmalarının başarımları değerlendirilmiştir.

4.2.1 Pearson KKD'ye göre GAÇ Yöntemlerinin Başarım Sonuçları

Pearson KKD için GAÇ algoritmalarının başarımları kanser türüne göre değişmekte olup başarım sonuçları kesinlik skorlarına göre Şekil 4.3'te verilmiştir. Kesinlik skorlarına göre yapılan değerlendirmede on iki farklı kanser türü için (BRCA, GBM, HNSC, LGG, LUAD, LUSC, OVA, PRAD, READ, SKCM, STAD ve UCEC) C3NET algoritması en yüksek kesinlik değerlerini elde etmiştir. C3NET algoritmasından sonra en iyi kesinlik değerlerini 4 farklı kanser türü için (BLCA, COAD, KIRC ve THCA) ARACNE algoritması elde etmiştir. C3NET algoritmasının kesinlik skorlarının diğer algoritmalara göre daha yüksek olmasının nedeni ilişki skorları üzerinde daha çok eleme yaparak daha seyrek ancak daha güvenilir gen ağı elde edebilmesidir. ARACNE algoritmasının da başarılı kesinlik skorları elde etmesinin sebebi VİE işlemi uygulanarak potansiyel olarak dolaylı (indirect) olduğu düşünülen bağlantıları elemesidir. Bu durumun daha iyi anlaşılması için BRCA veri seti için her bir GAÇ algoritması tarafından tahmin edilen ilişkilerle, bu tahminlerden gerçek pozitif (literatürdeki veritabanları ile örtüşen) olanların sayıları Çizelge 4.4'te verilmiştir.



Şekil 4. 3 Veri setleri için Pearson KKD'ye göre GAÇ algoritmalarının başarımı

Çizelge 4.4'ten de görüldüğü üzere CLR ve MRNET algoritmalarının ilişki tahmin sayıları ile bu tahminlerden gerçek pozitif olanların sayıları birbirlerine yakındır. Bu iki algoritmanın yüksek sayıda ilişki tahmini yapmasına karşın bu tahminlerdeki gerçek pozitif sayıları yeteri kadar yüksek olmadığı için kesinlik skorları düşük olmuştur. Buna karşın ARACNE ve C3NET algoritmaları daha az sayıda ilişki tahmini yapmalarına karşın bu tahminlerdeki gerçek pozitiflerin daha isabetli olması dolayısıyla daha yüksek kesinlik değerlerine erişmişlerdir.

Çizelge 4. 4 BRCA veri seti için GAÇ algoritmaları ile tahmin edilen etkileşimlerin sayıları (Pearson KKD ile) ve kesinlik değerleri

	Örtüşen Etkileşimler	**Tahmin Edilen Etkileşimler**	**Kesinlik Skoru**
MRNET	768	7213	0,10647
CLR	774	7243	0,10686
ARACNE	51	329	0,15502
C3NET	30	155	0,19355

Ayrıca, Pearson KKD kullanılarak GAÇ algoritmalarınca yapılmış olan bu tahminlerin istatistiksel olarak anlamlılığını değerlendirmek için Bağıntı 4.1 kullanılarak elde edilen p-değerleri Çizelge 4.5'te verilmiştir. Çizelge 4.5'te koyu yazılı değerleri içeren hücreler, ilgili kanser türü için istatistiksel olarak en anlamlı p-değerini göstermektedir. Şöyle ki en düşük p-değerine sahip olan tahmin istatistiksel olarak rastgelelikten en uzak tahmindir ve dolaysıyla en anlamlı değeri ifade etmektedir.

Bu sonuçlardan kesinlik değerleri ile p-değerleri arasında bir tutarsızlık olduğu düşünülebilir fakat aslında çoğu p-değeri istatistiksel olarak anlamlılığı ifade eden ve daha önceden belirlediğimiz 0,05 eşik değerinden daha küçüktür. MRNET ve CLR algoritmalarının daha düşük p-değerlerine sahip olmasının nedenini daha iyi anlatabilmek için Çizelge 4.6 hazırlanmıştır. Çizelge 4.6'dan görüleceği üzere literatürdeki tüm etkileşimlerin sayısı tüm olası etkileşimlerin sayısının sadece %12'sine denk gelmektedir. C3NET ve ARACNE algoritmaları tarafından tahmin edilen etkileşim sayısı CLR ve MRNET algoritmalarına karşın çok daha düşük kaldığı için, istatistiksel olarak anlamlı p-değerlerine sahip olmalarına karşın MRNET ve CLR algoritmalarından daha yüksek p-değerlerine sahiptirler. Fakat bu değerler 0,05 olarak belirlenmiş olan anlamlılık değerinden küçük olduğu için istatistiksel olarak anlamlı kabul edilir.

Çizelge 4. 5 Pearson KKD kullanılarak GAÇ algoritmalarınca yapılmış olan tahminlerin FKT p-değerleri

	MRNET	CLR	ARACNE	C3NET
BLCA	**3,1054E-04**	9,4193E-04	5,1464E-02	2,1582E-02
BRCA	**1,2463E-06**	2,3214E-06	5,8621E-02	8,4685E-03
COAD	**2,3598E-06**	3,7872E-06	1,8943E-03	1,3975E-02
GBM	2,1755E-06	**9,8068E-08**	5,5150E-02	1,4436E-02
HNSC	4,5470E-04	**1,0749E-04**	6,1926E-01	1,5154E-02
KIRC	3,2183E-01	5,3617E-01	**3,7488E-02**	3,8953E-02
LGG	1,4051E-08	**2,2098E-09**	2,7949E-02	9,1236E-03
LUAD	7,5555E-02	**4,5586E-03**	9,5673E-02	1,5154E-02
LUSC	4,0028E-01	2,1657E-01	1,8911E-02	**6,4595E-03**
OVCA	8,8880E-04	**8,6118E-04**	3,7298E-02	8,7655E-03
PRAD	**4,3797E-06**	**4,3797E-06**	4,6044E-01	3,8768E-02
READ	2,9629E-03	**9,0397E-05**	3,2162E-01	8,3551E-03
SKCM	7,5360E-04	**2,7415E-04**	2,7065E-01	4,7036E-02
STAD	2,9071E-03	**2,5643E-04**	8,6575E-01	1,7872E-02
THCA	9,4726E-05	**4,5691E-05**	1,9204E-01	9,0242E-03
UCEC	3,1614E-08	**2,5707E-08**	6,3648E-02	1,9723E-03

Sonuç olarak Pearson KKD ile bulunan ilişki skorları kullanılarak GAÇ algoritmaları tarafından yapılan etkileşim tahminlerinde C3NET ve ARACNE algoritmaları daha başarılı skorlar elde etmiştir.

Çizelge 4. 6 Pearson KKD ile BRCA veri seti için GAÇ algoritmaları elde edilen etkileşim bilgileri

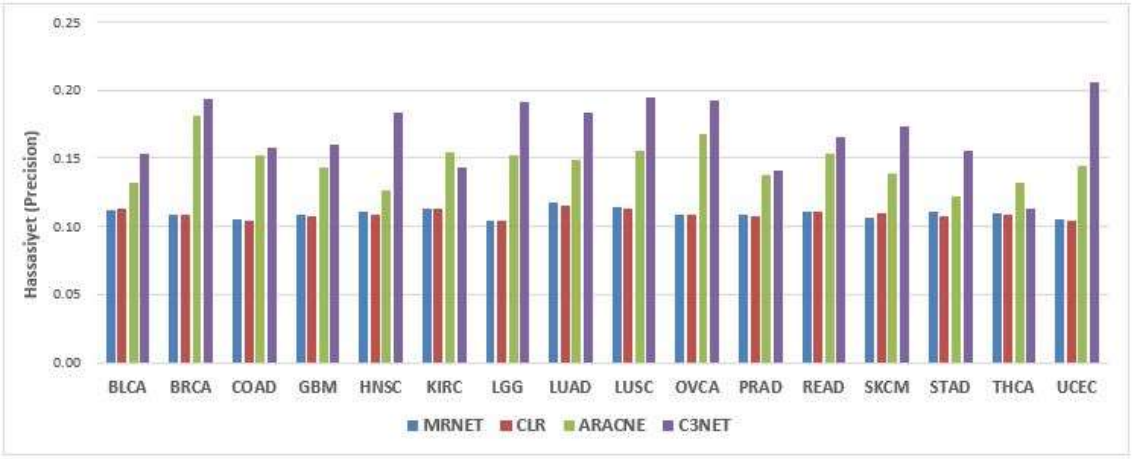
	Örtüşen Etkileşimler	Tahmin Edilen Etkileşimler	Literatürdeki Tüm Etkileşimler	Tüm olası Etkileşimler	P-değeri
MRNET	768	7213	1870	15576	1,246E-06
CLR	774	7243	1870	15576	2,321E-06
ARACNE	51	329	1870	15576	5,862E-02
C3NET	30	155	1870	15576	8,468E-03

4.2.2 Spearman KKD'ye göre GAÇ Yöntemlerinin Başarım Sonuçları

Spearman KKD ile hesaplanan ilişki değerleri kullanılarak çıkarımlanan gen ağlarında GAÇ algoritmalarının başarımları kanser türüne göre değişmekte olup başarım sonuçları kesinlik skorlarına göre Şekil 4.4'te verilmiştir. Kesinlik skorlarına göre yapılan değerlendirmede on dört farklı kanser türü için (BLCA, BRCA, COAD, GBM, HNSC, LGG, LUAD, LUSC, OVA, PRAD, READ, SKCM, STAD ve UCEC) C3NET algoritması en yüksek

değerleri elde etmiştir. C3NET algoritmasından sonra en iyi kesinlik değerlerini 2 farklı kanser türü için (KIRC ve THCA) ARACNE algoritması elde etmiştir. C3NET ve ARACNE algoritmalarının kesinlik skorlarının diğer algoritmalara göre daha yüksek olmasının nedeni bir önceki bölümde açıklandığı üzere ilişki skorları üzerinde yapmış oldukları eleme işlemi sonucu doğruluğu daha yüksek ağlar elde etmeleridir. Bu durumun daha iyi anlaşılması için bir önceki bölümde de açıklandığı üzere BRCA veri seti için her bir GAÇ algoritması tarafından tahmin edilen ilişkilerle, bu tahminlerden gerçek pozitif (literatürdeki veritabanları ile örtüşen) olanların sayıları Çizelge 4.7'de verilmiştir.

Çizelge 4.7'den de görüldüğü üzere Pearson KKD'de olduğu gibi CLR ve MRNET algoritmalarının ilişki tahmin sayıları ile bu tahminlerdeki gerçek pozitif olanların sayıları birlerine yakındır. Bu iki algoritmanın C3NET ve ARACNE algoritmalarına göre yüksek sayıda ilişki tahmini yapmasına karşın bu tahminlerdeki gerçek pozitif sayıları yeteri kadar yüksek olmadığı için kesinlik skorları düşük olmuştur. Buna karşın ARACNE ve C3NET algoritmaları daha az ilişki tahmini yapmalarına karşın bu tahminlerdeki gerçek pozitiflerin daha isabetli olması dolayısıyla daha yüksek kesinlik değerlerine erişmişlerdir. Ayrıca, C3NET algoritmasındaki örtüşen etkileşim sayısı ile tahmin edilen etkileşim sayısının Pearson KKD ile aynı olduğu gözlemlenmiştir.



Şekil 4. 4 Veri setleri için Spearman KKD'ye göre GAÇ algoritmalarının başarımı

Ek olarak, Spearman KKD kullanılarak GAÇ algoritmalarınca yapılmış olan bu tahminlerin istatistiksel olarak anlamlılığını değerlendirmek için Bağıntı 4.1 kullanılarak elde edilen p-değerleri Çizelge 4.8'de verilmiştir. Çizelge 4.8'de koyu yazılı değerleri içeren hücreler

Bölüm 4.2.1 de belirtildiği üzere ilgili kanser türü için istatistiksel olarak en anlamlı p-değerlerini göstermektedir.

Çizelge 4. 7 BRCA veri seti için GAÇ algoritmaları ile tahmin edilen etkileşimlerin sayıları (Spearman KKD ile) ve kesinlik değerleri

	Örtüşen Etkileşimler	**Tahmin Edilen Etkileşimler**	**Kesinlik Skoru**
MRNET	792	7296	0,10855
CLR	798	7346	0,10863
ARACNE	62	342	0,18128
C3NET	30	155	0,19354

Bölüm 4.2.1 de belirtildiği üzere bu sonuçlardan kesinlik değerleri ile p-değerleri arasında bir tutarsızlık olduğu düşünülebilir, fakat aslında çoğu p-değeri istatistiksel olarak anlamlılığı ifade eden ve daha önceden belirlenmiş olan 0,05 eşik değerinden daha küçüktür. MRNET ve CLR algoritmalarının daha düşük p-değerlerine sahip olmasının nedenini daha iyi anlatabilmek için Bölüm 4.2.1'de olduğu üzere Çizelge 4.9 hazırlanmıştır. Çizelge 4.9'dan görüleceği üzere literatürdeki tüm etkileşimlerin sayısı tüm olası etkileşimlerin sayısının sadece %12'sine denk gelmektedir. C3NET ve ARACNE algoritmaları tarafından tahmin edilen etkileşim sayısı, CLR ve MRNET algoritmalarının tahmin ettiği etkileşim sayısına göre çok daha düşüktür. Bu sebepten C3NET ve ARACNE algoritmaları istatistiksel olarak anlamlı p-değerlerine sahip olmalarına karşın MRNET ve CLR algoritmalarından daha yüksek p-değerlerine sahiptirler. Fakat bu değerler 0,05 olarak belirlediğimiz anlamlılık eşik değerinden küçük olduğu için istatistiksel olarak anlamlı kabul edilir.

Sonuç olarak Pearson KDD ile bulunan sonuçlarla benzer olarak, Spearman KKD ile bulunan ilişki skorları kullanılarak, GAÇ algoritmaları tarafından yapılan etkileşim tahminlerinde C3NET ve ARACNE algoritmaları daha başarılı sonuçlar elde etmiştir.

Çizelge 4. 8 Spearman KKD kullanılarak GAÇ algoritmalarınca yapılmış olan tahminlerin FKT p-değerleri

	MRNET	CLR	ARACNE	C3NET
BLCA	**2,0083E-03**	1,1780E-02	4,7051E-01	2,1582E-02
BRCA	**3,3084E-05**	3,3324E-05	9,6344E-04	8,4685E-03
COAD	5,9314E-08	**3,9255E-08**	6,8148E-02	1,3975E-02
GBM	5,7340E-05	**1,3214E-06**	1,5511E-01	1,4436E-02
HNSC	4,6138E-04	**1,0055E-05**	7,4156E-01	1,5154E-02
KIRC	**1,3350E-02**	2,2942E-02	3,5771E-02	3,2953E-02
LGG	**2,3481E-08**	5,4724E-08	7,3273E-02	9,1236E-03
LUAD	4,0174E-01	1,0901E-01	1,0212E-01	**1,5154E-02**
LUSC	4,8106E-02	1,0251E-02	3,5567E-02	**6,4595E-03**
OVCA	1,8541E-04	**1,7878E-04**	2,1958E-03	8,7655E-03
PRAD	7,5130E-05	**1,4810E-05**	2,8832E-01	3,8768E-02
READ	**8,1980E-04**	1,1681E-03	5,8309E-02	8,3551E-02
SKCM	**5,5373E-06**	2,9794E-04	2,4463E-01	4,7036E-02
STAD	1,0630E-03	**1,3667E-05**	8,6556E-01	1,7872E-02
THCA	1,5804E-04	**4,1514E-05**	4,9107E-01	9,0242E-03
UCEC	1,9490E-06	**1,6055E-07**	1,2620E-01	1,9723E-03

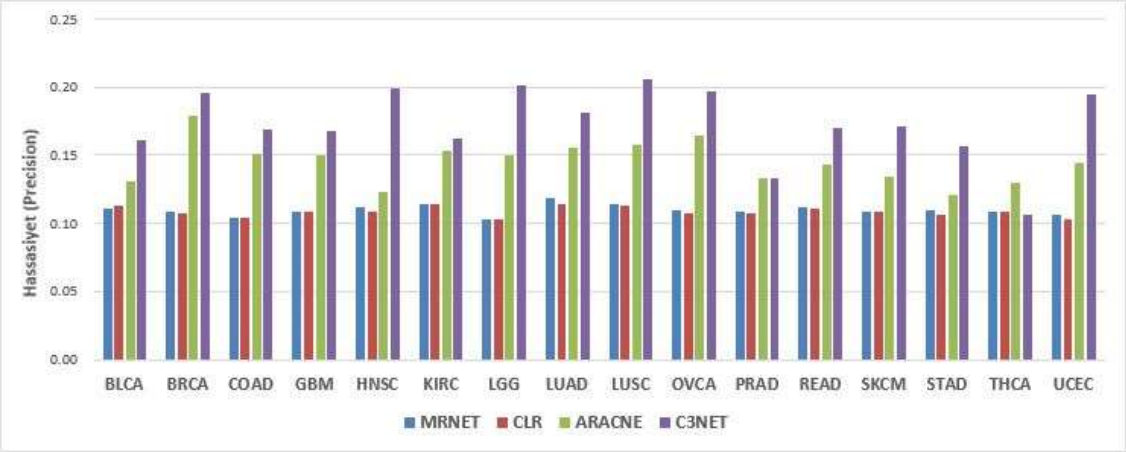
Çizelge 4. 9 Spearman KKD ile BRCA veri seti için GAÇ algoritmaları elde edilen etkileşim bilgileri

	Örtüşen Etkileşimler	Tahmin Edilen Etkileşimler	Literatürdeki Tüm Etkileşimler	Tüm olası Etkileşimler	P-değeri
MRNET	792	7296	1870	15576	3,308E-05
CLR	798	7346	1870	15576	3,332E-05
ARACNE	62	342	1870	15576	9,634E-04
C3NET	30	155	1870	15576	8,468E-03

4.2.3 Kendall KKD'ye göre GAÇ Yöntemlerinin Başarım Sonuçları

Kendall KKD ile hesaplanan ilişki değerleri kullanılarak çıkarımlanan gen ağlarında GAÇ algoritmalarının başarımları kanser türüne göre değişmekte olup başarım sonuçları kesinlik skorlarına göre Şekil 4.5'te verilmiştir. Kesinlik skorlarına göre yapılan değerlendirme de on beş farklı kanser türü için (BLCA, BRCA, COAD, GBM, HNSC, KIRC, LGG, LUAD, LUSC, OVA, PRAD, READ, SKCM, STAD ve UCEC) C3NET algoritması en yüksek değerleri elde etmiştir. C3NET algoritmasından sonra en iyi kesinlik değerlerini bir farklı kanser türü için (THCA) ARACNE algoritması elde etmiştir. C3NET ve ARACNE algoritmalarının kesinlik skorlarının diğer algoritmalara göre daha yüksek olmasının nedeni Bölüm 4.2.1 ve 4.2.2'de açıklandığı üzere ilişki skorları üzerinde yapmış oldukları

eleme işlemi sonucu doğruluğu daha yüksek ağlar elde etmeleridir. Bu durumun daha iyi anlaşılması için BRCA veri seti için her bir GAÇ algoritması tarafından tahmin edilen ilişkilerle, bu tahminlerden gerçek pozitif (literatürdeki veritabanları ile örtüşen) olanların sayıları Çizelge 4.10'da verilmiştir.



Şekil 4. 5 Veri setleri için Kendall KKD'ye göre GAÇ algoritmalarının başarımı

Çizelge 4.10'dan da görüldüğü üzere Pearson ve Spearman KKD'de olduğu gibi CLR ve MRNET algoritmalarının ilişki tahmin sayıları ile bu tahminlerdeki gerçek pozitif olanların sayıları birbirlerine yakındır. Bu iki algoritmanın C3NET ve ARACNE algoritmalarına göre yüksek sayıda ilişki tahmini yapmasına karşın bu tahminlerdeki gerçek pozitif sayıları yeteri kadar yüksek olmadığı için kesinlik skorları düşük olmuştur. Buna karşın ARACNE ve C3NET algoritmaları daha az ilişki tahmini yapmalarına karşın bu tahminlerdeki gerçek pozitiflerin oranının yüksek olmasından dolayı daha yüksek kesinlik değerlerine erişmişlerdir.

Ayrıca, Kendall KKD kullanılarak GAÇ algoritmalarınca yapılmış olan bu tahminlerin istatistiksel olarak anlamlılığını değerlendirmek için Bağıntı 4.1 kullanılarak elde edilen p-değerleri Çizelge 4.11'de verilmiştir. Çizelge 4.11'de koyu yazılı değerleri içeren hücreler Bölüm 4.2.1 ve 4.2.2 de belirtildiği üzere ilgili kanser türü için istatistiksel olarak en anlamlı p-değerlerini göstermektedir.

Çizelge 4. 10 BRCA veri seti için GAÇ algoritmaları ile tahmin edilen etkileşimlerin sayıları (Kendall KKD ile) ve kesinlik değerleri

	Örtüşen Etkileşimler	Tahmin Edilen Etkileşimler	Kesinlik Skoru
MRNET	788	7272	0,10836
CLR	790	7323	0,10787
ARACNE	62	346	0,17919
C3NET	31	158	0,1962

Çizelge 4. 11 Kendall KKD kullanılarak GAÇ algoritmalarınca yapılmış olan tahminlerin FKT p-değerleri

	MRNET	CLR	ARACNE	C3NET
BLCA	**1,4305E-03**	1,6619E-02	5,2814E-01	1,3435E-02
BRCA	2,6461E-05	**1,0901E-05**	1,0724E-03	6,2558E-03
COAD	3,9529E-08	**2,5549E-08**	8,0004E-02	6,5690E-02
GBM	1,1157E-05	**8,9857E-06**	7,7756E-02	6,7147E-02
HNSC	2,6276E-03	**1,5805E-05**	8,6923E-01	4,6125E-03
KIRC	**2,4371E-02**	4,5236E-02	4,3303E-02	3,1053E-02
LGG	**9,9971E-09**	1,2919E-08	8,8562E-02	2,9997E-03
LUAD	7,6718E-01	3,8359E-02	4,0089E-02	**2,6660E-02**
LUSC	5,0926E-02	8,8643E-03	2,9251E-02	**1,9723E-03**
OVCA	3,6748E-04	**7,9023E-05**	4,1940E-03	4,4110E-03
PRAD	2,8743E-05	**9,3622E-06**	4,1147E-01	6,2073E-02
READ	3,1894E-03	**1,2721E-03**	1,6417E-01	6,4460E-02
SKCM	**5,2105E-05**	1,4933E-04	3,9711E-01	6,3459E-02
STAD	2,6740E-04	**2,6117E-06**	9,3272E-01	1,7700E-02
THCA	**5,1162E-05**	6,3108E-05	5,4911E-01	7,1292E-02
UCEC	2,7817E-06	**4,5297E-08**	1,2710E-01	6,4595E-03

Bölüm 4.2.1 ve 4.2.2'de belirtildiği üzere bu sonuçlardan kesinlik değerleri ile p-değerleri arasında bir tutarsızlık olduğu düşünülebilir, fakat aslında çoğu p-değeri istatistiksel olarak anlamlılığı ifade eden ve daha önceden belirlediğimiz 0,05 eşik değerinden daha küçüktür. MRNET ve CLR algoritmalarının daha düşük p-değerlerine sahip olmasının nedenini daha iyi anlatabilmek için Bölüm 4.2.1 ve 4.2.2'de olduğu üzere Çizelge 4.12 hazırlanmıştır. Çizelge 4.12'den görüleceği üzere literatürdeki tüm etkileşimlerin sayısı tüm olası etkileşimlerin sayısının sadece %12'sine denk gelmektedir. C3NET ve ARACNE algoritmaları tarafından tahmin edilen etkileşim sayısı ise CLR ve MRNET algoritmalarına karşın çok daha düşük kaldığı için istatistiksel olarak anlamlı p-değerlerine sahip olmalarına karşın MRNET ve CLR algoritmalarından daha yüksek p-değerlerine

sahiptirler. Fakat bu değerler 0,05 olarak belirlenmiş olan anlamlılık eşik değerinden küçük olduğu için istatistiksel olarak anlamlı kabul edilir.

Çizelge 4. 12 Kendall KKD ile BRCA veri seti için GAÇ algoritmaları elde edilen etkileşim bilgileri

	Örtüşen Etkileşimler	Tahmin Edilen Etkileşimler	Literatürdeki Tüm Etkileşimler	Tüm olası Etkileşimler	P-değeri
MRNET	788	7272	1870	15576	2,646E-05
CLR	790	7323	1870	15576	1,090E-05
ARACNE	62	346	1870	15576	1,072E-03
C3NET	31	158	1870	15576	6,256E-03

Sonuç olarak Pearson ve Spearman KDD ile bulunan sonuçlarla benzer olarak, Kendall KKD ile bulunan ilişki skorları kullanılarak, GAÇ algoritmaları tarafından yapılan etkileşim tahminlerinde C3NET ve ARACNE algoritmaları daha başarılı sonuçlar elde etmiştir.

4.3 Bölüm Sonuçları

Bu bölümde literatürde sıklıkla kullanılan dört GAÇ algoritmasının proteomik kanser verisi kullanarak yaptıkları etkileşim tahminlerinin başarımlarında korelasyona dayalı ilişki tahmincilerinin etkisi incelenmiştir. İncelenen on altı kanser verisi için korelasyona dayalı ilişki tahmincilerin başarımları kesinlik ve FKT p-değerlerine göre incelendiğinde benzer değerler ürettikleri görülmektedir. Bir başka deyişle aynı kanser türü için aynı algoritma farklı korelasyon tabanlı ilişki tahmincisinde benzer başarım skorlarına erişmiştir. Buna karşılık C3NET algoritması kesinlik skorlarına göre en başarılı algoritma olmuştur. Bu sonuçların belirlenen 0,05 eşik değeri için FKT p-değerlerine göre değerlendirildiğinde istatistiksel olarak da anlamlı olduğu görülmüştür.

Korelasyon yöntemlerinin GAÇ algoritmaları üzerindeki başarımı değerlendirildiğinde MRNET, CLR ve C3NET yöntemleri üzerinde büyük bir etkisi olmadığı görülmüştür. Fakat korelasyon tabanlı ilişki tahmincisinin değişmesi ARACNE algoritmasının başarımını değiştirmektedir ve 8 veri setinde (BLCA, COAD, GBM, HNSC, LGG, LUSC, THCA, UCEC) Pearson KKD, 7 veri setinde (BRCA, KIRC, OVCA, PRAD, READ, SKCM, STAD) Spearman KKD, 1 veri setinde (LUAD) de Kendall Tau KKD en başarılı yöntem olmuştur.

Analiz işlemleri Ubuntu 15.10 Linux işletim sistemine, 4 GB belleğe ve i5-4460 3.2 GHz işlemciye sahip bir kişisel bilgisayar üzerinde yapılmıştır. Korelasyon tabanlı ilişki

tahmincilerinin algoritmaların çalışma zamanları üzerindeki etkisi değerlendirilmiş ve buna göre tüm GAÇ algoritmalarında ilişki tahmincilerin çalışma hızı iyiden kötüye doğru Pearson KKD, Spearman KKD ve Kendall KKD şeklinde olmuştur. Tüm veri setleri için algoritmaların ortalama çalışma süreleri saniye cinsinden Çizelge 4.13'te verilmiştir.

Çizelge 4. 13 GAÇ algoritmalarının korelasyon tabanlı ilişki tahmincilere göre ortalama çalışma süreleri (sn)

	MRNET	**CLR**	**ARACNE**	**C3NET**
Pearson KKD	0,0741	0,0651	0,0650	0,5289
Spearman KKD	0,0896	0,0721	0,0810	0,6443
Kendall KKD	27,5192	27,4916	27,4884	308,2056

Pearson KKD'ye göre GAÇ algoritmalarının çalışma süreleri düşükten yükseğe doğru sıralanması ARACNE, CLR, MRNET ve C3NET şeklinde olmuştur fakat burada ARACNE ve CLR'ın çok yakın çalışma sürelerine sahip olduğu unutulmamalıdır. Bu sıralama Spearman KKD'ye göre yapıldığında CLR, ARACNE, MRNET ve C3NET şeklinde olmuştur. Kendall KKD için benzer sıralama ARACNE, CLR, MRNET ve C3NET şeklinde yapılabilir. Kendall KKD yapısı gereği sıralama işlemi yaptığı için GAÇ algoritmalarının bu ilişki tahminci ile çalışma süreleri daha uzun olmuştur. Korelasyon tabanlı ilişki tahmincilerinde GAÇ algoritmalarının çalışma süreleri incelendiğinde ARACNE, CLR ve MRNET'in daha hızlı olduğu görülmekle birlikte C3NET'in daha uzun çalışma süresine sahip olduğu görülmektedir.

Bu bölümde yapılan çalışma sonucunda daha başarılı sonuçlar aldığımız C3NET algoritması tarafından anlamlı bulunan ve etkileşim veritabanlarında yer alan ilişkilerden en anlamlı değerlere sahip protein çiftleri Çizelge 4.14'te verilmiştir. Yeşil renkle doldurulmuş hücreler ilgili kanser türü için verilen protein çiftinin C3NET tarafından anlamlı bulunduğunu bildirmektedir. Literatürde ilgili kanser türü için önemli olduğu saptanmış protein çifti için ilgili hücre '*' işareti ile belirtilmiş ve kaynakları verilmiştir. Çizelge 4.14'te verilen protein çiftlerinin literatürde yer almayan diğer kanser türleri için de biyologlar tarafından araştırılması da faydalı olacaktır.

Çizelge 4. 14 C3NET tarafından kanser türleri için anlamlı bulunan etkileşimler

	Kanser Türü															
Etkileşimler (Gen-Gen Protein-Protein)	**BLCA**	**BRCA**	**COAD**	**GBM**	**HNSC**	**KIRC**	**LGG**	**LUAD**	**LUSC**	**OVCA**	**PRAD**	**READ**	**SKCM**	**STAD**	**THCA**	**UCEC**
CDH1-CTNNB1 E.Cadherin-beta.Catenin							*[64]									
FOXM1-CCNB1 FoxM1-Cyclin_B1	*[65]															
MSH6-MSH2										*[66], [67]						
SHC1-EGFR Shc_pY317-EGFR_pY1068								*[68]								

Nikuseva Matric vd. [64] yapmış oldukları çalışmada, merkezi sinir sisteminin 50 tümöründe E-cadherin (CDH1), adenomatous polyposis coli (APC) ve beta-catenin (CTNNB1) proteinlerinin değişimlerini araştırmışlardır. Sonuç olarak WNT biyolojik yolunun bileşenlerindeki genetik değişikliklerin beyin tümörü oluşumunda rol oynadığını tespit etmişlerdir ve E-cadherin (CDH1)'deki değişikliklerin beyin kanserinde etkin olabileceğini bildirmişlerdir. Kyu-Kim vd. [65] çalışmalarında 102 mesane kanseri hastasına ait genomik verileri üzerinde FOXM1-CCNB1 proteinlerinin etkileşimlerini araştırmışlardır. Yapmış oldukları gen ağ çıkarımının ve deneysel incelemelerin sonucunda mesane kanserinin oluşumunda FOXM1-CCNB1-Fanconi anemi biyolojik yolunun (Fanconi anemia pathways) aracılık edebileceğini belirtmişlerdir. Rambau vd. [66] araştırmalarında 602 yumurtalık kanseri hastasına ait mikro dizi ifade verilerini kullanarak onarım proteinlerinin (MLH1, PMS2, MSH2 ve MSH6) etkisini incelemişlerdir. Araştırma sonucunda yumurtalık kanserinde MSH2 / MSH6 anormalliğinin anlamlı sıklığının, yumurtalık kanserine özgü Lynch sendromu (Lynch syndrome)[1] taramasını desteklediğini bildirmişlerdir. Ramsoekh vd. [67] yapmış oldukları çalışmada Lynch sendromu kanıtlanmış 67 aileye ait verileri kullanarak MLH1, MSH2 ve MSH6 proteinlerindeki değişimlerin kanser üzerindeki etkisini incelemişlerdir. Sonuç olarak MLH1, MSH2 ve MSH6 proteinlerindeki mutasyonların ayırt edilebilir kanser risk profillerine neden olduğunu ve MSH6 proteinindeki mutasyonun yumurtalık kanserinde yüksek risk oluşturduğunu belirtmişlerdir. Zheng vd. [68] çalışmalarında aşağı regüle

[1] Lynch Sendromu cinsiyeti belirleyen kromozomlar dışında kalan kromozomlarda meydana gelen hasarlar sonucu oluşan kalıtsal bir hastalıktır.

adaptör proteini (Downregulated adaptor protein) $p66^{SHC}$ (SHC1)'nin akciğer kanseri üzerindeki etkileri incelemişlerdir. Çalışma sonucunda SHC1 eksikliğinin akciğer kanserinde etkili olabileceğini bildirmişlerdir.

Tüm bu sonuçlar bu tez çalışmasında önerilen yöntemin ve istatiksel olarak anlamlı bulunan sonuçların biyolojik olarak da anlamlı olduğunu göstermektedir.

BÖLÜM 5

GELİŞTİRİLEN YÖNTEM

Biyolojik ağ analizi için farklı algoritmalar geliştirilmiş olup, bu farklılığın ana sebebi kullandıkları istatistiksel yaklaşımlardır. Bu alanda yapılmış olan ilk çalışmalarda korelasyon tabanlı kümeleme yöntemleri kullanılmıştır [69]. WGCNA [22] literatürde çokça kullanılan ağ çıkarım yöntemlerinden biri olmakla birlikte hedef hastalıkla ilgili yüksek korelasyonlu gen/protein modüllerini (küme, alt-ağ, sub-network) ve bu modüllerdeki yüksek ilişki skorlarına sahip merkez genleri tespit etmek için kullanılmaktadır.

Bu bölümde ilişki tahmincilerinin entegrasyonunun ağ çıkarımı üzerindeki etkisini incelemek için önerdiğimiz ve sonuçlarını literatürle paylaştığımız yöntem açıklanacaktır.

5.1 Kullanılan Veri Seti

Biyoinformatik alanındaki çalışmalar, yüksek verimli tekniklerdeki gelişmeler ile birlikte büyük bir ivme kazanmıştır. Bu tekniklerden biri, yüksek verimli antikor bazlı bir teknik olan ve proteomik araştırmaları için protein ekspresyon verileri sağlayan ters faz protein dizisi (Reverse Phase Protein Array - RPPA) [58] teknolojisidir.

Bu çalışmada kullanılan RPPA verileri, TCPA'nın web sitesindeki indirme bölümünden (4. düzey veri) alınmıştır [70]. Veri seti 5 farklı kanser türüne aittir ve toplamda 2.230 kanser hastasının proteomik ifade verisinden oluşur. Veri setinde her bir hasta için 218'den fazla antikor bulunmaktadır. Verilerin bire-bir gen-protein eşleşmesini (Çizelge A-2.1) elde etmek için, TCPA'nın [70] web sitesinin My Protein bölümünden ilgili bilgiler alınmış ve sonrasında eşleme gerçekleştirilmiştir. Kanser türleri ve kısaltmaları, örnek sayıları, protein sayısı (antikor), seçilen protein sayısı Çizelge 5.1'de listelenmiştir.

Çizelge 5. 1 Kanser türü ve kısaltması, örnek sayısı, protein sayısı, seçilen proteinlerin sayısı

Kanser Türü	Kısaltması	Örnek Sayısı	Protein sayısı	Seçilen Proteinlerin Sayısı
Göğüs Kanseri (Breast invasive carcinoma)	BRCA	901	224	173
Beyin Kanseri (Glioblastoma multiforme)	GBM	205	223	173
Akciğer Kanseri (Lung squamous cell carcinoma)	LUSC	325	237	170
Böbrek Kanseri (Kidney clear cell carcinoma)	KIRC	445	233	169
Deri Kanseri (Skin Cutaneous Melanoma)	SKCM	354	223	173

Ayrıca, ilişki tahmin edicilerin performanslarını değerlendirmek için Hastalık-Gen ilişkileri entegrasyon platformu (DisGeNET) [71] ve Moleküler İmzalar Veritabanı (MSigDB) [72] altın standart olarak kullanılmıştır. DisGeNET literatürde yapılmış önceki çalışmalarda bildirilen ve UNIPROT veritabanı, Karşılaştırmalı Toksikogenomik Veritabanı (Comparative Toxicogenomics Database), GWAS kataloğu gibi çeşitli kamuya açık veri kaynaklarında saklanan gen-hastalık, genetik varyant-hastalık ilişkilerini içeren bir entegrasyon platformudur. DisGeNET 17.381 gen ve 15.093 hastalık arasındaki toplam 429.036 ilişkiyi barındırmaktadır. MSigDB ise 8 farklı koleksiyondan toplamda 17.779 gen seti barındırır ve kapsadığı koleksiyonlar; sıralı gen kümeleri, konumsal gen setleri, küratörlü gen setleri, motif gen setleri, hesaplamalı gen setleri, Gen Ontolojisi (GO) gen setleri, immünolojik (immunologic) gen setleri ve onkojenik gen kümeleridir.

Her bir kanser türüne göre verileri elde etmek için kullanılan arama terimleri, Birleşik Tıbbi Dil Sistemi - Kavram Benzersiz Tanımlayıcıları (Unified Medical Language System - Concept Unique Identifiers, UMLS-CUIs), en az literatürdeki iki yayında geçmesi kaydıyla (PMID ≥ 2) DisGeNET'ten elde edilen genlerin sayısı ve ilgili kanser veri setindeki genler ile DisGeNET arasındaki ortak genlerin sayısı Çizelge 5.2'de verilmiştir. Gen seviyesi analizi, seçilen kanserle ilgili DisGeNET verileri kullanılarak gerçekleştirilmiştir.

Çizelge 5. 2 DisGeNET'ten alınan verilerin bilgisi

	Kanser Türü	DisGeNET'teki arama terimleri	UMLS-CUIs	DisGeNET'ten alınan gen sayısı (PMIDs>=2)	Ortak gen sayısı
1	BRCA	Göğüs Kanseri (Breast invasive carcinoma)	C0346153	52	15
2	GBM	Beyin Kanseri (Glioblastoma multiforme)	C1621958	147	28
3	LUSC	Akciğer Kanseri (Lung squamous cell carcinoma)	C0149782	41	7
4	KIRC	Böbrek Kanseri (Kidney clear cell carcinoma)	C0022658	264	14
5	SKCM	Deri Kanseri (Skin Cutaneous Melanoma)	C0151779	104	25

DisGeNET'ten her bir kanser türü için seçilen genler alınarak BioCarta, KEGG ve Reactome veritabanlarından bu genleri içerme oranı en yüksek yüz biyolojik-yol (biological pathway) MSigDB çevrimiçi aracı kullanılarak, yanlış keşif oranı (False Discovery Rate – FDR-q) <0,05 ile tanımlanarak, alınmıştır. FDR skoru çoklu hipotez testi için ham p-değerinin ayarlanmış halidir. Burada, ham p değerlerini düzeltmek için Bonferroni [73] düzeltmesi kullanılarak p-değerleri test edilen hipotezlerin sayısı (100 biyolojik-yol alındığı için ham p-değeri 100 ile çarpılmıştır.) ile çarpılmıştır. Yol seviyesi analizi, DisGeNET'ten elde edilen hastalık genlerinin ilgili olduğu en önemli biyolojik-yollar kullanılarak gerçekleştirilmiştir.

5.2 Analiz

GAÇ algoritmaları gen ifade/proteomik verilerini kullanarak genler-proteinler arasındaki etkileşimleri bulmaya çalışırlar. Bunun için gen-protein çiftleri arasındaki ilişki değerlerini Bölüm 2'de verilen tahminciler gibi farklı istatistiksel yöntemlerle gerçekleştirirler. WGCNA yönteminde gen-protein çiftleri arasındaki ilişki değerleri Bölüm 3.6'da anlatılan, Pearson KKD ve Spearman KKD gibi korelasyon temelli *İlişki (adjacency)* fonksiyonu ile hesaplanır. Önerdiğimiz yöntemde Bölüm 3.6'da WGCNA yönteminin ilk adımında kullanılan *İlişki* fonksiyonu yerine Bölüm 2'de verilen ilişki tahmincilerinin entegrasyonu sağlanarak WGCNA yönteminin başarımı değerlendirilmiştir.

Bölüm 3.6'da ayrıntıları verilmiş olan WGCNA kullanılarak yapılmış olan analiz işleminin adımları sırasıyla aşağıda verilmiştir:

- İlk olarak TCPA'dan alınan veri setindeki proteomik verileri kullanılarak proteinler arasındaki ilişki skor matrisi hesaplanmıştır (*adjacency* fonksiyonu ile).

- Yüksek ilişki değerleriyle birbirine bağlı protein gruplarını (modüllerini) belirlemek için topolojik örtüşme ölçütünü kullanılarak benzerlik matrisi elde edilmiştir (*TOMsimilarity* fonksiyonu ile).
- Oluşturulan bu benzerlik matrisinden, hiyerarşik kümeleme yöntemi kullanılarak, birbiriyle ilişkili protein gruplarını tespit etmek için kümeleme ağacı oluşturulmuştur (*hclust* fonksiyonu ile).
- Oluşturulan kümeleme ağacından dinamik ağaç kesme yöntemi ile anlamlı gruplar bulunmuştur (*cutreeDynamic* fonksiyonu ile).
- Son adımda ise tespit edilen her bir gruptaki yüksek bağlı merkez gen/protein bulunmuştur (*selectOneHubInEachModule* fonksiyonu ile).

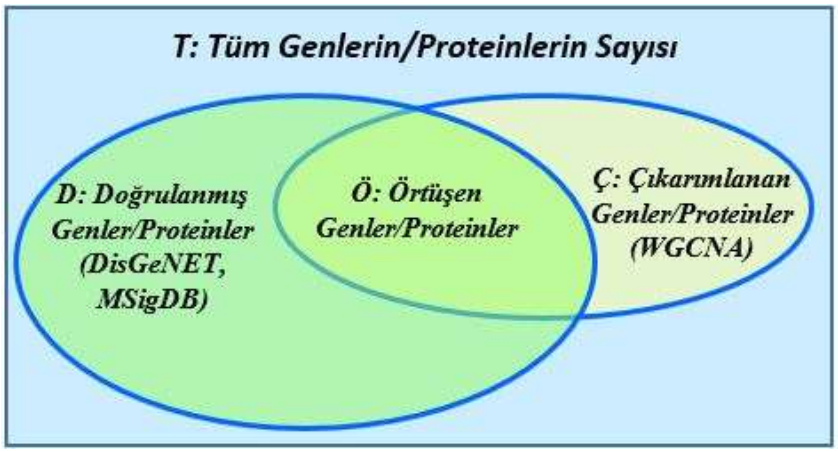
İncelenen kanser türü ile bulunan modüller arasındaki ilişkinin anlamlılığını istatistiksel olarak değerlendirmek için Bölüm 4.1'de ve Fury vd.'nin [63] yaptığı çalışmada anlatıldığı üzere FKT'nin tek yönlü (one-tailed) versiyonu olan hipergeometrik dağılım (hypergeometric distribution) ile hesaplanan örtüşme değeri (P-değeri) bulunarak kullanılmıştır. Bulunan p-değeri belirli bir eşik değerinin altındaysa bulunan modüldeki proteinlerin incelenen kanser türü ile olan ilişkisi istatistiksel olarak rastgelelikten uzak ve biyolojik olarak ta anlamlı kabul edilir.

Fury vd.'nin [63] yapmış olduğu çalışmada verilen hipergeometrik dağılım bağıntısının çalışmamıza uyarlanmış hali; Ç, ağ çıkarım yöntemi ile çıkarımlanan gen/protein sayısını; D, DisGeNET veritabanında doğrulama için kullanılacak gen/protein sayısını; Ö, Ç ve D arasındaki örtüşen gen/protein sayısını; T, insan genomu için bilinen tüm genlerin/proteinlerin sayısını belirtmek üzere Bağıntı 5.1'de verilmiştir.

$$\mathrm{P}(\ddot{\mathrm{O}}) = \frac{\binom{\text{Ç}}{\ddot{\mathrm{O}}}\binom{T-\text{Ç}}{D-\ddot{\mathrm{O}}}}{\binom{T}{D}} = \frac{\text{Ç!}\quad (T-\text{Ç})!\quad D!\quad (T-D)!}{\ddot{\mathrm{O}}!\quad (\text{Ç}-\ddot{\mathrm{O}})!\quad (D-\ddot{\mathrm{O}})!\quad (T-\text{Ç}-D+\ddot{\mathrm{O}})!\quad T!} \tag{5.1}$$

Hesaplanan p-değeri 0,05'ten küçükse, çıkarılan modüller ile hastalığa bağlı DisGeNET genleri arasındaki örtüşen proteinlerin ve bunların ilgili olduğu biyolojik-yolların rastgele olma ihtimalinin daha düşük olduğu, bu modüllerin hastalık ile ilişkili ve biyolojik olarak ilginç olma ihtimalinin daha yüksek olduğu düşünülür. Ayrıca, Şekil 5.1 ve Çizelge 5.3,

ilişkilendirme tahmincilerinin performanslarının genel değerlendirmesinde kullanılan FKT'nin daha iyi anlaşılmasını sağlamak için verilmiştir.



Şekil 5. 1 FKT ile örtüşme analizi

Şekil 5.1'de verildiği gibi, verilen iki kümeye (ör: çıkarımlanmış ve onaylanmış genler/proteinler) dayalı olarak, örtüşme işleminin literatürdeki tüm genlere/ proteinlere (T) göre istatistiksel olarak anlamlı olup olmadığı belirlenmiştir. Literatürdeki son çalışmalar insan genomundaki gen sayısının 19.000 civarında olduğunu ortaya koymuştur [74], dolayısıyla bu değer çalışmamızda T olarak kullanılmıştır. Çizelge 5.3, Şekil 5.1'de gösterilen bölgeleri özetler ve bu bölgeler FKT'de bir girdi olarak kullanılır.

Çizelge 5. 3 Gen seviyesinde analiz için FKT Parametreleri

	Literatürde Doğrulanmış Genler/Proteinler	Literatürde Doğrulanmamış Genler/Proteinler
Çıkarımlanmış Genler/Proteinler	*Ö*	*Ç – Ö*
Çıkarımlanmamış Genler/Proteinler	*D - Ö*	*T – Ç – D + Ö*

Değerlendirme sürecinde kullanılan bir diğer metrik kesinlik (precision) skorudur ve Bölüm 4'te Bağıntı 4.2 ile belirtilmiştir. Kesinlik değerinin hesaplanmasında gerçek pozitifler, biyolojik-yol seviyesindeki veya gen seviyesindeki analizde gerekli koşulları sağlayan modülleri gösterir (P değeri <0,05). Yanlış pozitifler ise biyolojik yol seviyesindeki veya gen seviyesindeki analizlerde gerekli koşulları karşılamayan modüllerin sayısını gösterir. Biyolojik-yol seviyesindeki ve gen seviyesindeki analizlerde belirlenen gerekli koşullar Bölüm 5.3'te açıklanmıştır.

Analizler için MINET [8], DepEst [75] ve WGCNA [22] R paketleri kullanılmıştır. Seçilen ilişki tahmincileri MINET [8]'ten *build.mim* fonksiyonu ve DepEst [75]' den *get.mim* fonksiyonu tarafından sağlanan belirli parametre değerleri ile hesaplanmıştır. Ayrıca, *adjacency*, *TOMsimilarity*, *hclust*, *cutreeDynamic* ve *selectOneHubInEachModule* fonksiyonları analiz sürecinde WGCNA [22] paketinden kullanılan fonksiyonlardır.

5.3 Geliştirilen Yöntem

Bu bölümde, ilişki tahmincilerinin performanslarını analiz etmek için geliştirilen yöntem Şekil 5.2'de gösterilmiştir.

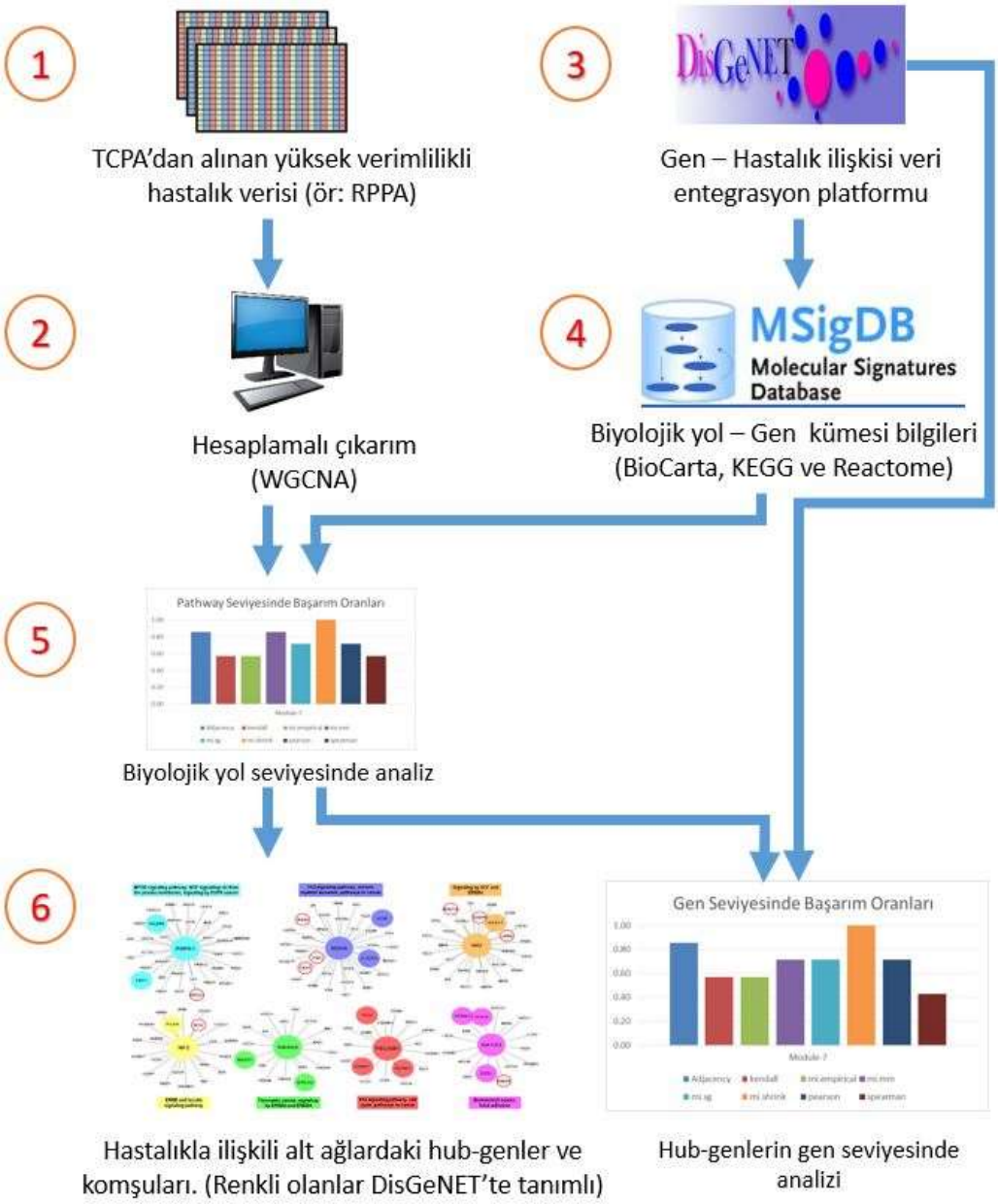
1. İlk adımda, TCPA'dan alınan proteomik veriler analiz için hazırlanır (bkz. Bölüm 4.1).
2. İkinci adımda, Bölüm 5.2'de açıklandığı üzere WGCNA [22] ile analiz işlemleri gerçekleştirilerek ortak ifade ağları ve modüller bulunmuştur. Sonraki aşamalarda kullanılacak olan ilişki skor matrisi WGCNA'da [22] adjacency fonksiyonu ve ayrıca, MINET [8] ve DepEst [75] paketleri tarafından sağlanan diğer ilişkilendirme tahmin edicileri kullanılarak hesaplanmıştır. Ayrıca, her bir ilişki tahmincisi için en az 7 modül (alt ağ) üreten parametreler, belirli parametreler kullanılarak bulunmuştur. Kullanılan parametreler Çizelge 5.4'te verilmiştir.

Çizelge 5. 4 Kullanılan R paketleri, fonksiyonlar, parametreler ve test edilen değerler

R Paketi	**Fonksiyon**	**Parametreler**	**Test edilen değerler**
MINET	build.mim	estimator	"spearman", "pearson", "kendall", "mi. empirical", "mi.mm", "mi.shrink", ve "mi.sg".
		disc	"none", "equalfreq" ve "equalwidth"
DepEst	obtain.mim	estimator	"b.spline" ve "KDE".
		cop.transform	TRUE, FALSE.
WGCNA	TOMsimilarity	TOMDenom	min ve mean.
	hclust	method	"ward.D", "ward.D2", "single", "complete", "average" ve "mcquitty".
	cutreeDynamic	deepSplit	0'dan 4'e kadar 1 artarak.
		minClusterSize	10'dan 20'ye kadar 1 aratarak.
		cutHeight	0,7'den 0,999'a kadar 0,001 artarak.

Her bir veri kümesi ve her bir ilişki tahmincisi için ortak ifade ağlarından 7, 8 ve 9 modül oluşturulur. Modül sayısının 7-8-9 olarak seçilmesinin nedeni detaylarıyla Bölüm 5.4'te açıklanmıştır. Oluşturulan modüller içinden, modüllerdeki protein

sayıları daha eşit dağılmış olan modülleri üreten parametreler, her bir ilişki tahminci için seçilmiştir. Seçilen bu parametreler kullanılarak oluşturulan modüller İlişki tahmincilerinin başarım karşılaştırılmasında kullanılmıştır. Çizelge 5.5'te her bir veri seti için seçilen en iyi ilişki tahmincisi ve kullanılan parametreler verilmiştir.



Şekil 5. 2 Geliştirilen yöntem

3. Üçüncü adımda, DisGeNET web sayfası [76] kullanılarak her kanser türü için en az iki farklı çalışmada (PMID ≥ 2) verilen kanser tipi ile ilişkili olduğu doğrulanmış olan genler elde edilmiştir.

Çizelge 5. 5 Kullanılan veri setine göre fonksiyonlardaki en iyi sonuçları veren parametre değerleri

		Fonksiyonlar, parametreler ve seçilen değerler								
Modül boyutu	Veri seti	build.mim		obtain.mim		TOMsimilarity	hclust	cutreeDynamic		
		İlişki tahminci	discretization method	İlişki tahminci	cop. transform	TOMDenom	method	min Cluster Size	cut Height	deep Split
7	BRCA	Shrink	globalequalwidth	-	-	mean	ward.D2	14	0,994	2
8	BRCA	Shrink	equalfreq	-	-	min	complete	10	0,995	2
9	BRCA	Shrink	globalequalwidth	-	-	mean	ward.D2	10	0,986	2
7	GBM	SG	equalfreq	-	-	mean	ward.D	13	0,887	1
8	GBM	SG	globalequalwidth	-	-	mean	ward.D2	12	0,991	3
9	GBM	SG	globalequalwidth	-	-	mean	ward.D2	11	0,991	3
7	KIRC	Shrink	equalfreq	-	-	min	complete	14	0,995	3
8	KIRC	Shrink	equalfreq	-	-	min	complete	11	0,995	3
9	KIRC	Shrink	equalfreq	-	-	min	complete	10	0,994	4
7	LUSC	-	-	BS	TRUE	mean	ward.D2	10	0,848	2
8	LUSC	-	-	BS	TRUE	mean	ward.D	10	0,866	2
9	LUSC	-	-	BS	TRUE	mean	ward.D	11	0,866	3
7	SKCM	MM	equalfreq	-	-	min	complete	13	0,945	1
8	SKCM	MM	equalwidth	-	-	min	complete	10	0,854	0
9	SKCM	MM	globalequalwidth	-	-	min	complete	10	0,985	1

4. Dördüncü adımda, üçüncü adımda seçilen genler alınarak BioCarta, KEGG ve Reactome veritabanlarından bu genleri içerme oranı en yüksek yüz biyolojik-yol (biological-pathway) MSigDB web sayfası [77] kullanılarak, yanlış keşif oranı (False Discovery Rate – FDR-q) <0,05 ile tanımlanarak, bulunmuştur.

5. **FKT ile biyolojik-yol (pathway) seviyesi analizi**: Bu adımda, 4'ncü adımda elde edilen biyolojik-yollar ile ikinci adımda elde edilen modüller arasındaki ilişkinin örtüşme oranları ve ilgili p-değerleri, FKT kullanılarak bulunmuştur. 4'ncü adımda DisGeNET genleri ile ilişkisi en yüksek 100 biyolojik-yolu seçildiğinden çoklu hipotez testi için ham p-değerlerini düzeltmek üzere, FDR q-değeri, p-değeri 100 ile çarpılarak hesaplanmıştır. Bulgularımızı hastalık ilgisi açısından biyolojik-yol seviyesinde değerlendirmek için, 4'ncü adımda FDR q-değeri <0,05 olan ve hastalıkla ilişkili biyolojik-yolların en az iki tanesine sahip ve en az 5 ortak gen (DisGeNet'ten alınan genlerle çıkarımlanan modüller arasındaki ortak gen) içeren WGCNA modülleri gerçek pozitif olarak kabul edilmiştir. Eşik değeri olarak en az 2 biyolojik-yol ve en az 5 ortak genin seçilmesinin nedeni Bölüm 5.5'te açıklanmıştır. Bu koşulları sağlayan modüllerin sayısı (gerçek pozitif), toplam modül sayısına (gerçek

pozitif + yanlış pozitif) bölünerek, ilişki tahmincilerinin biyolojik-yol seviyesindeki performans puanları (kesinlik değerleri) elde edilmiştir. İncelenen kanser türlerine ait biyolojik-yol seviyesindeki kesinlik skorları her bir ilişki tahmincisi için modül sayılarına göre ayrı ayrı ve ortalama olarak (ort.) Çizelge 5.6'da verilmiştir. Ayrıca, Çizelge 5.6'da her bir kanser türü için en yüksek ortalama kesinlik skoru koyu renkle belirtilmiştir.

Çizelge 5. 6 Kanser türleri için modül sayılarına göre ilişki tahmincilerin FKT ile biyolojik-yol seviyesindeki kesinlik skorları

	BRCA				**GBM**				**KIRC**				**LUSC**				**SKCM**			
	7	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**
Adjacency	0,86	0,63	0,33	0,61	0,57	0,63	0,56	0,58	0,43	0,63	0,33	0,46	0,71	0,75	0,56	0,67	0,86	0,63	0,56	0,68
Kendall Tau KKD	0,57	0,50	0,44	0,51	0,71	0,63	0,67	0,67	0,57	0,38	0,33	0,43	0,57	0,50	0,44	0,51	0,71	1,00	0,56	**0,757**
Pearson KKD	0,71	0,50	0,56	0,59	0,43	0,75	0,44	0,54	0,57	0,38	0,33	0,43	0,57	0,75	0,44	0,59	0,71	0,75	0,67	0,71
Spearman KKD	0,57	0,50	0,67	0,58	0,57	0,38	0,44	0,46	0,71	0,25	0,33	0,43	0,57	0,38	0,44	0,46	0,71	0,75	0,67	0,71
BS	0,71	0,63	0,44	0,59	0,57	0,63	0,67	0,62	0,43	0,63	0,33	0,46	0,86	0,75	0,67	**0,76**	0,86	0,63	0,56	0,68
Ampirik	0,71	0,75	0,56	0,67	0,71	0,50	0,56	0,59	0,57	0,38	0,22	0,39	0,86	0,75	0,44	0,68	0,86	0,63	0,56	0,68
ÇYT	0,43	0,50	0,33	0,42	0,86	0,63	0,44	0,64	0,43	0,25	0,33	0,34	0,71	0,63	0,56	0,63	0,86	0,63	0,56	0,68
MM	0,86	0,38	0,33	0,52	0,86	0,75	0,56	0,72	0,71	0,38	0,33	0,47	0,71	0,75	0,56	0,67	0,86	0,75	0,67	**0,758**
SG	0,71	0,63	0,56	0,63	0,86	0,88	0,78	**0,84**	0,57	0,50	0,22	0,43	0,86	0,63	0,44	0,64	0,71	0,63	0,67	0,67
Shrink	1,00	0,88	0,67	**0,85**	0,71	0,63	0,56	0,63	0,71	0,63	0,44	**0,59**	0,71	0,88	0,67	**0,75**	1,00	0,63	0,56	0,73

6. **FKT ile gen seviyesi analizi**: Bu adımda 5'nci adımda her bir kanser türü için biyolojik-yol seviyesinde gerçek pozitif kabul edilen modüllerdeki genlerle ilgili kanser türüyle ilişkili olduğu DisGeNet'ten belirlenen genlerin örtüşme oranları FKT ile hesaplanmıştır ve 0,05'ten küçük p-değerine sahip modüllerin ilgili kanser türüyle ilişkili olabileceği değerlendirilip, bu modüller gerçek pozitif kabul edilmiştir. Gerçek pozitif modüllerin sayısı, toplam modül sayısına (gerçek pozitif + yanlış pozitif) bölünerek, ilişki tahmincilerinin gen seviyesindeki performans puanları (kesinlik değerleri) elde edilmiştir. İncelenen kanser türlerine ait gen seviyesindeki kesinlik skorları her bir ilişki tahmincisi için modül sayılarına göre ayrı ayrı ve ortalama olarak (ort.) Çizelge 5.7'de verilmiştir. Ayrıca, Çizelge 5.7'de her bir kanser türü için en yüksek ortalama kesinlik skoru koyu renkle belirtilmiştir.

Çizelge 5. 7 Kanser türleri için modül sayılarına göre ilişki tahmincilerin FKT ile gen seviyesindeki kesinlik skorları (p-değeri<0,05)

	BRCA				GBM				KIRC				LUSC				SKCM			
	7	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**	**7**	**8**	**9**	**Ort.**
Adjacency	0,86	0,63	0,33	0,61	0,57	0,63	0,56	0,58	0,29	0,50	0,33	0,37	0,57	0,75	0,56	0,63	0,86	0,63	0,56	0,68
Kendall Tau KKD	0,57	0,50	0,44	0,51	0,71	0,63	0,67	0,67	0,43	0,25	0,33	0,34	0,57	0,38	0,33	0,43	0,71	1,00	0,56	**0,757**
Pearson KKD	0,71	0,50	0,56	0,59	0,43	0,75	0,44	0,54	0,43	0,38	0,33	0,38	0,57	0,63	0,44	0,55	0,71	0,75	0,67	0,71
Spearman KKD	0,43	0,38	0,56	0,45	0,57	0,38	0,44	0,46	0,57	0,25	0,11	0,31	0,57	0,25	0,33	0,38	0,57	0,63	0,67	0,62
BS	0,57	0,50	0,44	0,51	0,57	0,63	0,67	0,62	0,43	0,63	0,33	0,46	0,57	0,75	0,67	**0,66**	0,86	0,63	0,56	0,68
Ampirik	0,71	0,75	0,56	0,67	0,71	0,50	0,56	0,59	0,57	0,38	0,11	0,35	0,71	0,63	0,44	0,59	0,86	0,63	0,56	0,68
ÇYT	0,43	0,50	0,33	0,42	0,86	0,50	0,44	0,60	0,43	0,25	0,33	0,34	0,57	0,38	0,44	0,46	0,86	0,63	0,56	0,68
MM	0,86	0,38	0,33	0,52	0,86	0,75	0,56	0,72	0,71	0,25	0,33	0,43	0,43	0,75	0,44	0,54	0,86	0,75	0,67	**0,758**
SG	0,71	0,63	0,56	0,63	0,86	0,88	0,67	**0,80**	0,43	0,25	0,22	0,30	0,71	0,63	0,44	0,59	0,71	0,63	0,67	0,67
Shrink	1,00	0,88	0,67	**0,85**	0,71	0,63	0,56	0,63	0,57	0,50	0,44	**0,51**	0,57	0,63	0,67	0,62	1,00	0,63	0,56	0,73

5.4 Modül Sayısının Belirlenmesi

Literatürde, WGCNA tarafından analiz edilen veri setleri yaklaşık 1.000 veya daha fazla gen/protein içerir ve bu genler/proteinler farklı sayıda modülde kümelenir. Örneğin, 862 protein içeren Drosophila veri seti 28 modülde kümelenmiştir ve 2.050 protein içeren Maya (Yeast) veri kümesi 44 modülde kümelenmiştir [78]. Bu çalışmada analiz edilen veri kümelerindeki gen/protein sayısı 169 ile 173 arasında değişmektedir ve bu gen/protein sayıları literatürdeki birçok çalışmaya kıyasla çok küçüktür. Bu yüzden, veri kümesinin boyutuna göre ideal modül sayısını belirlemek için aşağıdaki adımlar uygulanmıştır.

1. İlk adımda, deneysel verilere yaklaşan yapay gen ifade verileri üreten bir biyolojik ağ oluşturma aracı olan SynTReN [4] paketini kullanarak e.coli ve maya (yeast) organizmalarından simüle edilmiş gen ekspresyon veri kümelerini ürettik. Oluşturulan veri kümelerinin özellikleri ve veri kümesi üretiminde kullanılan SynTReN parametreleri sırasıyla Çizelge 5.8 ve 5.9'da verilmiştir.

Çizelge 5. 8 Oluşturulan veri kümelerinin özellikleri

Veri seti	Topoloji	Örnek Sayısı	Gen Sayısı
ecoli100	E.coli	200	100
ecoli200			200
ecoli300			300
yeast100	Maya (yeast)		100
yeast200			200
yeast300			300

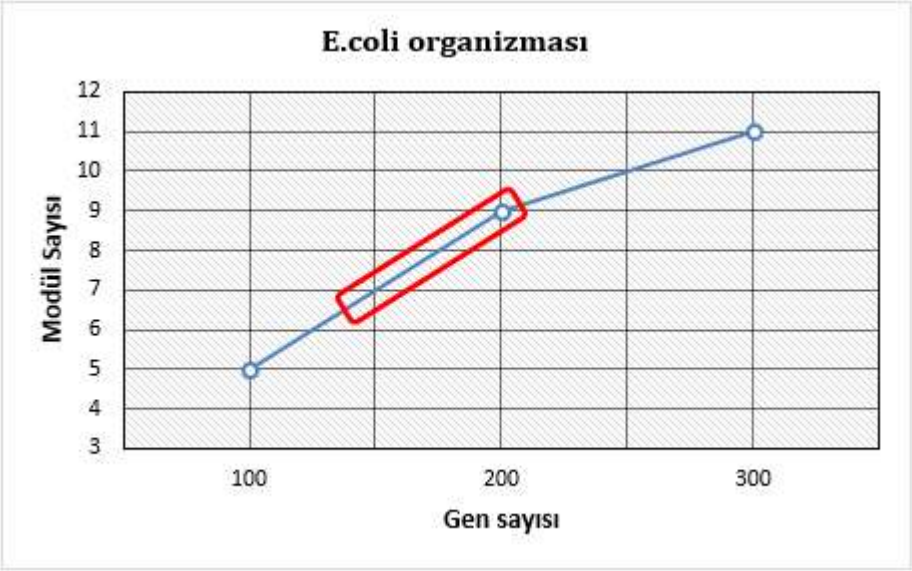
Çizelge 5. 9 SynTReN parametreleri

Parametre Adı	**Değer**
Örnek sayısı (Number of experiments)	200
Gen sayısı (Number of nodes)	100'den 300'e kadar100'er arttırarak

SynTReN paketindeki bulunan diğer parametreler için varsayılan değerler ve gen ifade verisi üretimi için SynTReN sürüm 1.2'i kullanılmıştır.

2. İkinci adımda, WGCNA [22] R paketinin varsayılan parametre ayarlarını kullanarak önceki adımda oluşturulan gen ifade veri kümelerini analiz ettik. Bu analize göre, yaklaşık 150-200 gen içeren veri setlerinin, ölçek içermeyen (scale-free) topoloji ile 7-8-9 modül ürettiklerini gördük. WGCNA aracı tarafından rapor edilen ölçeğin topolojiyle ilişkili r^2 puanını ağın serbest bir topoloji sergilediğini değerlendirmek için kullandık. $r^2 > 0,75$ puanı, ağ topolojisinin birçok biyolojik ağda olduğu gibi ölçeksiz olduğu anlamına gelir [22]. Ayrıca, veri kümesindeki gen sayısı arttıkça, aynı ölçek içermeyen topoloji eşik değerine ulaşmak için modül sayısı artar. Modül sayısını belirleme yaklaşımının daha iyi anlaşılması için Şekil 5.3 ve 5.4 eklenmiştir.

   Şekil 5.3'te ve 5.4'te görüldüğü üzere bu çalışmada analiz edilen veri kümeleri 169-173 protein içerdiğinden, ilişkilendirme tahmincilerinin kanser veri setleri üzerindeki performanslarını değerlendirmek için 7-8-9 sayıda modül içeren biyolojik ağlar incelenmek üzere seçilmiştir.



Şekil 5. 3 E. coli organizması için veri kümesindeki gen sayısına göre üretilen modüllerin sayısı (Ölçek içermeyen (scale-free) topoloji puanı r2> 0,75)

Şekil 5. 4 Maya (yeast) organizması için veri kümesindeki gen sayısına göre üretilen modüllerin sayısı (Ölçek içermeyen (scale-free) topoloji puanı r2> 0,75)

5.5 FKT'ye Dayanan Örtüşme Analizi İçin Eşik Değerlerinin Belirlenmesi

Bir modülde en az beş geni ortak olmak üzere en az iki hastalıkla ilişkili biyolojik-yol varsa, bu modülün (hastalıkla ilişkili bir modül) doğru bir seçim olduğunu varsaymak için gerekçemiz, istatistiksel ve sayısal analizimizin doğruluğunu arttırmaktır. Şekil 5.5'te görüldüğü gibi, DisGeNET genleri ile incelemiş olduğumuz kanser veri kümeleri arasındaki ortalama örtüşen gen sayısı, 7 ile 9 arasında değişen modül sayıları için 1 ile 4 arasında değişmektedir. Dolayısıyla örtüşme oranının değerlendirilmesinde eşik değeri olarak 5 genin seçilmesi ortalama örtüşen gen sayılarına göre (1-4) daha sıkı bir eşik değeri sağlamaktadır. Bu analize dayanarak, hastalıkla ilişkili genlerin geliştirilen yöntem ile saptanması, bu eşik değerleri ile yapmış olduğumuz seçim kriterlerimizin kesinliğini ve doğruluğunu göstermektedir.

DisGeNET'ten alınan hastalıkla ilişkisi onaylanmış genlerle kanser veri kümelerindeki toplam örtüşen gen sayıları; BRCA için 15; GBM için 28; LUSC için 7; KIRC için 14; SKCM için 25'dir.

Şekil 5. 5 DisGeNET ve kanser veri kümeleri arasındaki modül büyüklüğüne göre değişen ortalama örtüşen gen sayısı.

5.6 Bölüm Sonuçları

Geliştirilen yöntemle yapılan analiz işlemleri sonucunda ağ çıkarım algoritmalarında kullanılan ilişki tahmincilerinin farklı kanser türlerine ait proteomik verileri üzerindeki başarımı WGCNA, DepEst ve MINET R paketleri kullanılarak biyolojik-yol ve gen seviyesinde incelenmiştir.

Bölüm 5.3'te açıklandığı gibi tamamlayıcı bir sisteme sahip olmak için sıkı kesme işlemleriyle 2 seviyeli bir değerlendirme işlemi (yani biyolojik-yol ve gen seviyesi) tasarladık. Biyolojik-yol seviyesindeki değerlendirmede, hastalıkla ilişkili modülleri araştırdık. Daha sonra, gen seviyesindeki değerlendirmede, hastalıkla ilişkili modüller üzerinde odaklandık ve bu hastalık ile ilişkili modüllerdeki genler ile daha önce rapor edilen hastalıkla ilişkili genleri (DisGeNET'ten alınan) arasındaki örtüşme oranlarını inceledik.

Bölüm 5.4.1 ve 5.4.2'de biyolojik-yol ve gen seviyesindeki analiz sonuçları her bir kanser türü için ayrı ayrı açıklanmıştır.

5.6.1 Biyolojik-Yol Seviyesi Analiz Sonuçları

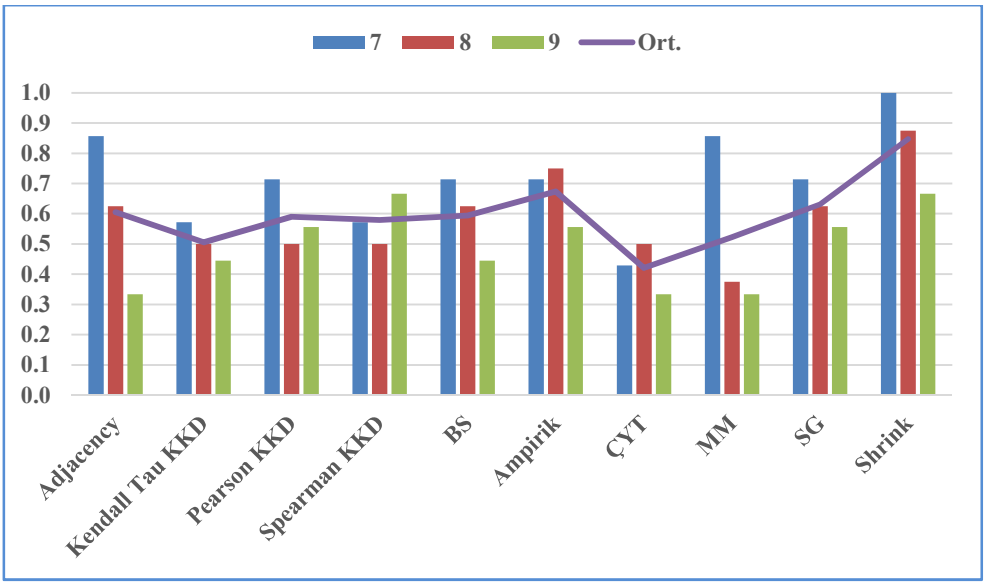
Çizelge 5.6'dan görüleceği gibi, ortalama performans skoru (Ort.) ile belirlenen ilişkilendirme tahmin edicilerinin başarım sıralamaları, modül boyutlarına ve kanser tipine göre değişmektedir. Biyolojik-yol seviyesi analizinde, kanser türüne göre en iyi performans gösteren yöntemler şöyledir: BRCA için Shrink, GBM için SG, KIRC için Shrink, LUSC için BS ve SKCM için MM. Sonuçlardan da görüleceği üzere biyolojik-yol seviyesi analizinde ortalama kesinlik skorlarına göre en başarılı ilişki tahmincini veri kümesine göre değişmektedir. Fakat MI tabanlı ilişki tahmincileri BRCA, GBM, KIRC, LUSC veri kümelerinde diğer yöntemleri geride bırakmışlardır. Ancak, sadece SKCM veri kümesinde MM ve Kendall Tau KKD yöntemleri çok yakın skorlar elde ederek başarılı bulunmuştur.

Çizelge 5.6 ve Şekil 5.6'dan de anlaşılacağı üzere BRCA veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: Shrink, Ampirik, Adjacency, BS, Pearson KKD, Spearman KKD, MM, Kendall KKD ve ÇYT.

BRCA veri seti için Shrink yöntemi, hastalık ile ilişkili alt ağların belirlenmesi açısından 0.85 ortalama kesinlik puanıyla en doğru yöntemdir. Bu yöntemi Ampirik, SG, Adjacency, BS, Pearson KKD, Spearman KKD yöntemleri sırasıyla 0,67; 0,63; 0,61; 0,59; 0,59; 0,58'lik daha düşük ortalama kesinlik skorlarıyla takip etmektedir. Ayrıca, MM ve Kendall Tau KKD yöntemleri 0,52; 0,51 gibi nispeten düşük ortalama kesinlik skoruna sahip olurken, ÇYT metodu 0,42 ile en düşük ortalama kesinlik skorunu elde etmiştir.

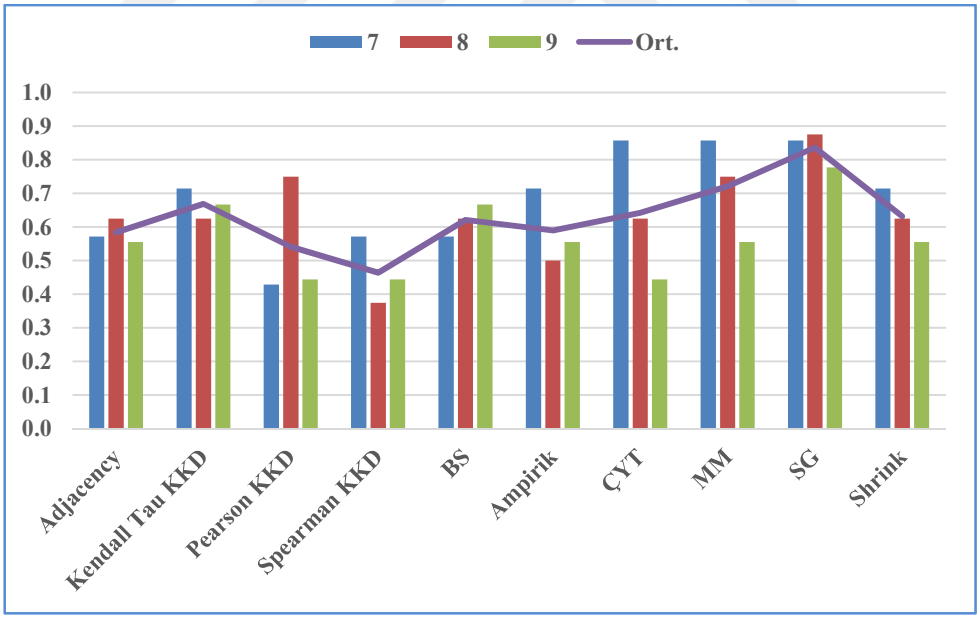
Çizelge 5.6 ve Şekil 5.7'den de görüleceği üzere GBM veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: SG, MM, Kendall Tau KKD, ÇYT, Shrink, BS, Ampirik, Adjacency, Pearson KKD ve Spearman KKD.

GBM veri seti için SG yöntemi, hastalık ile ilgili alt ağların belirlenmesi açısından 0,84 ortalama kesinlik puanıyla en doğru ilişki yöntemidir. Bu yöntemi MM, Kendall Tau KKD, ÇYT, Shrink, BS yöntemleri sırasıyla 0,72; 0,67; 0,64; 0,63; 0,62'lik daha düşük ortalama kesinlik skorlarıyla takip etmektedir. Ayrıca, Ampirik, Adjacency ve Pearson KKD metotları sırasıyla 0,59; 0,58; 0,54 gibi daha da düşük ortalama kesinlik skorlarına sahip olurken, Spearman KKD yöntemi 0,46 ile en düşük ortalama kesinlik skorunu elde etmiştir.



Şekil 5. 6 İlişki tahmincilerin BRCA veri seti için biyolojik-yol seviyesindeki başarımları

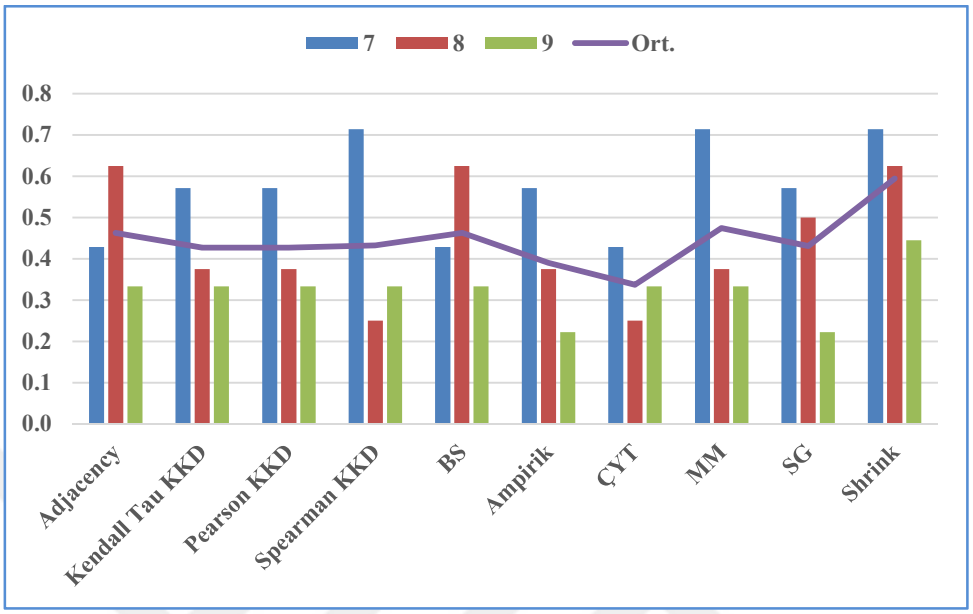
Çizelge 5.6 ve Şekil 5. 8'den de anlaşılacağı üzere KIRC veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: Shrink, MM, Adjacency, BS, Spearman KKD, SG, Kendall Tau KKD, Pearson KKD, Ampirik ve ÇYT.



Şekil 5. 7 İlişki tahmincilerin GBM veri seti için biyolojik-yol seviyesindeki başarımları

KIRC veri seti için Shrink yöntemi, hastalık ile ilgili alt ağların belirlenmesi açısından, 0,59 ortalama kesinlik puanıyla en doğru ilişki yöntemi olarak tanımlanmaktadır. Bu yöntemi MM, Adjacency, BS, Spearman KKD, SG, Kendall Tau KKD, Pearson KKD yöntemleri sırasıyla 0,47; 0,46; 0,46; 0,43; 0,43; 0,43; 0,43'lük daha düşük ortalama kesinlik

puanlarıyla izlemektedir. Ayrıca, Ampirik yöntem 0,39 gibi daha da düşük ortalama kesinlik skoru alırken, KDE yöntemi 0,34 kesinlik skoru ile en düşük puana sahip olduğu görülmektedir.

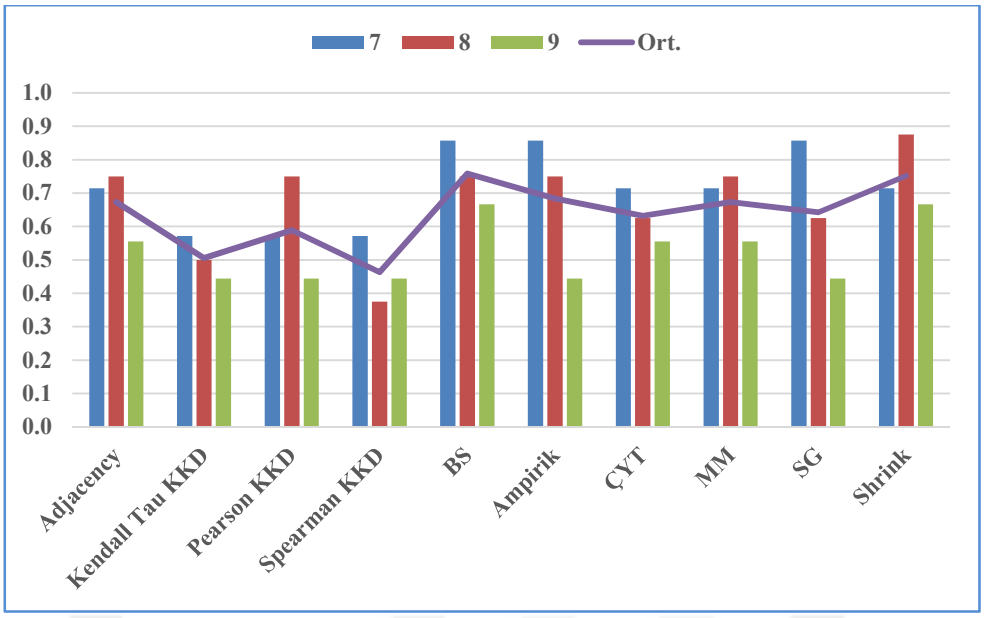


Şekil 5. 8 İlişki tahmincilerin KIRC veri seti için biyolojik-yol seviyesindeki başarımları

Çizelge 5.6 ve Şekil 5.9'dan da görüleceği üzere LUSC veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: BS, Shrink, Ampirik, Adjacency, MM, SG, ÇYT, Pearson KKD, Kendall Tau KKD and Spearman KKD.

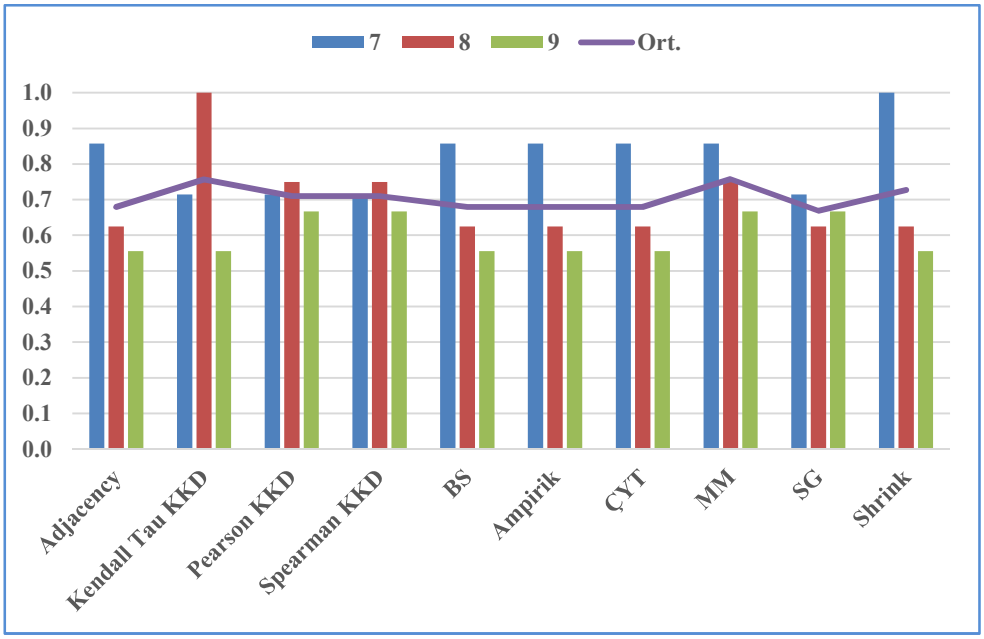
LUSC veri seti için hastalık ile ilgili alt ağları BS ve Shrink yöntemleri, sırasıyla, 0,76 ve 0,75'lik çok yakın ortalama kesinlik değerleri ile tanımlayabildiler. Bu yöntemleri Ampirik, Adjacency, MM, SG, ÇYT metotları sırasıyla 0,68; 0,67; 0,67; 0,64; 0,63 gibi daha düşük ortalama kesinlik skorlarıyla izlemiştir. Ek olarak, Pearson KKD ve Kendall Tau KKD metotları 0,59; 0,51 gibi daha düşük ortalama kesinlik puanları alırken, Spearman KKD yöntemi 0,46 ile en düşük ortalama kesinlik skoruna sahip olmuştur.

Çizelge 5.6 ve Şekil 5.10'dan de anlaşılacağı üzere SKCM veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: MM, Kendall Tau KKD, Shrink, Pearson KKD, Spearman KKD, Adjacency, BS, Ampirik, ÇYT ve SG.



Şekil 5. 9 İlişki tahmincilerin LUSC veri seti için biyolojik-yol seviyesindeki başarımları

SKCM veri seti için hastalık ile ilgili alt ağları MM ve Kendall yöntemleri sırasıyla, 0,758 ve 0,757'lik çok az farklı ortalama kesinlik değerleriyle tanımlayabildiler. Bu yöntemleri Shrink, Pearson KKD, Spearman, Adjacency, BS, Ampirik, ÇYT ve SG yöntemleri sırasıyla 0,73; 0,71; 0,71; 0,68; 0,68; 0,68; 0,68 ve 0,67'lik daha düşük ortalama kesinlik puanlarıyla takip etmiştir.



Şekil 5. 10 İlişki tahmincilerin SKCM veri seti için biyolojik-yol seviyesindeki başarımları

Sonuçlardan da görüleceği üzere biyolojik-yol seviyesindeki analiz işlemleri sonucunda KB tabanlı ilişki tahmin yöntemleri korelasyon tabanlı yöntemlere karşı daha başarılı olmuştur.

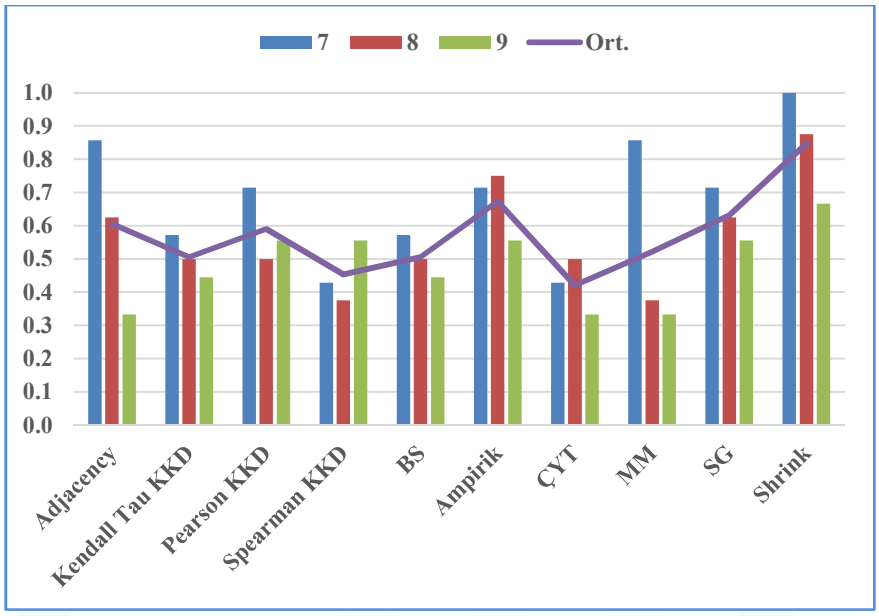
5.6.2 Gen-Seviyesi Analiz Sonuçları

Çizelge 5.7'den de görüldüğü üzere, gen-seviyesi analizindeki ortalama kesinlik skorlarına göre en iyi performans gösteren ilişki tahminci her veri seti için değişmektedir. Gen seviyesi analizinde kanser türüne göre en iyi performans gösteren yöntemler şöyledir: BRCA için Shrink, GBM için SG, KIRC için Shrink, LUSC için BS ve SKCM için MM. Ayrıca, biyolojik-yol seviyesi analizinde olduğu gibi, her bir veri seti için en başarılı ilişki tahminci değişse de, MI tabanlı ilişki tahmincileri BRCA, GBM, KIRC, LUSC veri kümelerinde daha başarılı bulunmuştur. Sadece SKCM veri setinde, MM ve Kendall Tau KKD yöntemi hem biyolojik-yol seviyesi hem de gen seviyesi (p-değeri<0,05) analizine göre çok yakın ortalama kesinlik değerlerine sahip olup, başarılı bulunmuştur.

Çizelge 5.7 ve Şekil 5.11'den de anlaşılacağı üzere BRCA veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: Shrink, Ampirik, SG, Adjacency, Pearson KKD, MM, Kendall Tau KKD, BS, Spearman KKD ve ÇYT.

BRCA veri seti için Shrink ilişki tahmincisi hastalıkla ilişkili genleri tanımlamak açısından 0,85 ortalama kesinlik değeriyle en başarılı yöntem olarak bulunmuştur. Bu yöntemi Ampirik, SG, Adjacency, Pearson KKD yöntemleri sırasıyla 0,67; 0,63; 0,61; 0,59 gibi daha düşük ortalama kesinlik skorlarıyla takip etmektedir. Ayrıca, MM, Kendall Tau KKD, BS yöntemleri sırasıyla 0,52; 0,51; 0,51 gibi daha da düşük ortalama kesinlik skorları elde ederken, Spearman KKD ve ÇYT yöntemleri sırasıyla 0,45 ve 0,42 ile en düşük ortalama kesinlik puanlarını elde etmiştir.

Çizelge 5.7 ve Şekil 5.12'den de görüldüğü üzere GBM veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: SG, MM, Kendall Tau KKD, Shrink, BS, ÇYT, Ampirik, Adjacency, Pearson KKD ve Spearman KKD.

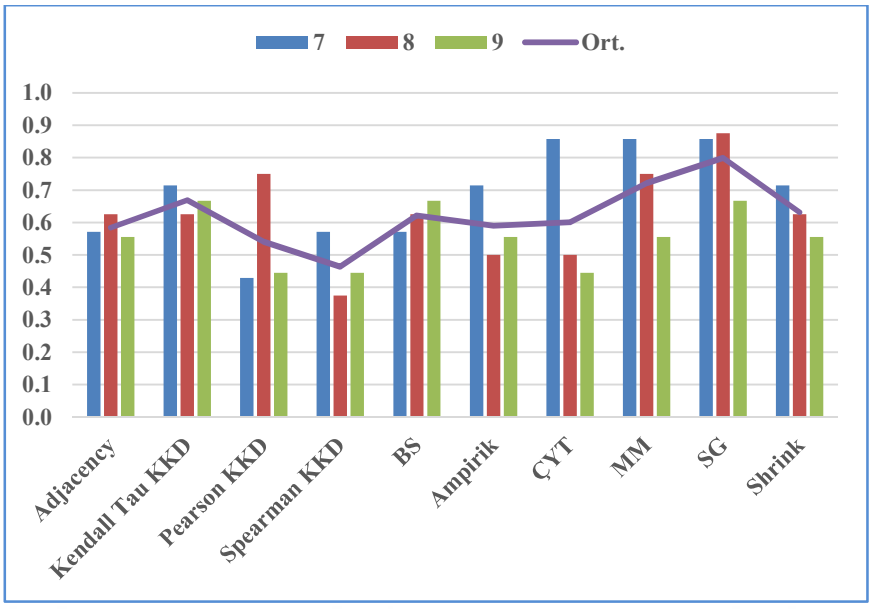


Şekil 5. 11 İlişki tahmincilerin BRCA veri seti için gen seviyesindeki başarımları

GBM veri seti için SG ilişki tahmincisi hastalıkla ilişkili genleri tanımlamak açısından 0,80 ortalama kesinlik değeriyle en başarılı yöntem olarak bulunmuştur. Bu yöntemi MM, Kendall Tau KKD, Shrink, BS, ÇYT yöntemleri sırasıyla 0,72; 0,67; 0,63; 0,62; 0,60 gibi daha düşük ortalama kesinlik skorlarıyla takip etmektedir. Ayrıca, Ampirik, Adjacency, Pearson KKD yöntemleri sırasıyla 0,59; 0,58; 0,54 gibi daha da düşük ortalama kesinlik skorları elde ederken, Spearman KKD 0,46 ile en düşük ortalama kesinlik puanını elde etmiştir.

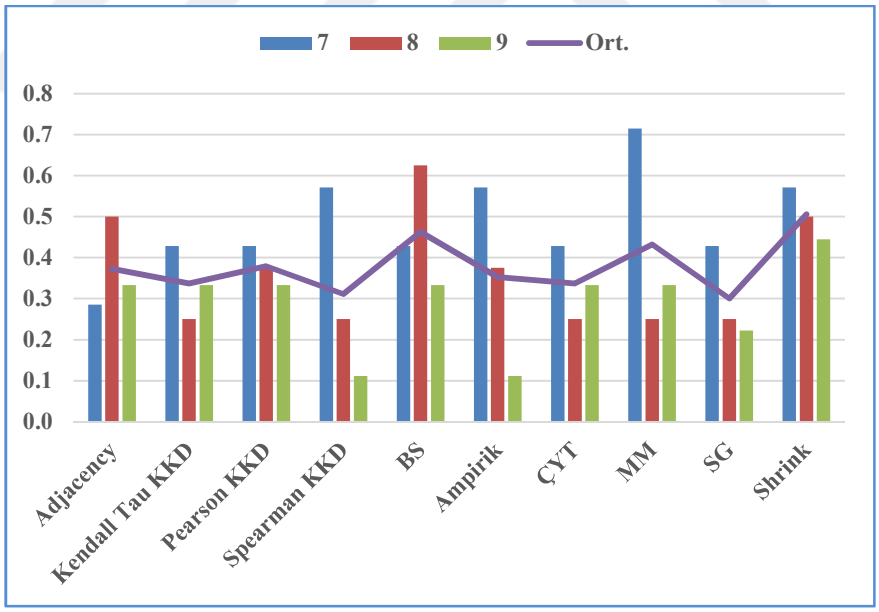
Çizelge 5.7 ve Şekil 5.13'ten de anlaşıldığı üzere KIRC veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: Shrink, BS, MM, Pearson KKD, Adjacency, Ampirik, Kendall Tau KKD, ÇYT, Spearman KKD ve SG.

KIRC veri seti için Shrink ilişki tahmincisi hastalıkla ilişkili genleri tanımlamak açısından 0,51 ortalama kesinlik değeriyle en başarılı yöntem olarak bulunmuştur. Bu yöntemi BS ve MM yöntemleri sırasıyla 0,46; 0,43 gibi daha düşük ortalama kesinlik skorlarıyla takip etmektedir. Ayrıca, Pearson KKD, Adjacency, Ampirik, Kendall Tau KKD, ÇYT yöntemleri sırasıyla 0,38; 0,37; 0,35; 0,34; 0,34 gibi daha da düşük ortalama kesinlik skorları elde ederken, Spearman KKD ve SG yöntemleri sırasıyla 0,31 ve 0,30 ile en düşük ortalama kesinlik puanlarını elde etmiştir.



Şekil 5. 12 İlişki tahmincilerin GBM veri seti için gen seviyesindeki başarımları

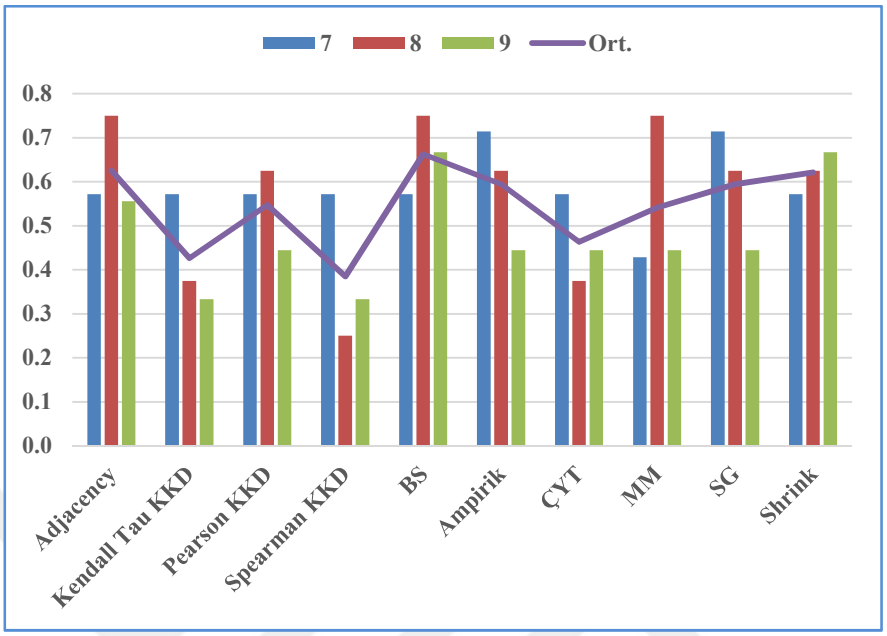
Çizelge 5.7 ve Şekil 5.14'ten de görüldüğü üzere LUSC veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: BS, Adjacency, Shrink, Ampirik, SG, Pearson KKD, MM, ÇYT, Kendall Tau KKD ve Spearman KKD.



Şekil 5. 13 İlişki tahmincilerin KIRC veri seti için gen seviyesindeki başarımları

LUSC veri seti için BS ilişki tahmincisi hastalıkla ilişkili genleri tanımlamak açısından 0,66 ortalama kesinlik değeriyle en başarılı yöntem olarak bulunmuştur. Bu yöntemi Adjacency, Shrink, Ampirik, SG yöntemleri sırasıyla 0,63; 0,62; 0,59; 0,59 gibi daha düşük ortalama kesinlik skorlarıyla takip etmektedir. Ayrıca, Pearson KKD, MM, ÇYT, Kendall

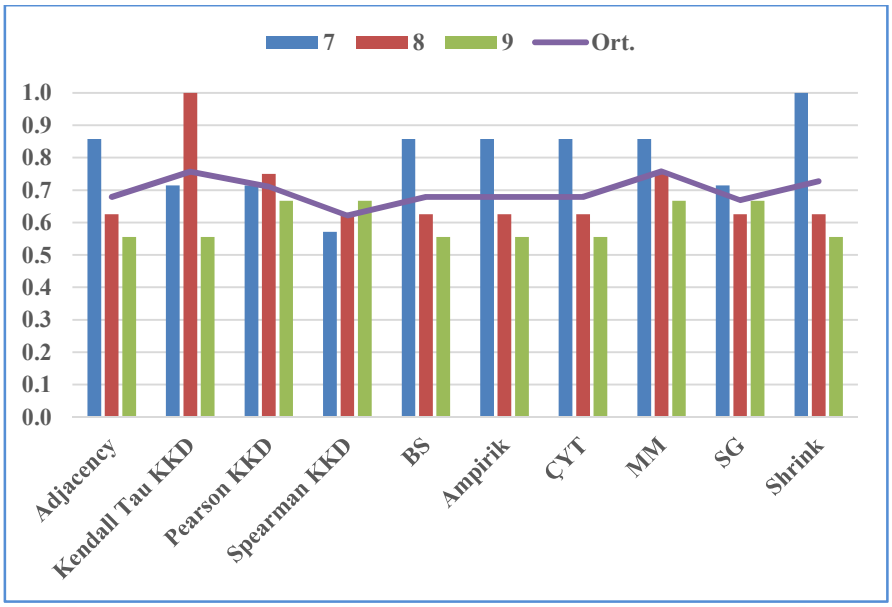
Tau KKD yöntemleri sırasıyla 0,55; 0,54; 0,46; 0,43 gibi daha da düşük ortalama kesinlik skorları elde ederken, Spearman KKD 0,38 ile en düşük ortalama kesinlik puanını elde etmiştir.



Şekil 5. 14 İlişki tahmincilerin LUSC veri seti için gen seviyesindeki başarımları

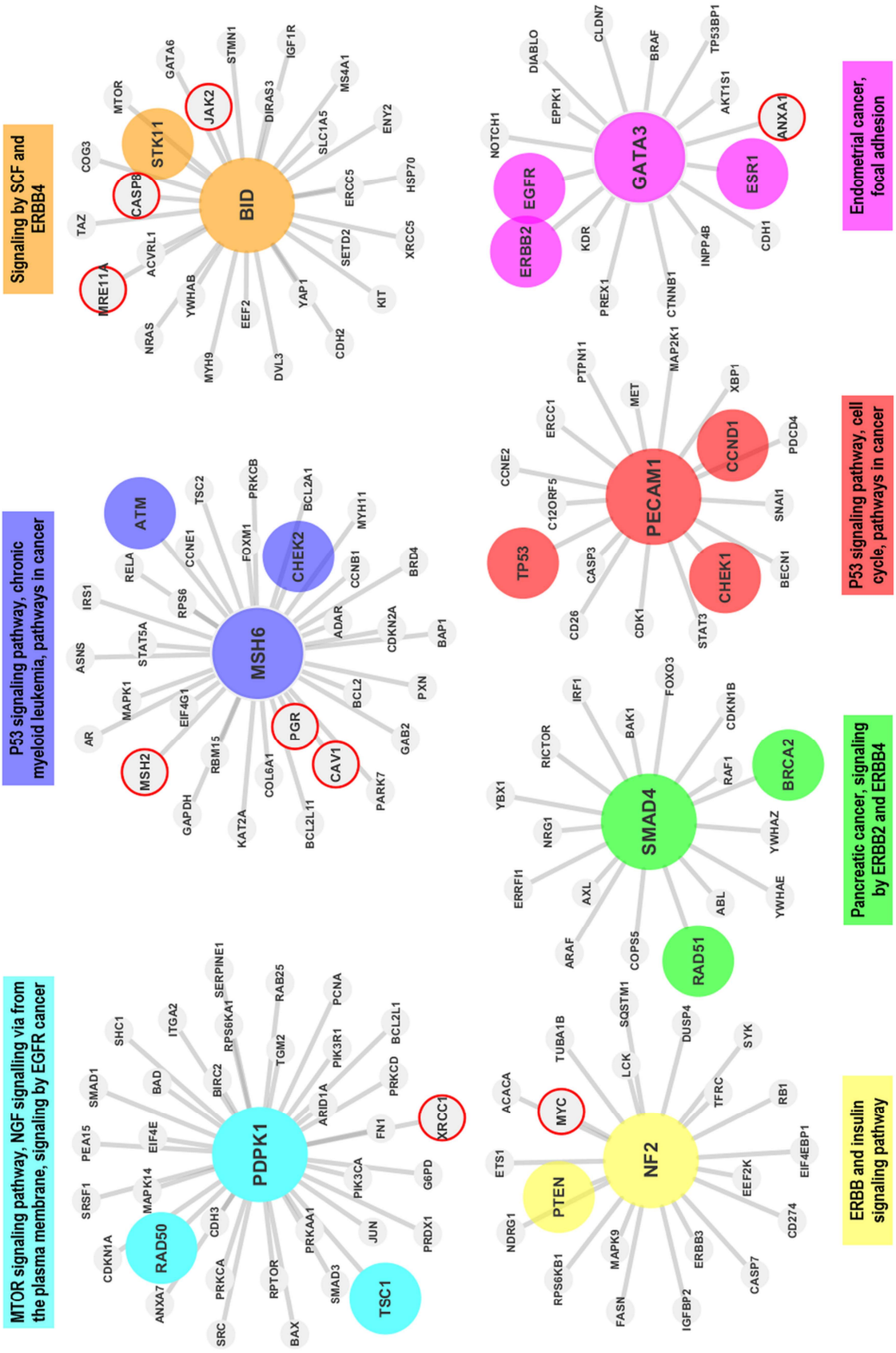
Çizelge 5.7 ve Şekil 5.15'ten de görüldüğü üzere SKCM veri seti için tahmin edicilerin ortalama kesinlik skorlarına göre başarı sıralaması şu şekildedir: MM, Kendall Tau KKD, Shrink, Pearson KKD, Adjacency, BS, Ampirik, ÇYT, SG ve Spearman KKD.

SKCM veri seti için MM ve Kendall Tau KKD ilişki tahmincileri hastalıkla ilişkili genleri tanımlamak açısından sırasıyla 0,758 ve 0,757 gibi ortalama kesinlik değerleriyle en başarılı yöntemler olarak bulunmuştur. Bu yöntemi Shrink, Pearson KKD, Adjacency, BS, Ampirik, ÇYT ve SG yöntemleri sırasıyla 0,73; 0,71; 0,68; 0,68; 0,68; 0,68 ve 0,67 gibi daha düşük ortalama kesinlik skorlarıyla takip etmektedir. Ayrıca, Spearman KKD 0,62 ile en düşük ortalama kesinlik puanını elde etmiştir.

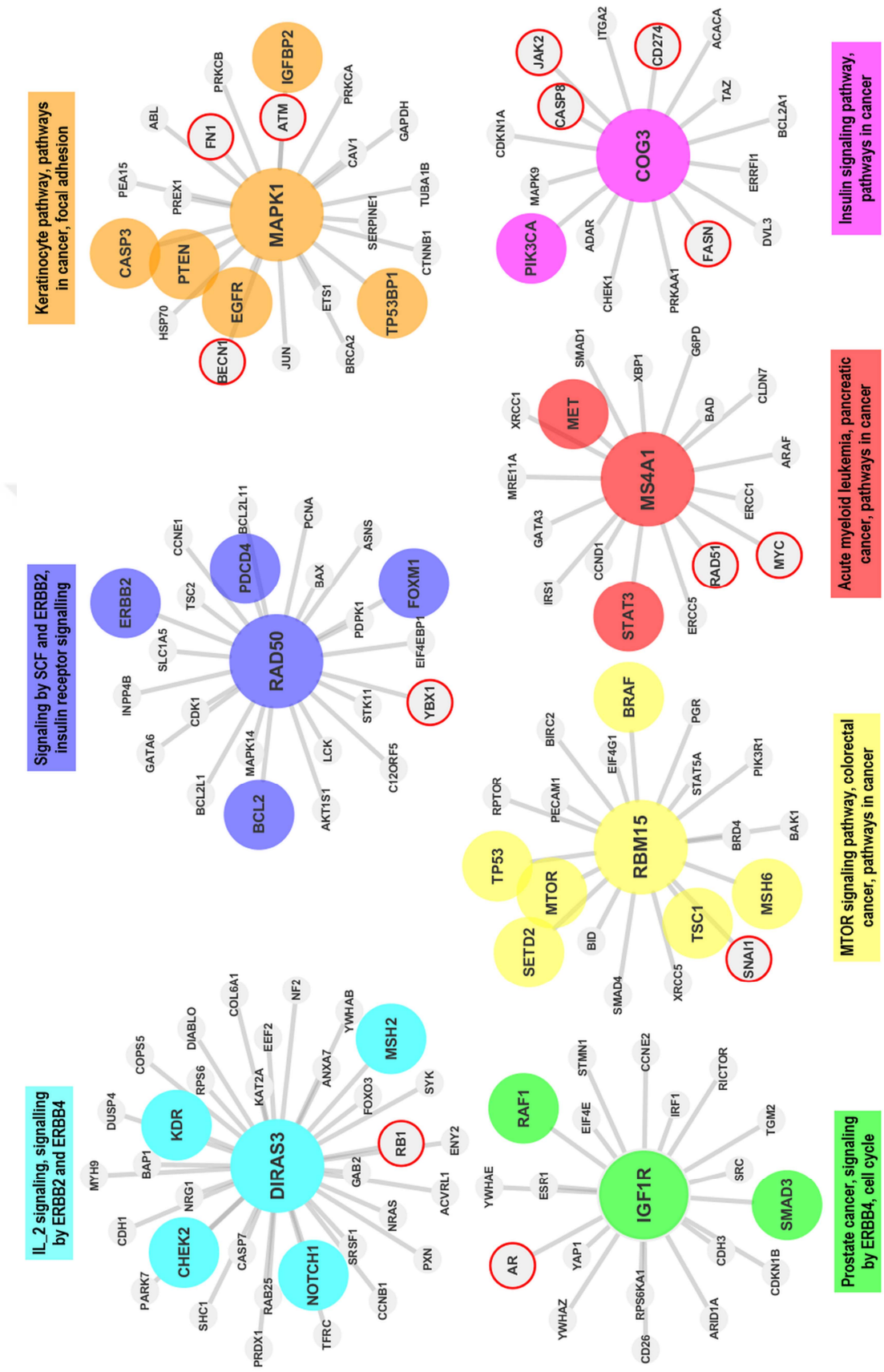


Şekil 5. 15 İlişki tahmincilerin SKCM veri seti için gen seviyesindeki başarımları

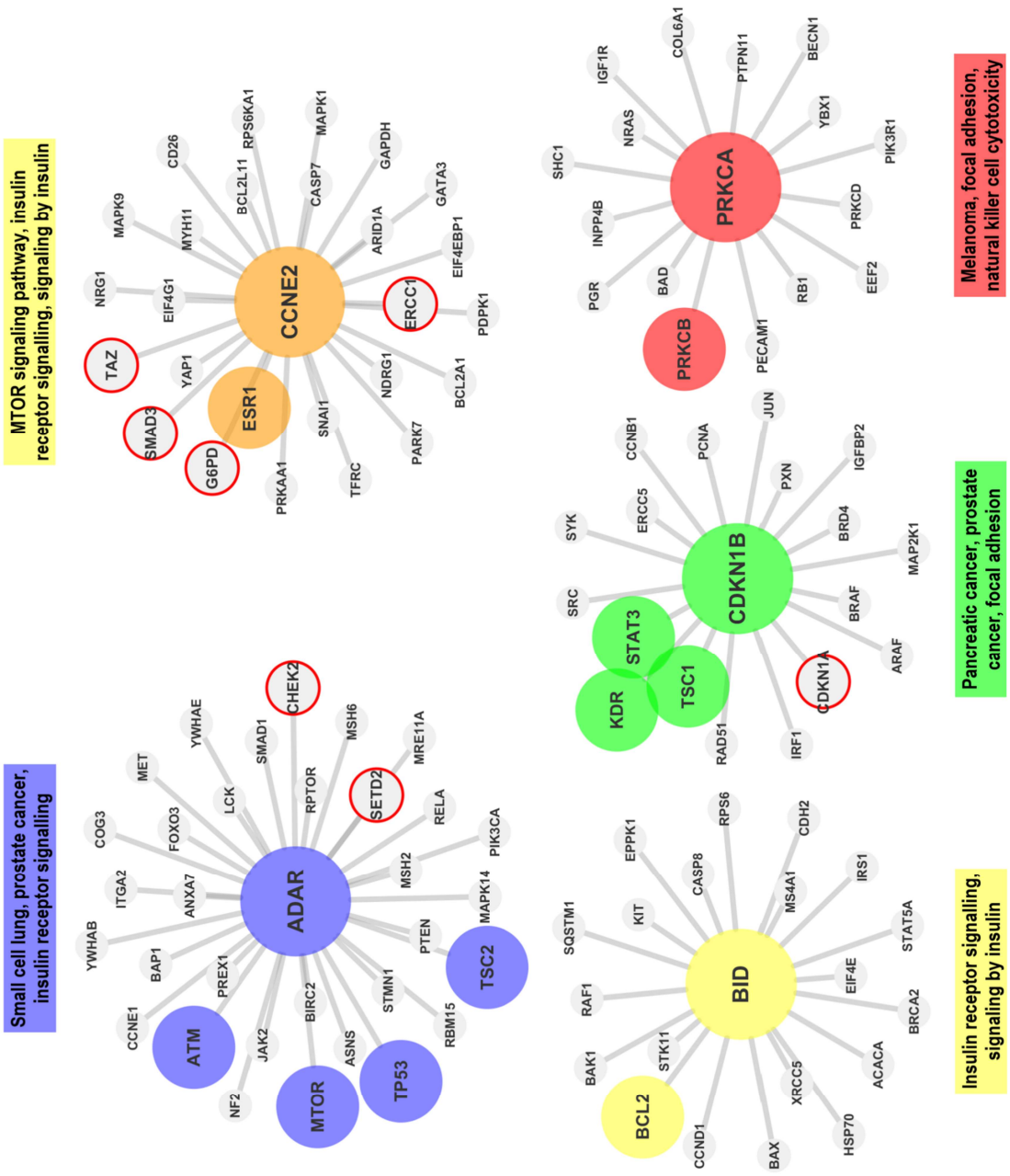
Son olarak, her bir kanser türü için en yüksek kesinlik skorlarını sağlayan ilişki tahmincileri tarafından tanımlanan modüllerin merkezi genlerinin alt ağları Cytoscape [79] ile oluşturularak Şekil 5.16 - 5.20'de gösterilmiştir. Bu alanda çalışan diğer araştırmacıların da değerlendirmesi adına tüm kanser türleri için DisGeNET'te kayıtlı olan ve deneysel olarak hastalıkla ilişkisi olduğu onaylanan genler, Şekil 5.16 - 5.20'de renkli ve büyük düğümlerle gösterilmiştir. Bunlar arasında, renkli olmayan ancak kırmızı bir çerçeveye sahip olan genlerin PMID değeri 1'dir. Şekil 5.16 – 5.20'deki gri renkli düğümler için DisGeNET'te herhangi bir bilgiye rastlanmamıştır. Ayrıca, her bir modülün ilişkili olduğu ilk üç biyolojik yol (düzeltilmiş p-değeri<0,05 şartını sağlayarak en küçük p-değeri alan ilk üç modül), her bir modüle açıklama yapmak için ilgili modülün üstünde veya altında verilmektedir.



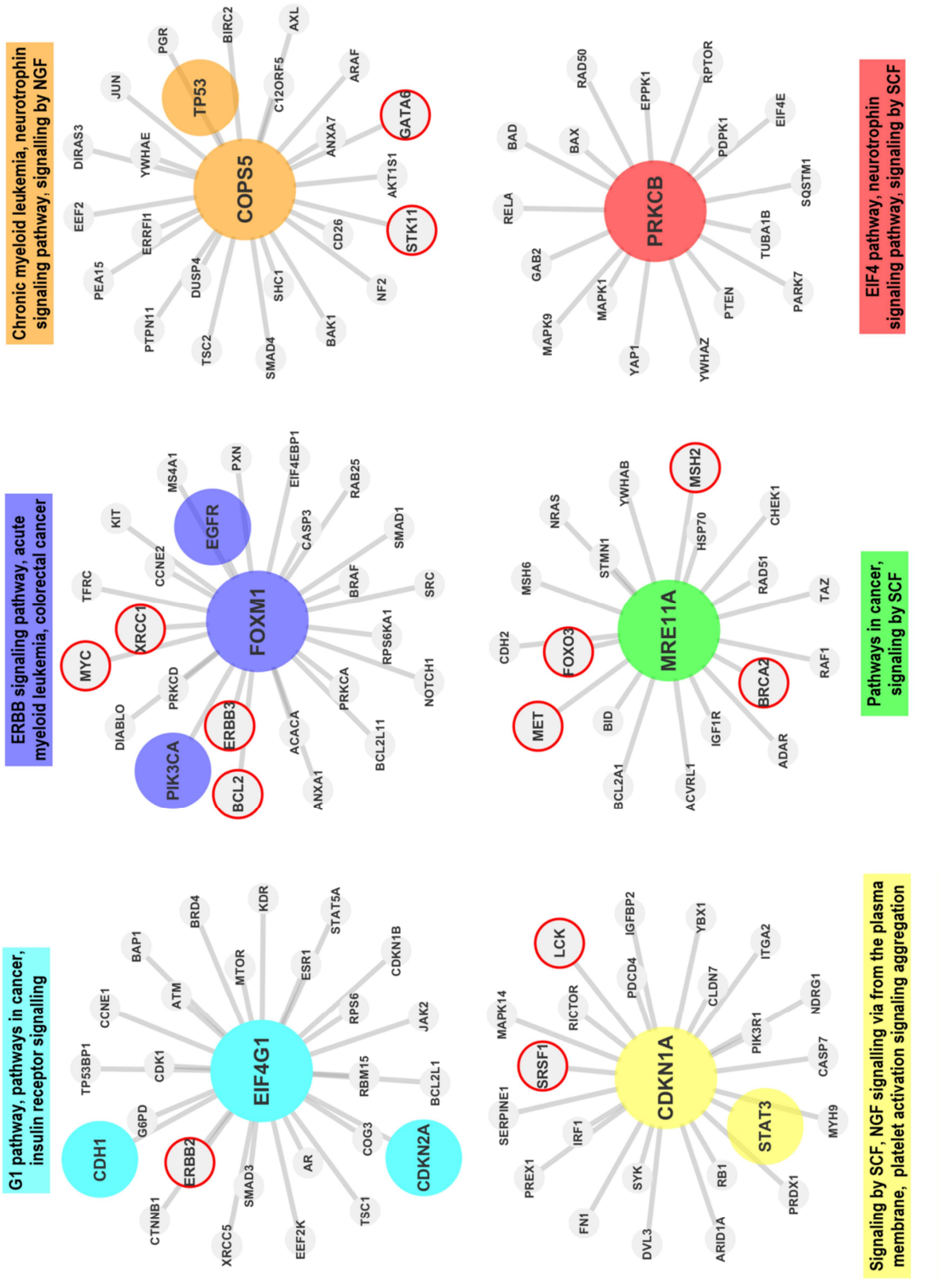
Şekil 5. 16 BRCA veri seti için en başarılı Shrink ilişki tahmin yöntemiyle elde edilen hastalıkla ilişkili alt ağlardaki merkez genler ve komşuları [80].



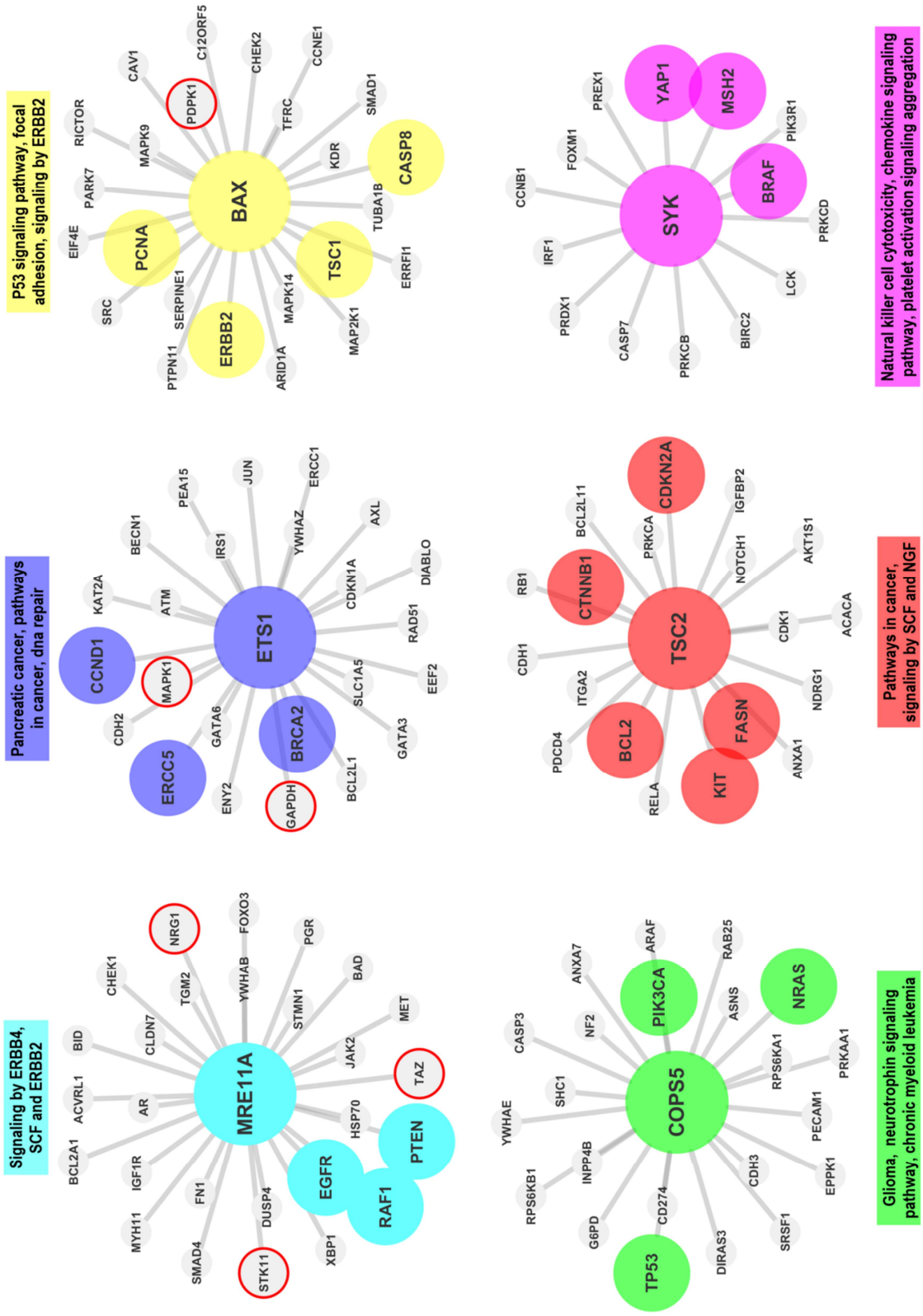
Şekil 5. 17 GBM veri seti için en başarılı SG ilişki tahmin yöntemiyle elde edilen hastalıkla ilişkili alt ağlardaki merkez genler ve komşuları [80].



Şekil 5. 18 KIRC veri seti için en başarılı Shrink ilişki tahmin yöntemiyle elde edilen hastalıkla ilişkili alt ağlardaki merkez genler ve komşuları [80].



Şekil 5. 19 LUSC veri seti için en başarılı BS ilişki tahmin yöntemiyle elde edilen hastalıkla ilişkili alt ağlardaki merkez genler ve komşuları [80].



Şekil 5. 20 SKCM veri seti için en başarılı MM ilişki tahmin yöntemiyle elde edilen hastalıkla ilişkili alt ağlardaki merkez genler ve komşuları [80].

5.6.3 Genel Değerlendirme

Bu bölümde GAÇ algoritmalarında kullanılan dokuz ilişkilendirme tahmincisinin performansları, hem biyolojik-yol hem de gen düzeyinde beş farklı kanser türünün proteomik verileri üzerinde incelenmiştir. Bu değerlendirmeyi yapmak için, Pearson veya Spearman korelasyonuna dayanan ve literatürde birçok farklı çalışmada kullanılmış olan WGCNA'da ilişki matrisi oluşturma prosedürü olan *İlişki* fonksiyonu yerine seçilmiş ilişkilendirme tahmin edicileri kullanılmıştır. Sonuç olarak, Çizelge 5.6 ve 5.7'de verilen kesinlik skorları açısından, KB temelli yöntemlerin, korelasyon-esaslı yöntemlerden ve WGCNA'da varsayılan olarak kullanılan *İlişki* prosedüründen daha iyi sonuçlar vermiştir. Literatürdeki çalışmalara paralel olarak [81], bulgularımız korelasyon temelli yöntemlerin, kanserle ilişkili kompleks moleküler etkileşimler gibi hücre molekülleri arasındaki doğrusal olmayan ilişkileri tahmin etmek için yeterli olmadığını doğrulamıştır.

Shrink ilişki tahmincisi kullanılarak yapılan WGCNA analizi sonucunda MSH6, BID, PDPK1 (PDK1), SMAD4, PECAM1 (CD31), NF2 ve GATA3 merkez genlerinin (proteinlerinin), meme kanseri (BRCA) üzerinde önemli etkileri olabileceği gözlenmiştir. DisGeNET'teki veriler bu genler arasından MSH6 ve GATA3 genlerinin tek bir çalışmada BRCA ile ilişkili olduğunu ve ayrıca, CHEK2, RAD50, TSC1, STK11, ATM, PTEN, BRCA2, RAD51, TPD3, CHEK1, ERBB2, CCND1, ESR1 ve EGFR çoklu çalışmalarda (PMIDs>=2) doğrulanmış olduğunu göstermiştir. Ayrıca, DisGeNET'te rapor edilmemiş olmasına karşın PDPK1 (PDK1) [82], [83] ve SMAD4 [84] genlerinin (proteinlerinin) farklı çalışmalarda BRCA ile ilişkili olduğu belirtilmiştir.

SG ilişki tahmincisi ile yapılan WGCNA analizi sonucunda RAD50, DIRAS3, MAPK1 (ERK2), IGF1R (IGF1R_pY1135Y1136), RBM15, COG3 ve MS4A1 (CD20) hub genlerinin (proteinlerinin), GBM üzerinde önemli etkileri olabileceği bulunmuştur. DisGeNET'teki veriler bu genler arasından IGF1R geninin sadece bir çalışmada GBM ile ilişkili olduğunu ve ayrıca, KDR, CHEK2, MSH2, NOTCH1, ERBB2, BCL2, FOXM1, PDCD4, CASP3, PTEN, EGFR, IGFBP2, TP53BP1, SMAD3, RAF1, MTOR, SETD2, BRAF, TP53, TSC1, MSH6, MET, STAT3 ve PIK3CA genlerinin çoklu çalışmalarda (PMIDs>=2) doğrulanmış olduğunu göstermiştir. Ayrıca, DisGeNET'te bildirilmemiş olmasına rağmen literatürde yapılan bir çalışmada RAD50 [85] geninin (protein) GBM ile ilişkili olduğu bildirilmiştir.

Shrink ilişki tahmincisi kullanılarak yapılan WGCNA analizi sonucunda BID, ADAR (ADAR1), CDKN1B (P27_Pt198), CCNE2 (CYCLINE2) ve PRKCA (PKCALPHA) merkez genlerinin (proteinlerinin) KIRC üzerinde önemli etkileri olabileceği gözlenmiştir. DisGeNET'teki veriler ışığında bu genlerin KIRC ile ilişkili olduğuna dair bir bilgi bulamadık. DisGeNET'teki veriler MTOR, ATM, TSC2, TP53, ESR1, KDR, BCL2, TSC1, STAT3 ve PRKCB genlerinin KIRC ile ilgili olduğunu çoklu (PMIDs>=2) çalışmalarda doğrultusunda göstermiştir.

BS ilişki tahmincisi ile yapılan WGCNA analizi sonucunda FOXM1, EIF4G1 (EIF4G), COPS5 (JAB1), MRE11A (MRE11), CDKN1A (P21) ve PRKCB (PKCPANBETAII_pS660) hub genlerinin (proteinlerinin), LUSC üzerinde önemli etkileri olabileceği bulunmuştur. DisGeNET'teki veriler bu genler arasından EIF4G1 (EIF4G) geninin (proteinin) iki çalışmada LUSC ile ilişkili olduğunu, CDH1, PIK3CA, CDKN2A, TP53, EGFR ve STAT3 genlerinin çoklu çalışmalarda hastalık ilişkili genler olarak doğrulandığını göstermiştir. Ayrıca, DisGeNET'te bildirilmemiş olmasına son yıllarda yapılan çalışmalarda FOXM1 [86], [87] ve CDKN1A (P21) [88] genlerinin (proteinlerinin) LUSC ile ilişkili olduğu gösterilmiştir.

MM ilişki tahmincisi kullanılarak yapılan WGCNA analizi sonucunda ETS1, MRE11A (MRE11), BAX, TSC2 (TUBERIN_pT1462), COPS5 (JAB1) ve SYK merkez genlerinin (proteinlerinin) SKCM üzerinde önemli etkileri olabileceği gözlenmiştir. DisGeNET'teki veriler ışığında bu genlerin SKCM ile ilişkili olduğuna dair bir bilgi bulamadık. Fakat DisGeNET'te RAF1, EGFR, ERCC5, PTEN, BRCA2, CCND1, PCNA, ERBB2, CASP8, TSC1, PIK3CA, TP53, NRAS, KIT, BCL2, CTNNB1, FASN, CDKN2A, YAP1, BRAF ve MSH2 çoklu çalışmalarda hastalık genleri olarak vurgulanmıştır. Ayrıca, DisGeNET'te bildirilmemiş olmasına rağmen, son çalışmalarda ETS1 [89] geninin SKCM ile ilişkili olduğu gösterilmiştir.

Sonuç olarak, DisGeNET'te yer almamasına rağmen, literatürde güncel çalışmalarda hastalık ile ilişkili olduğu tespit edilen genler (RAD50, FOXM1, PDK1, CDKN1A, SMAD4, ETS1 [82], [83], [84], [85], [86], [87], [88], [89]) geliştirilen yöntem ile tespit edilebilmiştir ve ilgili çalışmalar, bu genlerin (proteinler) kansere bağlı süreçlerle ilişkili olduğunu doğrulamaktadır. Du vd. [82] çalışmalarında, PDK1'i BRCA için

potansiyel bir tedavi edici hedef gen olarak tanımlamıştır. Ayrıca, Dupuy vd. [83] araştırmalarında PDK1'i, BRCA'da metabolizmanın ve metastatik potansiyelin bir anahtar regülatörü olarak belirlemişlerdir. Liu vd. [84] yapmış oldukları çalışmada, SMAD4 protein düzeyinin değerlendirilmesinin BRCA hakkında ek prognostik bilgi sağlayabileceğini belirtmişlerdir. Mishima vd. [85] araştırmalarında, MRE11-RAD50-NBS1 kompleks inhibitörünün GBM'de radiosensitiviteyi etkin bir şekilde artırabildiğini bulmuşlardır. Sun vd. [87] çalışmalarında, FOXM1 ekspresyonunun prognostik önemini LUSC'da göstermişlerdir. Zhang vd. [86] yapmış oldukları araştırma sonucunda FOXM1'i LUSC'nin yeni bir biyobelirleyicisi olarak nitelendirmişlerdir. Fukazawa vd. [88] çalışmalarında, SOX2'nin LUSC üzerindeki tümörijenik etkisinin CDKN1A'nın kısmen baskılanmasıyla gerçekleştiğini bildirmişlerdir. Keehn vd. [89] araştırmalarında, ETS1'in invazif SKCM'nin patogenezinde önemli olabileceğini belirtmişlerdir. Bu sonuçlar geliştirdiğimiz yöntemin istatistiksel olarak başarılı bulduğu ilişki tahmincisi ile oluşturulan alt ağların biyolojik olarak ta anlamlı olduğunu göstermektedir.

BÖLÜM 6

ÇIKARIMLANAN GEN VE PROTEİN AĞLARININ ENTEGRASYONU

Bu bölümde, Bölüm 5'te detayları verilen ve bu tez çalışmasında geliştirilen yöntemle çıkarımlanmış gen ve protein ağlarının entegrasyonu için sunulan yaklaşım anlatılmıştır.

6.1 Kullanılan Veri Seti

Bu bölümde akciğer kanserine ait The Cancer Genome Atlas (TCGA) tarafından sağlanan gen ifade verileri ve TCPA tarafından sağlanan proteomik verileri UCSC Xena [90] web tarayıcısı kullanılarak indirilmiştir. Genişleme süreci devam etmekte olan UCSC Xena veri setleri şu anda Pan-Kanser veri kümelerinin yanı sıra 35'ten fazla kanser türünden 1400'den fazla veri kümesine ev sahipliği yapmaktadır [91]. Bu veri kümelerinden literatürde sıklıkla kullanılanları TCPA, GDC (Genomic Data Commons), ICGC (International Cancer Genome Consortium), GTEx (Genotype-Tissue Expression), TARGET (Therapeutically Available Research to Generate Effective Treatments), TOIL (Scalable and Efficient Workflow Engine) gibi veri kataloglarıdır. Gen ifade veri setinde toplam 155 hastadan alınmış örnekler için 17.815 gene ait ifade verisi yer almaktadır. Proteomik veri setinde ise 328 hastadan alınmış örnekler için 277 proteine ait veri yer almaktadır. Bu farklı veri setlerinden, hem gen ifade hem de proteomik veri kümelerinde bulunan ortak genler seçilmiştir. İki veri kümesindeki ortak genleri seçmek için, protein adları gen sembolleriyle eşleştirilmiştir (Çizelge A-2. 1). Buna göre ortak genlerin ve proteinlerin sayısı 159 olarak bulunmuştur. Sonrasında ilgili veri setlerindeki hastalardan her iki veri setinde ortak olanlar bulunmuştur. Bulunan ortak hasta sayısı 84'tür. Sonuç olarak entegrasyon işleminde kullanılmak üzere 84 ortak hastaya ait 159 gen ve 159 protein içeren veri setleri oluşturulmuştur.

6.2 Analiz ve Entegrasyon

Entegrasyon için hazırlanan gen ifade ve proteomik veri setleri Bölüm 5'te detaylı bir şekilde anlatılmış olan yöntemle hem biyolojik yol ve hem de gen seviyesinde ayrı ayrı analiz edilmiştir. Entegrasyon için hazırlamış olduğumuz veri setlerinin analizinde çizelge 5.4'te verilen parametre ve fonksiyonlar kullanılarak, Bölüm 5.3'te anlatıldığı üzere her bir ilişki tahminci için en anlamlı sonucu veren parametreler belirlenmiştir. Sonrasında bu parametreler kullanılarak her bir ilişki tahmincinin biyolojik yol ve gen seviyesinde en yüksek kesinlik değerine sahip olduğu değerler ile gen ve proteomik alt ağları oluşturulmuştur. Oluşturulan bu ağların entegrasyonu için öncelikle etkileşim matrisleri bulunmuştur. *EM,* etkileşim matrisini; *G,* gen alt ağındaki genleri belirtmek üzere, G_i ve G_j genleri arasında bir ilişki olduğu geliştirilen yöntemle bulunmuşsa $EM_G(ij)$ = 1, eğer bir ilişki bulunmamışsa $EM_G(ij)$ = 0 olacak şekilde gen etkileşim matrisi oluşturulmuştur. Proteomik alt ağları için de aynı yöntemle; *P*, proteomik alt ağındaki proteinleri belirtmek üzere P_i ve P_j proteinleri arasında bir ilişki geliştirilen yöntemle bulunmuşsa $EM_P(ij)$ = 1, bulunmamışsa $EM_P(ij)$ = 0 olacak şekilde protein etkileşim matrisi oluşturulmuştur. Oluşturulmuş olan EM_G ve EM_P matrisleri kullanılarak da Bağıntı 6.1 ile entegrasyon matrisi (E) oluşturulmuştur.

$$E = \frac{EM_G + EM_P}{2} \tag{6.1}$$

Son olarak, oluşturulan *E* matrisi kullanılarak, E_{ij} >= 0,5 eşit değerini sağlayan ilişki değerlerine sahip gen/protein çiftleri kullanılarak her bir ilişki tahminci için entegrasyonu sağlanmış olan ağ oluşturulmuştur. Yapılan entegrasyon işleminin daha iyi anlaşılması için ayrıca Şekil 6.1 verilmiştir.

Şekil 6. 1 Gen ve protein ağlarının entegrasyonu

Oluşturulan bu ağların biyolojik olarak anlamlı olup olmadığını bulmak için EK-B'de verilen web uygulaması kullanılmıştır. Her bir ilişki tahmincinin elde ettiği kesinlik değeri

ve Bölüm 4.1'de detayları verilmiş olan FKT ile hesaplanmış p-değeri Çizelge 6.1'de verilmiştir.

Çizelge 6. 1 İlişki tahmincilerin biyolojik ağların entegrasyonundaki başarım değerleri

	Kesinlik	FKT p-değeri
Adjacency	0,14	8,27E-001
Kendall Tau KKD	0,18	3,23E-001
Pearson KKD	0,15	5,26E-001
Spearman KKD	0,17	3,19E-001
BS	0,21	8,92E-002
Ampirik	0,18	1,75E-001
ÇYT	0,23	1,38E-003
MM	0,2	4,79E-002
SG	0,15	4,15E-001
Shrink	0,25	7,70E-005

6.3 Bölüm Sonuçları

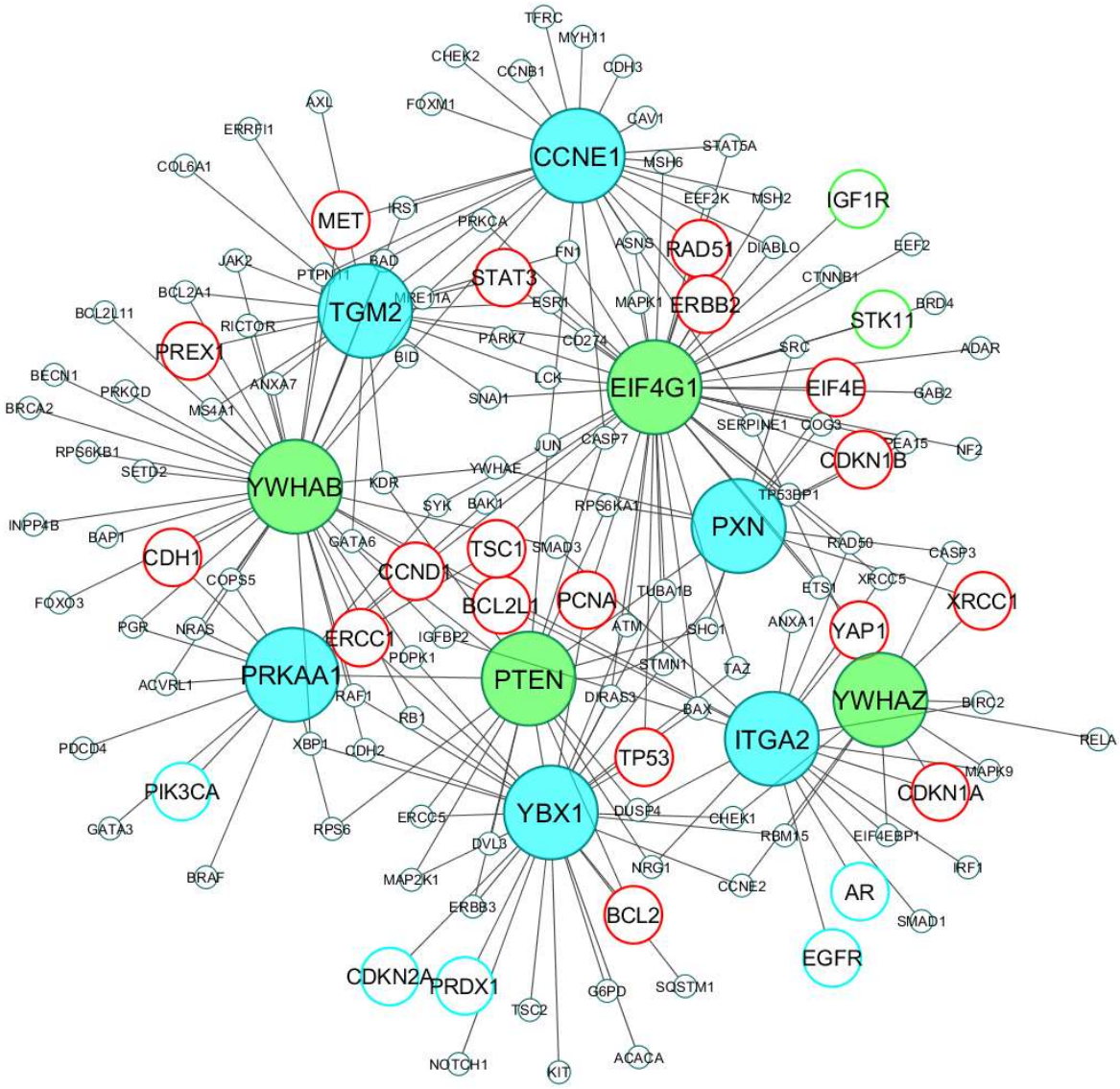
Bu bölümde Bölüm 5'te ayrıntısı verilen ve bu tez çalışmasında önerilmiş olan metodoloji kullanılarak ilişki tahmincilerinin biyolojik ağların entegrasyonu üzerindeki etkisi incelenmiştir. Bu işlem için 84 akciğer kanseri hastasından elde edilen 159 gen ve proteine ait veriler analiz edilmiştir. İlişki tahmincilerinin karşılaştırılması için her bir ilişki tahmincinin geliştirilen metodolojiye göre en iyi sonuçları elde ettiği gen ve protein alt ağları kullanılmıştır.

Sonuç olarak, Çizelge 6.1 incelendiğinde, entegrasyon işlemi sonucu oluşturulan ağlar her bir ilişki tahminci için kesinlik değerlerine göre değerlendirildiğinde, Shrink ilişki tahmincisinin en iyi skora eriştiği görülmüştür. Bu sonucun FKT p-değerine göre istatistiksel olarak da anlamlı olduğu değerlendirilmiştir. Kesinlik skorlarına göre Shrink ilişki tahmincisini sırasıyla ÇYT, BS ve MM ilişki tahmincileri takip etmiştir. Korelasyon tabanlı ilişki tahmincileri olan Kendall Tau, Spearman ve Pearson KKD yöntemleri kısmen daha düşük kesinlik skorlarına sahip olmuştur. Bu sonuçlar Bölüm 5'te elde ettiğimiz sonuçlarla örtüşmektedir ve KB tabanlı ilişki tahmincilerinin kanser gibi karmaşık yapıya sahip hastalıklarla ilişkili olabilecek yapıların tespitinde korelasyon tabanlı yöntemlere göre daha başarılı olduğunu göstermiştir.

Son olarak, entegrasyon sonucu en yüksek kesinlik skorunu sağlayan Shrink ilişki tahmincisi tarafından tanımlanan modüllerin merkez genlerinin alt ağları Cytoscape [79] ile oluşturularak Şekil 6.2'de gösterilmiştir.

Bu alanda çalışan diğer araştırmacıların da değerlendirmesi adına akciğer kanseri için DisGeNET'te kayıtlı olan ve deneysel olarak hastalıkla ilişkisi olduğu onaylanan genler, Şekil 6.2'de kırmızı, yeşil ve mavi renkli çerçevelerle gösterilmiştir. Yeşil çerçeveye sahip genler yalnızca proteomik verilerin analizinden, mavi çerçeveye sahip genler sadece gen ifade verilerinin analizinden ve kırmızı çerçeveye sahip genler proteomik ve gen ifade verilerinin analizinden elde edilmiş ortak genleri göstermektedir. Ayrıca, yeşil renkli merkez genler proteomik verilerinden, mavi renkli merkez genler ise gen ifade verilerinden elde edilmiştir. Şekil 6.2'deki gri renkli düğümler için DisGeNET'te herhangi bir bilgiye rastlanmamıştır.

Bölüm 5'te verilen yöntem kullanılarak yapılan analizde Shrink ilişki tahmincisi ile yapılan entegrasyon işlemi sonucunda EIF4G1 (EIF4G), PTEN, YWHAZ, YWHAB, CCNE1, TGM2, PXN, PRKAA1, YBX1 ve ITGA2 hub genlerinin, akciğer kanseri üzerinde önemli etkileri olabileceği bulunmuştur. DisGeNET'teki veriler bu genler arasından PTEN geninin beş çalışmada akciğer kanseri ile ilişkili olduğunu göstermiştir. Bununla birlikte ERBB2, IGF1R, PCNA, RAD51, STAT3, STK11, TP53, TSC1, YAP1, CCND1, CDKN1, BBCL2L1, CDH1, ERCC1, MET, PREX1, BCL2, CDKN1A, XRCC1, PTEN, PRDX1, CDKN2A, AR, EGFR, PIK3CA ve EIF4E genlerinin çoklu çalışmalarda hastalık ilişkili genler olarak doğrulandığını DisGeNET'teki veriler doğrulamıştır.



Şekil 6. 2 Akciğer kanseri proteomik ve gen ifadesi veri setleri için entegrasyon sonucu en yüksek kesinlik değerine ulaşan Shrink tahmin yöntemiyle elde edilen hastalıkla ilişkili alt ağlardaki merkez genler ve komşuları

BÖLÜM 7

SONUÇ VE ÖNERİLER

Bu çalışmada kanser gibi karmaşık biyolojik ilişkilere sahip hastalıkların analizinde daha doğru sonuçlara ulaşılmasına izin veren proteomik kanser verileri kullanılmıştır [31]. Verilerin analizinde ilişki tahmincilerin etkisi literatürde sıklıkla kullanılan beş GAÇ algoritmasıyla incelenmiştir. Bunun için literatürde birçok çalışmada kullanılmış olan dokuz farklı ilişki tahmincisi seçilmiştir.

Bölüm 4'te anlatıldığı üzere korelasyon tabanlı ilişki tahmincilerin proteomik verilerin analizi üzerindeki başarımı dört farklı GAÇ algoritması ile incelenmiştir. Yöntemlerin başarımını ölçmek altın standart olarak, PC [59] tarafından oluşturulan ve içinde 25 farklı biyolojik veritabanındaki bilgilerin entegrasyonu sonucu elde edilmiş BioPAX [60] formatındaki dosyada yer alan protein-protein ilişki verileri kullanılmıştır. BioPAX dosyasında yer alan 19.000'e yakın proteine ait bir buçuk milyondan fazla protein-protein ilişkisi Babur vd. [61] tarafından geliştirilen Java uygulaması kullanılarak tespit edilmiştir. Elde edilen bu protein-protein ilişkileri verilerinin diğer araştırmacıların da kullanımına sunulması için Ek A-4'te anlatılan bir web uygulaması da geliştirilmiştir. Elde ettiğimiz protein-protein ilişki verilerini kullanarak ARACNE, MRNET, CLR ve C3NET GAÇ algoritmalarının korelasyon tabanlı ilişki tahmincilerine göre elde ettikleri başarımları Bölüm 4.1'de detaylarını verdiğimiz analiz işlemi ile değerlendirilmiştir. Elde ettiğimiz sonuçlar kesinlik değerlerine göre incelendiğinde korelasyon tabanlı ilişki tahmincilerden MRNET, CLR ve C3NET GAÇ algoritmalarının başarımında büyük bir etkisi olmadığı görülmüştür. ARACNE algoritmasında ise BRCA veri seti için kesinlik değerlerine göre başarım sırası 0,18128; 0,17919; 0,15502 değerleriyle sırasıyla *Spearman KKD > Kendall Tau KKD > Pearson KKD* şeklinde olmuştur ve diğer veri setlerinde de başarım

sırası değişmekle birlikte benzer sonuçlar alınmıştır. Bu kesinlik değerlerinin istatiksel olarak rastgelelikten uzak olup olmadığını kontrol etmek için uyguladığımız FKT testi de 0,05 eşik değerine göre sonuçlarımızın anlamlı olduğunu göstermiştir. Bununla birlikte, kanser türleri için kesinlik değerlerine göre algoritmaların başarımları incelendiğinde Pearson KKD ile yapılan analizde 12 kanser türünde C3NET, 4 kanser türünde ARACNE; Spearman KKD ile yapılan analizde 14 kanser türünde C3NET, 2 kanser türünde ARACNE; Kendall Tau KKD ile yapılan analizde 15 kanser türünde C3NET, 1 kanser türünde ARACNE en yüksek skorları elde etmiştir. Ayrıca, kesinlik değerine göre GAÇ algoritmalarının genel başarım sıralaması *C3NET > ARACNE > MRNET ≈ CLR* şeklinde olmuştur. C3NET ve ARACNE algoritmalarının kesinlik skorlarının diğer algoritmalara göre daha yüksek olmasının nedeni Bölüm 4.2.1 ve 4.2.2'de açıklandığı üzere ilişki skorları üzerinde yapmış oldukları eleme işlemi sonucu doğruluğu daha yüksek ağlar elde etmeleridir. Buna karşın, CLR ve MRNET algoritmalarının ilişki tahmin sayıları ile bu tahminlerden gerçek pozitif olanların sayıları birlerine yakındır. Bu iki algoritmanın yüksek sayıda ilişki tahmini yapmasına karşın bu tahminlerdeki gerçek pozitif sayıları yeteri kadar yüksek olmadığı için kesinlik skorları düşük olmuştur. Bu değerlendirme sonucunda hastalıkla ilgili etkin genlerin/proteinlerin tespitinde C3NET algoritmasının diğer GAÇ yöntemlerine göre daha yüksek kesinlik değerlerine sahip olduğu görülmüştür. Alınan sonuçların istatistiksel olarak anlamlılığı FKT ile değerlendirilmiş ve 0,05 eşik değerine göre anlamlı bulunmuştur. Ayrıca, C3NET ile tespit ettiğimiz ve Bölüm 4'te verdiğimiz protein-protein ilişkilerinin biyolojik olarak ta anlamlı olduğu yapılan literatür taraması sonucunda tespit edilmiştir ve ilgili literatür çalışmaları Çizelge 4.14'te belirtilmiştir. Algoritmalar analiz sürelerine göre değerlendirildiğinde ARACNE, CLR ve MRNET bir birine yakın sürelerde analiz işlemlerini gerçekleştirmişlerdir. C3NET algoritması ise yapmış olduğu eleme işlemelerinin çok zaman almasından dolayı diğer algoritmalara göre 10 kat daha yavaştır.

Sonrasında KB tabanlı ilişki tahmincilerin proteomik verilerin analizindeki başarımını incelemek için Bölüm 5'te açıklaması yapılan yöntem geliştirilmiştir. Bu bölümde incelenen beş kanser türü, Amerikan Kanser Derneği tarafından yayınlanan sonuçlara göre en sık görülen kanser türleri arasındadır [92]. Hedef hastalıkla ilgili yüksek ilişkili gen/protein kümelerini ve bu kümelerdeki yüksek ilişki değerlerine sahip merkez genleri bulmak için WGCNA yöntemini kullanılmıştır. WGCNA yönteminde, incelenen veri

setinde yer alan genlerin/proteinlerin ilişkisini tespit etmek için varsayılan olarak kullanılan korelasyon tabanlı *İlişki* yöntemi yerine KB tabanlı ilişki tahmincilerinin entegrasyonunu gerçekleştirilmiştir. Bu sayede doğrusal olmayan ilişkilerin daha doğru bir şekilde tespit edilmesi hedeflenmiş ve bu hedefin doğruluğunu tespit etmek için Bölüm 5'te açıklandığı üzere biyolojik-yol ve gen seviyesinde ilişki tahmincilerin başarımı karşılaştırılmıştır. İlişki tahmincilerin başarımını belirlemek için kullandığımız ve hastalıkla ilişkili olduğu literatürde yapılan önceki çalışmalarca tespit edilmiş genler DisGeNET'ten elde edilmiştir. Ardından bu genlerin yer aldığı biyolojik-yolları belirlemek için BioCarta, KEGG ve Reactome veritabanlarından bu genleri içerme oranı en yüksek yüz biyolojik-yol (biological pathway) MSigDB çevrimiçi aracı kullanılarak elde edilmiştir. DisGeNET ve MSigDB'den elde edilen bu veriler altın standart olarak kabul edilmiş ve ilişki tahmincilerin başarımlarının hesaplanmasında kullanılmıştır. Sonuç olarak beş kanserin türünün dördünde, KB tabanlı yöntemler biyolojik-yol ve gen seviyesindeki analiz sonuçlarında daha başarılı kesinlik skorlarına erişmiştir ve bunlar BRCA için Shrink, GBM için SG, KIRC için Shrink, LUSC için BS'dir. SKCM kanser türünde ise kesinlik skorlarına göre ilişki tahmincilerin başarımı incelendiğinde, MM daha yüksek skora sahip olmakla birlikte Kendall Tau KKD ile çok yakın başarım skorlarına erişmiştir. Sonrasında, her bir kanser türü için ayrı ayrı olmak üzere en başarılı kabul edilen bu kestirimciler kullanılarak elde edilen modüllerin ve hub genlerin grafiksel gösterimi biyologların ve bu alanda çalışan araştırmacıların incelemesi adına Şekil 5.16 – 5.20'de sunulmuştur. Şekil 5.16 – 5.20'de yer alan ve önerdiğimiz yöntemce önemli oldukları varsayılan merkez genler için literatür taraması yapıldığında, bu genlerden bazılarının (PDK1, SMAD4, RAD50, FOXM1, CDKN1A, ETS1 [82], [83], [84], [85], [86], [87], [88], [89]) DisGeNET'te yer almamasına rağmen, literatürde ilgili hastalıklarla ilişkili olduklarının tespit edildiği görülmüştür. Bu nedenle, bizim önerdiğimiz yöntem, daha önce çalışılmış olan bu genleri yakalayabildiğinden, önceki çalışmalarda gözden kaçan hastalıklarla ilişkili gen-gen etkileşimlerinin daha kapsamlı bir listesini yakalayabileceği öngörülmüştür.

Bölüm 5'te yer alan bu çalışmamızın en önemli katkısı, WGCNA yöntemi ile entegre edildiğinde hastalıkla ilişkili modüllerinin belirlenmesinde önemli bir iyileşme sağlayabilen, biyolojik ağ oluşturma metodolojilerinde farklı ilişkilendirme tahmin edicilerinin kullanılmasıdır. Ayrıca, her bir tahmin edicinin biyolojik-yol seviyesi ve gen

seviyesi analizindeki benzer kesinlik puanları da çalışmamızın tutarlılığını göstermektedir.

Son olarak, Bölüm 6'da gerçekleştirilen entegrasyon işlemi sonucunda Bölüm 5'te verilen sonuçlarla örtüşen neticeler elde edilmiştir. Yani, KB tabanlı ilişki tahmincileri korelasyon tabanlı yöntemlere karşı entegrasyon işlemi sonucunda da daha yüksek kesinlik skorlarına sahip biyolojik ağların elde edilmesine olanak sağlamıştır.

Gelecekteki çalışmalarımızda farklı proteomik verilerin analizinde farklı ilişki tahmincilerinin ve farklı ağ çıkarım yöntemlerinin incelenmesi yapılacak olup, doğruluğu daha yüksek biyolojik ağların tahmin edilebilmesi için yeni yöntemler geliştirilmeye çalışılacaktır.

KAYNAKLAR

[1] Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C. Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. ve Vidal, M., (2005). "Towards a proteome-scale map of the human protein–protein interaction network", Nature, 437: 1173.

[2] Schadt, E.E., (2009). "Molecular networks as sensors and drivers of common human diseases", Nature, 461: 218.

[3] Olsen, C., Meyer, P.E. ve Bontempi, G., (2008). "On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information", EURASIP Journal on Bioinformatics and Systems Biology, 2009: 308959.

[4] Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B. ve Marchal, K., (2006). "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms", BMC Bioinformatics, 7: 43.

[5] Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. ve Califano, A., (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context", BMC Bioinformatics, 3: 1-7.

[6] Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., ve Gardner, T.S., (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles", PLoS Biology, 5: 8-15.

[7] Peng, H., Long, F. ve Ding, C., (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, 27: 1226-1238.

[8] Meyer, P.E., Lafitte, F. ve Bontempi, G., (2008). "minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information", BMC Bioinformatics, 9: 461.

[9] de Matos Simoes, R. ve Emmert-Streib, F., (2011). "Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks", PLoS One, 6: e29279.

[10] Altay, G. ve Emmert-Streib, F., (2010). "Inferring the conservative causal core of gene regulatory networks", BMC Systems Biology, 4: 132.

[11] Daub, C.O., Steuer, R., Selbig, J. ve Kloska, S., (2004). "Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data", BMC Bioinformatics, 5: 118.

[12] Kurt, Z., Aydin, N. ve Altay, G., (2014). "A comprehensive comparison of association estimators for gene network inference algorithms", Bioinformatics, 30: 2142-2149.

[13] Butte, A.J. ve Kohane, I.S., (1999). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements", Pacific Symposium on Biocomputing 2000, 4-9 Ocak 2000, Hawaii.

[14] Ish-Horowicz, J. ve Reid, J., (2017). "Mutual information estimation for transcriptional regulatory network inference", BioRxiv, 7: 132647.

[15] Schaffter, T., Marbach, D. ve Floreano, D., (2011). "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods", Bioinformatics, 27: 2263-2270.

[16] Kong, W., Zhang, J., Mou, X. ve Yang, Y., (2014). "Integrating gene expression and protein interaction data for signaling pathway prediction of Alzheimer's disease", Computational and Mathematical Methods in Medicine, 2014: 340758.

[17] Shi, X., Wang, X., Shajahan, A., Hilakivi-Clarke, L., Clarke, R. ve Xuan, J., (2015). "BMRF-MI: integrative identification of protein interaction network by modeling the gene dependency", BMC Genomics, 16: 10.

[18] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G. ve Milanesi, L., (2016). "Methods for the integration of multi-omics data: mathematical aspects", BMC Bioinformatics, 17: 15.

[19] Zhao, X.M., Wang, R.S., Chen, L. ve Aihara, K., (2008). "Uncovering signal transduction networks from high-throughput data by integer linear programming", Nucleic Acids Research, 36: e48.

[20] Şenbabaoğlu, Y., Sümer, S.O., Sánchez-Vega, F., Bemis, D., Ciriello, G., Schultz, N. ve Sander, C., (2016). "A multi-method approach for proteomic network inference in 11 human cancers", PLoS Computational Biology, 12: e1004765.

[21] Madhamshettiwar, P.B., Maetschke, S.R., Davis, M.J., Reverter, A. ve Ragan, M.A., (2012). "Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets", Genome Medicine, 4: 41.

[22] Langfelder, P. ve Horvath, S., (2008). "WGCNA: an R package for weighted correlation network analysis", BMC Bioinformatics, 9: 559.

[23] Horvath, S., Zhang, B., Carlson, M., Lu, K., Zhu, S., Felciano, R., Laurance, M., Zhao, W., Qi, S. ve Chen, Z., (2006). "Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target", Proceedings of the National Academy of Sciences, 103: 17402-17407.

[24] Carlson, M.R., Zhang, B., Fang, Z., Mischel, P.S., Horvath, S. ve Nelson, S.F., (2006). "Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks", BMC Genomics, 7: 40.

[25] Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B. ve Gunnarsdottir, S., (2008). "Genetics of gene expression and its effect on disease", Nature, 452: 423.

[26] Fuller, T.F., Ghazalpour, A., Aten, J.E., Drake, T.A., Lusis, A.J. ve Horvath, S., (2007). "Weighted gene coexpression network analysis strategies applied to mouse weight", Mammalian Genome, 18: 463-472.

[27] Keller, M.P., Choi, Y., Wang, P., Davis, D.B., Rabaglia, M.E., Oler, A.T., Stapleton, D.S., Argmann, C., Schueler, K.L. ve Edwards, S., (2008). "A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility", Genome Research, 18: 706-716.

[28] Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P.L., Liu, W., Yang, J.Y., Broom, B.M., Verhaak, R.G. ve Kane, D.W., (2013). "TCPA: a resource for cancer functional proteomics data", Nature Methods, 10: 1046.

[29] Ruyssinck, J., Geurts, P., Dhaene, T., Demeester, P. ve Saeys, Y., (2014). "Nimefi: gene regulatory network inference using multiple ensemble feature importance algorithms", PLoS One, 9: e92709.

[30] Anderson, L., ve Seilhamer, J., (1997). "A comparison of selected mRNA and protein abundances in human liver", ELECTROPHORESIS, 18: 533-537.

[31] Arthur, J.M., (2003). "Proteomics", Current Opinion in Nephrology and Hypertension, 12: 423-430.

[32] Moon, Y.I., Rajagopalan, B. ve Lall, U., (1995). "Estimation of mutual information using kernel density estimators", Physical Review, 52: 2318.

[33] Paninski, L., (2003). "Estimation of entropy and mutual information", Neural computation, 15: 1191-1253.

[34] Miller, G.A., (1955). "Note on the bias of information estimates", Information theory in psychology: Problems and Methods, 2: 100.

[35] Stuart, A., (1956). "Rank Correlation Methods. By M. G. Kendall", British Journal of Statistical Psychology, 9: 68-68.

[36] Hausser, J. ve Strimmer, K., (2009). "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks", Journal of Machine Learning Research, 10: 1469-1484.

[37] Schürmann, T. ve Grassberger, P., (1996). "Entropy estimation of symbol sequences", Chaos: An Interdisciplinary Journal of Nonlinear Science, 6: 414-427.

[38] Cakır, T., Hendriks, M.M., Westerhuis, J.A. ve Smilde, A.K., (2009). "Metabolic network discovery through reverse engineering of metabolome data", Metabolomics, 5: 318-329.

[39] De La Fuente, A., Bing, N., Hoeschele, I. ve Mendes, P., (2004). "Discovery of meaningful associations in genomic data using partial correlation coefficients", Bioinformatics, 20: 3565-3574.

[40] Soper, H., Young, A., Cave, B., Lee, A. ve Pearson, K., (1917). "On the distribution of the correlation coefficient in small samples.", Biometrika, 11: 328-413.

[41] Casadei, D., (2012). "Estimating the selection efficiency", Journal of Instrumentation, 7: 8021.

[42] Qiu, P. ve Zhang, L., (2012). "Identification of markers associated with global changes in DNA methylation regulation in cancers", BMC Bioinformatics, 13: 7.

[43] Veitch, J., Mandel, I., Aylott, B., Farr, B., Raymond, V., Rodriguez, C., van der Sluys, M., Kalogera, V. ve Vecchio, A., (2012). "Estimating parameters of coalescing compact binaries with proposed advanced detector networks", Physical Review, 85: 104045.

[44] Ayadi, W., Elloumi, M. ve Hao, J.-K., (2012). "Pattern-driven neighborhood search for biclustering of microarray data", BioMed Central, 13: 11.

[45] Mahanta, P., Ahmed, H.A., Bhattacharyya, D.K. ve Kalita, J.K., (2012). "An effective method for network module extraction from microarray data", BMC Bioinformatics, 13: 4.

[46] Wu, C., Zhu, J. ve Zhang, X., (2012). "Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes", BMC Bioinformatics, 13: 182.

[47] Ioannidis, J.P., Trikalinos, T.A., Ntzani, E.E. ve Contopoulos-Ioannidis, D.G., (2003). "Genetic associations in large versus small studies: an empirical assessment", The Lancet, 361: 567-571.

[48] Lapata, M., (2006). "Automatic evaluation of information ordering: Kendall's tau", Computational Linguistics, 32: 471-484.

[49] Zotenko, E., Mestre, J., O'Leary, D.P. ve Przytycka, T.M., (2008). "Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality", PLoS Computational Biology, 4: e1000140.

[50] Meyer, P.E., Kontos, K. ve Bontempi, G., (2007). "Biological network inference using redundancy analysis", Bioinformatics Research and Development, 12: 16-27.

[51] Vu, V.Q., Yu, B. ve Kass, R.E., (2007). "Coverage-adjusted entropy estimation", Statistics in Medicine, 26: 4039-4060.

[52] Murakami, Y. ve Mizuguchi, K., (2010). "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites", Bioinformatics, 26: 1841-1848.

[53] Chang, D.T.H., Wang, C.C. ve Chen, J.-W., (2008). "Using a kernel density estimation based classifier to predict species-specific microRNA precursors", BMC Bioinformatics, 9: 2.

[54] Taylor, C.C., Mardia, K.V., Di Marzio, M. ve Panzera, A., (2012). "Validating protein structure using kernel density estimates", Journal of Applied Statistics, 39: 2379-2388.

[55] Yip, A.M. ve Horvath, S., (2007). "Gene network interconnectedness and the generalized topological overlap measure", BMC Bioinformatics, 8: 22.

[56] Li, A. ve Horvath, S., (2006). "Network neighborhood analysis with the multi-node topological overlap measure", Bioinformatics, 23: 222-231.

[57] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. ve Barabási, A.-L., (2002). "Hierarchical organization of modularity in metabolic networks", Science, 297: 1551-1555.

[58] Akbani, R., Becker, K.F., Carragher, N., Goldstein, T., de Koning, L., Korf, U., Liotta, L., Mills, G.B., Nishizuka, S.S. ve Pawlak, M., (2014). "Realizing the Promise of Reverse Phase Protein Arrays for Clinical, Translational, and Basic Research: A Workshop Report The RPPA (Reverse Phase Protein Array) Society", Molecular & Cellular Proteomics, 13: 1625-1643.

[59] Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D. ve Sander, C., (2010). "Pathway Commons, a web resource for biological pathway data", Nucleic Acids Research, 39: 685-690.

[60] Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'eustachio, P., Schaefer, C. ve Luciano, J., (2010). "The BioPAX community standard for pathway data sharing", Nature Biotechnology, 28: 935.

[61] Babur, Ö., Aksoy, B.A., Rodchenkov, I., Sümer, S.O., Sander, C. ve Demir, E., (2013). "Pattern search in BioPAX models", Bioinformatics, 30: 139-140.

[62] Fisher R.A., (1992). Statistical Methods for Research Workers, Springer, New York.

[63] Fury, W., Batliwalla, F., Gregersen, P.K. ve Li, W., (2006). "Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion", International Conference of the IEEE Engineering in Medicine and Biology Society, 1-3 Eylül 2006, New York.

[64] Nikuševa-Martić, T., Beroš, V., Pećina-Šlaus, N., Pećina, H.I. ve Bulić-Jakuš, F., (2007). "Genetic changes of CDH1, APC, and CTNNB1 found in human brain tumors", Pathology-Research and Practice, 203: 779-787.

[65] Kim, S.K., Roh, Y.G., Park, K., Kang, T.H., Kim, W.J., Lee, J.S., Leem, S.H. ve Chu, I.-S., (2014). "Expression Signature Defined by FOXM1–CCNB1 Activation Predicts Disease Recurrence in Non–Muscle-Invasive Bladder Cancer", Clinical Cancer Research, 20: 3233-3243.

[66] Rambau, P.F., Duggan, M.A., Ghatage, P., Warfa, K., Steed, H., Perrier, R., Kelemen, L.E. ve Köbel, M., (2016). "Significant frequency of MSH2/MSH6

abnormality in ovarian endometrioid carcinoma supports histotype-specific Lynch syndrome screening in ovarian carcinomas", Histopathology, 69: 288-297.

[67] Ramsoekh, D., Wagner, A., van Leerdam, M.E., Dooijes, D., Tops, C.M., Steyerberg, E.W. ve Kuipers, E.J., (2009). "Cancer risk in MLH1, MSH2 and MSH6 mutation carriers; different risk profiles may influence clinical management", Hereditary cancer in clinical practice, 7: 17.

[68] Zheng, Z., Yang, J., Zhao, D., Gao, D., Yan, X., Yao, Z., Liu, Z. ve Ma, Z., (2013). "Downregulated adaptor protein p66Shc mitigates autophagy process by low nutrient and enhances apoptotic resistance in human lung adenocarcinoma A549 cells", The FEBS Journal, 280: 4522-4530.

[69] Eisen, M.B., Spellman, P.T., Brown, P.O. ve Botstein, D., (1998). "Cluster analysis and display of genome-wide expression patterns", Proceedings of the National Academy of Sciences, 95: 14863-14868.

[70] The University of Texas MD Anderson Cancer Center, http://www.tcpaportal.org/tcpa/, 30 Ağustos 2017.

[71] Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F. ve Furlong, L.I., (2016). "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants", Nucleic Acids Research, 45: 943.

[72] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R. ve Lander, E.S., (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", Proceedings of the National Academy of Sciences, 102: 15545-15550.

[73] McDonald, J.H., (2014). Handbook of Biological Statistics, Third Edition, Sparky House Publishing, Maryland.

[74] Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. ve Tress, M.L., (2014). "Multiple evidence strands suggest that there may be as few as 19000 human protein-coding genes", Human Molecular Genetics, 23: 5866-5878.

[75] Altay, G., Kurt, Z., Altay, N. ve Aydin, N., (2017). "DepEst: an R package of important dependency estimators for gene network inference algorithms", BioRxiv, 3: 102871.

[76] Wee, D., Integrative Biomedical Informatics Group GRIB/IMIM/UPF, http://www.disgenet.org/web/DisGeNET/menu/, 19 Mayıs 2017.

[77] Sig, G., GSEA and MSigDB Team, http://software.broadinstitute.org/gsea/msigdb/index.jsp, 19 Mayıs 2017.

[78] Dong, J. ve Horvath, S., (2007). "Understanding network concepts in modules", BMC Systems Biology, 1: 24.

[79] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. ve Ideker, T., (2003). "Cytoscape: a software environment for

integrated models of biomolecular interaction networks", Genome Research, 13: 2498-2504.

[80] Erdoğan, C., Kurt, Z. ve Diri, B., (2017). "Estimation of the proteomic cancer co-expression sub networks by using association estimators", PLoS One, 12: e0188016.

[81] Song, L., Langfelder, P. ve Horvath, S., (2012). "Comparison of co-expression measures: mutual information, correlation, and model based indices", BMC Bioinformatics, 13: 328.

[82] Du, J., Yang, M., Chen, S., Li, D., Chang, Z. ve Dong, Z., (2016). "PDK1 promotes tumor growth and metastasis in a spontaneous breast cancer model", Oncogene, 35: 3314.

[83] Dupuy, F., Tabariès, S., Andrzejewski, S., Dong, Z., Blagih, J., Annis, M.G., Omeroglu, A., Gao, D., Leung, S. ve Amir, E., (2015). "PDK1-dependent metabolic reprogramming dictates metastatic potential in breast cancer", Cell Metabolism, 22: 577-589.

[84] Liu, N., Yu, C., Shi, Y., Jiang, J. ve Liu, Y., (2015). "SMAD4 expression in breast ductal carcinoma correlates with prognosis", Oncology Letters, 10: 1709-1715.

[85] Mishima, K., Mishima-Kaneko, M., Saya, H., Ishimaru, N., Yamada, K., Fukada, J., Nishikawa, R. ve Kawata, T., (2014). "Rt-21mre11-rad50-nbs1 Complex Inhibitor Mirin Enhances Radiosensitivity In Human Glioblastoma Cells", Neuro-oncology, 16: 192.

[86] Zhang, J., Zhang, J., Cui, X., Yang, Y., Li, M., Qu, J., Li, J. ve Wang, J., (2015). "FoxM1: A novel tumor biomarker of lung cancer", International Journal of Clinical and Experimental Medicine, 8: 3136.

[87] Sun, Q., Dong, M., Chen, Y., Zhang, J., Qiao, J. ve Guo, X., (2016). "Prognostic significance of FoxM1 expression in non-small cell lung cancer", Journal of Thoracic Disease, 8: 1269.

[88] Fukazawa, T., Guo, M., Ishida, N., Yamatsuji, T., Takaoka, M., Yokota, E., Haisa, M., Miyake, N., Ikeda, T. ve Okui, T., (2016). "SOX2 suppresses CDKN1A to sustain growth of lung squamous cell carcinoma", Scientific Reports, 6: 20113.

[89] Keehn, C.A., Smoller, B.R. ve Morgan, M.B., (2004). "Ets-1 immunohistochemical expression in non-melanoma skin carcinoma", Journal of Cutaneous Pathology, 31: 8-13.

[90] University of California, UCSC Xena, https://xenabrowser.net/datapages/, 30 Ağutos 2017.

[91] Goldman, M., Craft, B., Zhu, J. ve Haussler, D., (2017). "The UCSC Xena system for cancer genomics data visualization and interpretation", AACR Annual Meeting, 1-5 Nisan 2017, Washington.

[92] Siegel, R.L., Miller, K.D. ve Jemal, A., (2018). "Cancer statistics, 2018", CA: A Cancer Journal for Clinicians, 68: 7-30.

[93] Erli, F., Fen Sitesi, https://www.fenehli.com/, 18 Mart 2018.

[94] Watson, J.D. ve Crick, F.H., (1953). "The structure of DNA", Cold Spring Harbor Symposia on Quantitative Biology, 18:123-31.

[95] Yıldırım, A., Bardakçı, F., Karataş, M. ve Tanyolaç, B., (2010). "Moleküler Biyoloji", Nobel Yayın Dağıtım, 613: 2-13.

[96] Başaran, E., Aras, S. ve Cansaran-duman, D., (2010). "Genomik, proteomik, metabolomik kavramlarına genel bakış ve uygulama alanları", Türk Hijyen ve Deneysel Biyoloji Dergisi, 67: 85-96.

[97] Kuska, B., (1998). "Beer, Bethesda, and biology: how genomics came into being", Journal of the National Cancer Institute, 902: 93.

[98] Bal, S.H. ve Budak, F., (2013). "Genomik, Proteomik Kavramlarına Genel Bakış ve Uygulama Alanları", Uludağ Üniversitesi Tıp Fakültesi Dergisi, 39: 65-69.

[99] Del Boccio, P. ve Urbani, A., (2005). "Homo sapiens proteomics: clinical perspectives", Annali dell'Istituto Superiore Di Sanita, 41: 479-482.

[100] Kurban, S., (2010). "Proteomik", Yeni Tıp Dergisi, 27: 70.

[101] Collins, D., Joe, J., Hiroyuki, N., R.NET, http://jmp75.github.io/rdotnet/index.html/, 18 Mart 2016.

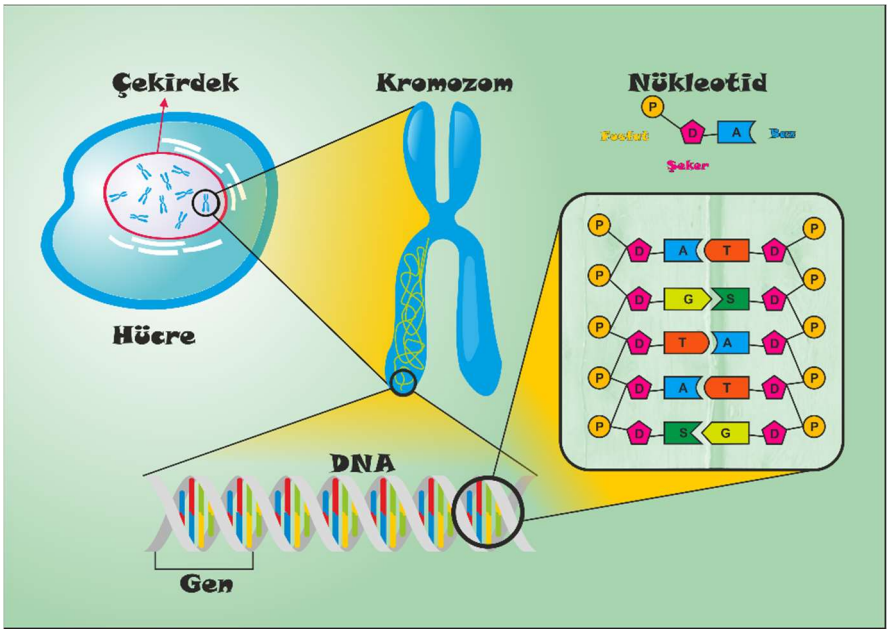
EK-A

BİYOLOJİK BİLGİ

Bu bölümde tezimizde kullandığımız biyolojik terimlerin daha iyi anlaşılması için temel biyolojik ön bilgiler verilmiştir

A-1 Kromozom, DNA, Gen ve Nükleotid

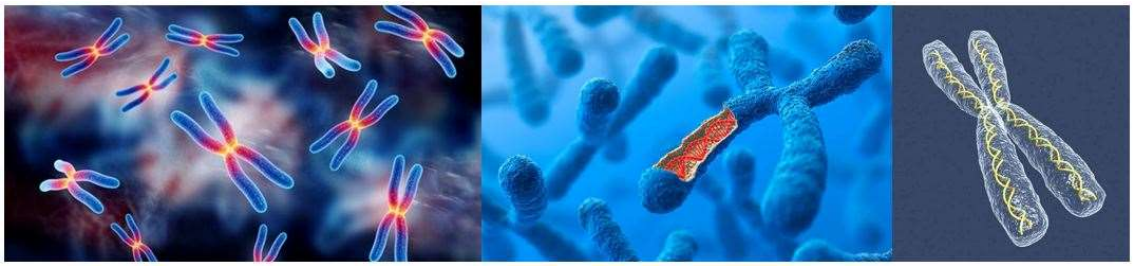
Canlıların en küçük yapı taşı olan hücre, çekirdek adı verilen bir merkezden yönetilmekte ve çekirdek içerisinde de bu yönetimi gerçekleştiren ve ayrıca, canlıya ait kalıtımsal şifreleri taşıyan yapılar yer almaktadır. Bu yapılardan en büyük ve en komplike olanı kromozomdur. DNA ve proteinlerin birleşmesiyle kromozomlar oluşmaktadır. DNA üzerinde de kalıtımsal özelliklere tesir eden genler yer almakta ve genler de nükleotid adı verilen mikro moleküllerden oluşmaktadır (Şekil A-1.1).



Şekil A-1. 1 Hücre, çekirdek, kromozom, DNA, gen ve nükleotid [93]

A-1.1 Kromozom

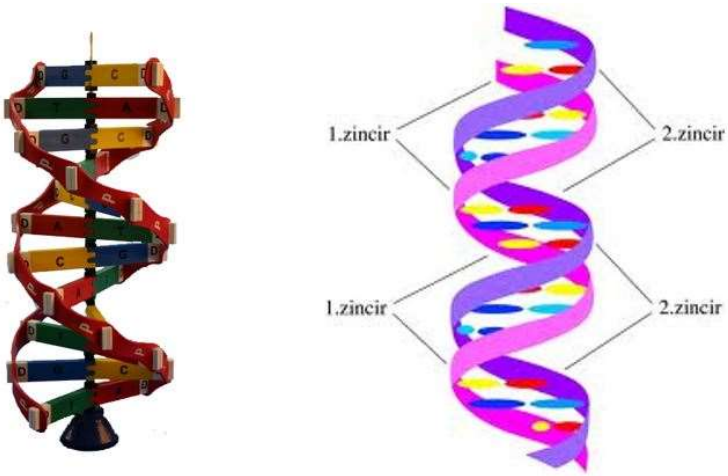
DNA ve proteinlerin birleşimi sonucu oluşan kromozomlar canlılara ait kalıtsal özelliklerin (cinsiyet, göz rengi, saç rengi, vb.) anne ve babadan çocuklarına aktarılmasını sağlayan komplike yapılardır. İnsan vücudunda hücre çekirdeğinde yer alan kromozomlar sıkıca sarılmış DNA paketleri olarak ta bilinir. Kormozomlar çekirdeksiz hücrelerde ise sitoplazma içerisinde dağılmış halde yer alırlar.



Şekil A-1. 2 Kromozom [93]

A-1.2 DNA (Deoksiribo Nükleik Asit)

Canlının özelliklerini nesilden nesile aktarmasını sağlayan DNA [94], kromozomlar üzerinde yer alan ipliksi yapıdaki kalıtım molekülüdür. Hücredeki yönetici molekül olan ve hücrenin genetik bilgisini taşıyan DNA, binlerce genden oluşur. Her bir gen, bir protein molekülünün nasıl inşa edileceğine dair bir reçete olarak hizmet eder. Proteinler hücre fonksiyonları için önemli görevleri yerine getirirler veya yapı taşları olarak görev yaparlar. Genlerden gelen bilgi akışı, protein bileşimini ve dolayısıyla hücrenin işlevlerini belirler.

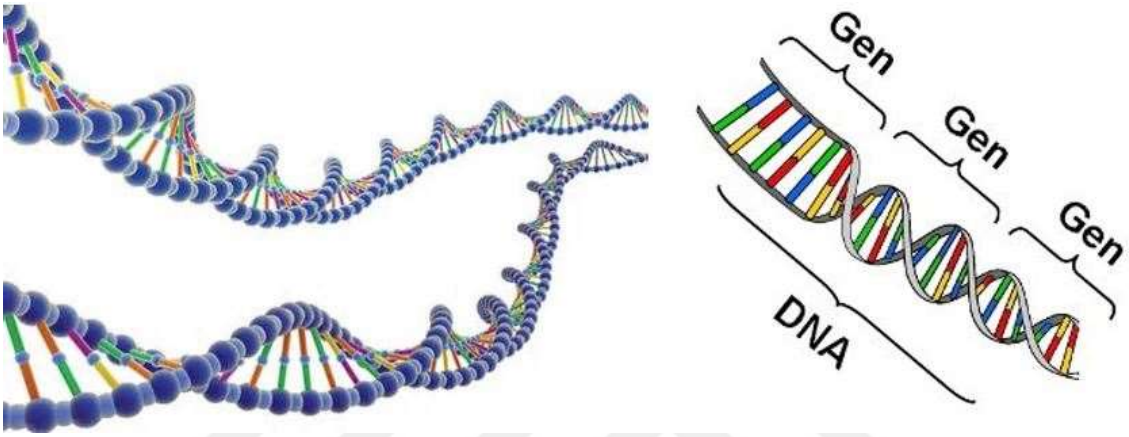


Şekil A-1. 3 DNA'nın yapısı [93]

Çift zincirli sarmal bir yapıda olan DNA zincirlerinin biri anneden diğeri de babadan yavru canlıya aktarılır.

A-1.3 GEN

Proteinlerin kodlanmasını sağlayan genler, genom içindeki küçük DNA bölümleridir (Şekil A-1.4). Organizmanın cinsiyet, göz rengi, saç rengi, ten rengi, kan grubu vb. gibi bireysel özelliklerine ilişkin yönergeleri içerirler.



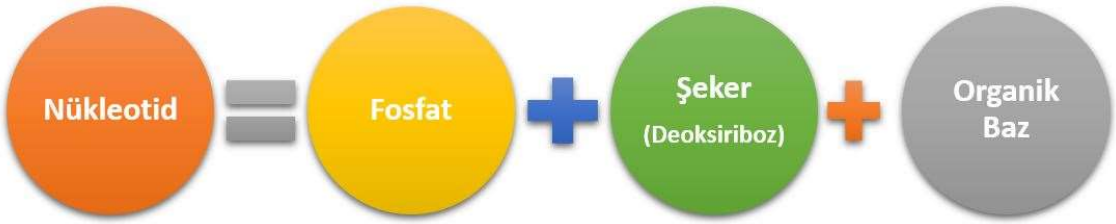
Şekil A-1. 4 Gen [93]

Canlıların kromozomları üzerindeki genlerinin birbirinden farklı olması sayesinde canlı çeşitliliği sağlanır. Nükleotid adı verilen yapılardan oluşan genler, bir canlı için gerekli olan proteinlerin inşasında görev alırlar ve DNA'nın en küçük yapı birimidirler. İnsan hücresinde birkaç yüz bazdan oluşan genler bulunduğu gibi, iki milyon civarında baz içeren genler de bulunmaktadır [95].

A-1.4 Nükleotid

Genleri oluşturan nükleotidler, DNA'nın en küçük yapı birimidir. Nükleik asitlerin yapı taşı olan nükleotidler; yapılarında bir adet beş karbonlu deoksiriboz şekeri, bir adet fosfat ve bir adet organik baz bulundururlar (Şekil A-1.5).

DNA'nın yapısında aynı fosfat ve deoksiriboz şekerinden oluşan dört farklı nükleotid bulunmaktadır (Şekil A-1.6). Nükleotidler yapılarında bulundurdukları bu organik bazlara göre (Şekil A-1.7) isimlendirilirler.

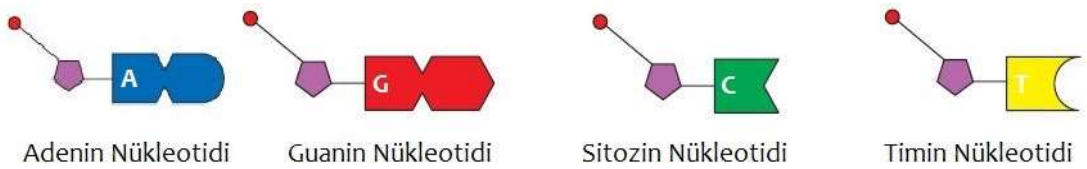


Şekil A-1. 5 Nükleotidin yapısı [93]

Nükleotidleri birbirinden ayırmamıza yarayan organeller yapılarında bulundurdukları Adenin (A), Guanin (G), Sitozin (S veya C) ve Timin (T) organik bazlarıdır.

Şekil A-1. 6 Nükleotidin yapısındaki bazlar ve diğer organeller [93]

Nükleotidler birleşerek DNA'nın yapısında yer alan genleri oluştururlar. DNA molekülünü meydana getiren zincirlerdeki nükleotidlerin yan yana dizilişleri belirli bir kural çerçevesinde olur ve Guanin Nükleotidi ile Sitozin Nükleotidi, Adenin Nükleotidi ile de Timin Nükleotidi eşleşir. Bir DNA molekülündeki adenin (A) ve timin (T) nükleotidlerinin sayıları ile guanin (G) ve sitozin (C) nükleotidlerinin sayıları da eşittir. Farklı bireyler farklı sayıda nükleotide sahiptirler ve ayrıca, bu nükleotidlerin dizilişleri de farklıdır. Nükleotidler, DNA zincirini oluşturmak için kimyasal bir bağ olan hidrojen bağı ile birbirlerine bağlanırlar. Adenin – timin ve guanin - sitozin nükleotid ikililerinin farklı sıralarda, alt alta dizilmesi ile de farklı DNA zincirleri meydana gelmektedir.

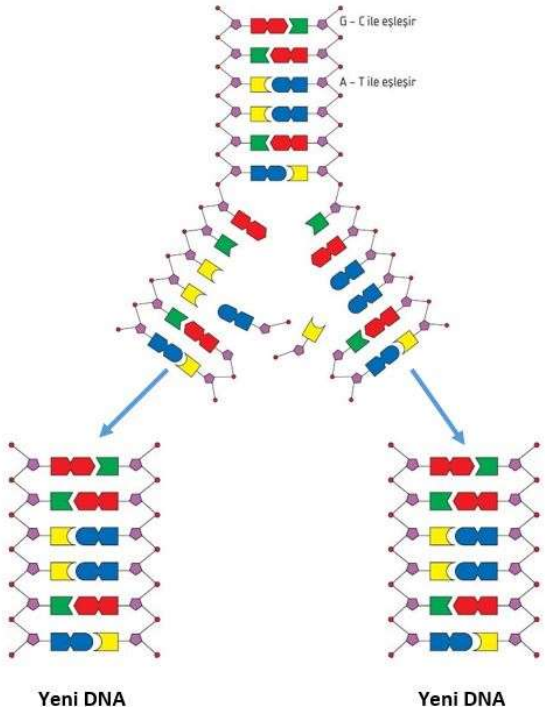


Şekil A-1. 7 Nükleotidler [93]

A-1.5 DNA'nın Kendini Eşlemesi

Hücre içerisinde sarmal bir yapıda bulunan DNA, çift zincirli bir yapıya sahiptir. DNA'nın sahip olduğu kalıtsal özellikleri yavru hücreye aktarması için kendini eşlemesi

gerekmektedir ve bu eşlemeyi gerçekleştirmesi için de DNA zincirleri birbirinden koparılmaktadır. Bu koparılma sonucu oluşan her bir zincir kendini eşleyerek, yine iki tane çift sarmal yapıda DNA oluşturmaktadır ve bu olaya DNA sentezi (replikasyon) adı verilmektedir. DNA'nın kendini eşlemesi sonucunda oluşan bu kopya sayesinde canlıya ait tüm kalıtsal bilgiler yavru hücrelere aktarılmaktadır. DNA'nın kendini eşlemesini (Şekil A-1.8) birkaç adımda özetleyecek olursak;



Şekil A-1. 8 DNA'nın kendini eşlemesi (Replikasyon) [93]

- DNA'yı oluşturan zincirli yapı, iki zinciri bir arada tutan hidrojen bağlarının koparılmasıyla bir fermuar dişlerinin birbirinden ayrılması gibi ayrılır.
- Ayrılan her bir nükleotid hücre sitoplazmasında serbest halde bulunan nükleotid ile birleşerek, A-1.4'te belirtildiği üzere adenin'in karşısına timin ve guanin karşısına sitozin gelecek şekilde birleşerek, birbiriyle aynı dizilime sahip olan çift sarmal yapıdaki DNA'yı oluştururlar.
- Sonuç olarak aynı özelliğe sahip çift sarmal yapıdaki iki DNA oluşmuş olur ve bu işlem sırasında oluşabilecek kodlama hatalarına da mutasyon adı verilir.

A-2 Genom, Genomik, Proteom ve Proteomik

Bir canlının kromozomlarında bulunan genetik şifrelerin tümünü ifade etmek için kullanılan Genom tanımı, ilk olarak Alman botanikçi Hans Winkler tarafından kullanılmıştır [96]. Fakat sonrasında unutulmuş olan bu terim 1986 yılında genetikçi Thomas Roderick tarafından, genomun haritalanması ve karakterizasyonunun tanımlanması için tekrar kullanılmıştır [97].

Genomda bulunan genetik şifrelerin bilgisinin ortaya çıkarılması için yapılan tüm ilgili çalışmalara genomik adı verilmektedir [98]. Genomik, bir canlıya ait tüm genlerin tek tek tanımlanmasına, genlerin zaman-yer-miktar olarak üretim ve aktivasyonlarının incelenmesine ve ayrıca, genlerin birbirleriyle ve çevreleriyle olan etkileşimlerinin araştırılmasına olanak tanır [96].

Ayrıca, genomik sayesinde farklı canlılara ait genetik bilgiler kıyaslanabilmekte, canlılar arasındaki ortak noktalar araştırılabilmekte, canlıların ürettikleri proteinlerin sayıları, çeşitleri ve bu proteinlerin fonksiyonları hakkında bilgi edinile bilinmektedir [96].

Genomik alanında yapılan çalışmaların amacı organizmalardaki hastalık ya da fizyolojik süreçlerle ilişkili genleri tanımlanmaktır [99]. Ancak, genomik alanında yapılan bu araştırmalar, tanımlanan bu genlerin organizmada hangi oranda kullanıldığı hakkında bilgi vermemektedir ve bir gen farklı biyolojik işleve sahip birçok proteini kodlayabilmektedir. Ayrıca, bir gen tarafından kodlanan bu proteinler, sentezlendikten sonra da değişimlere uğrayabilmektedir. Araştırmalar sentez sonrası gözlemlenebilen bu değişimlerin de gen işlevinden bağımsız olduğunu göstermiştir [31]. Bu sebeplerden dolayı, büyük boyutlarda bilimsel veri elde edilmesine karşın, sadece genomik çalışmalar baz alınarak yapılan teşhis yaklaşımları kliniksel kullanım için yeterli olmamaktadır [31][99][100]. Ayrıca, gen verilerinden elde edilen mRNA'ların ifade düzeyleriyle, bu mRNA'lar tarafından kodlanan proteinlerin miktarları arasında istatistiksel olarak zayıf bir doğrusal ilişkinin bulunduğu bildirilmiştir [30]. Tüm bu nedenlerden dolayı, genler tarafından üretilen proteinlerin, sentez sonrası değişimlerinin, hücrede bulundukları yerlerin ve göreceli miktarlarının anlaşılması için genomik sonrası bilgilere ihtiyaç duyulmaktadır. Bahsedilen bu ihtiyaçları karşılamak için de proteom ve proteomik teknolojisi geliştirilmiştir.

İlk kez 1994 yılında Marc Wilkins tarafından önerilmiş ve kabul görmüş olan proteom kelimesi, bir canlıda ya da bir hücrede belirli bir zaman (hastalık – sağlık, ergenlik – gençlik – yaşlılık, hücre döngüsünün farklı evreleri gibi) ve belirli bir ortamda (farklı hücre tipleri vs.) bulunan proteinlerin tümünü ifade etmektedir [100]. Bir canlıya ait protein ekspresyon verileri vücudun farklı bölgelerinde, hücre döngüsünün farklı evrelerinde ve farklı çevre koşullarında değişmektedir [31], [99], [100]. Genom, sabit bir yapıdadır ve bir organizma için çok iyi tanımlanabilmektedir. Proteom ise hücreden hücreye farklılık göstermekte, iç - dış uyarılarla biyokimyasal etkileşimlere girmekte ve bu etkileşimler sonucu sürekli değişim halindedir [100].

Proteomik ise belirli koşullar dâhilinde, belirli bir hücrede genom tarafından sentezlenen tüm proteinlerin incelenmesi, başka bir deyişle proteomun bir bütün halinde ele alınması işlemidir. Proteinlerin translasyon[1] sonrası değişimleri, doku ve hücrelerdeki görevleri, diğer proteinlerle ve makro yapılarla olan etkileşimleri proteomik ile ortaya çıkarılmaktadır. Yapılan biyolojik çalışmalarda hücrelerin fonksiyonlarının genler ve mRNA'lardan ziyade proteinler tarafından düzenlendiği bildirilmektedir [31]. Proteinlerin fonksiyonlarının belirlenmesinde hücredeki konumları, fizyolojik uyarılar sonucu konumlarında meydana gelen değişiklikler ve translasyon sonrası modifikasyonlar önemli yer tutmaktadır. Bu bilgilerde ancak proteomik çalışmalar ile tespit edilebilmektedir.

Bu tez çalışmasında proteomik verilerle çalışılmasının ana sebebi kanser gibi kompleks yapıya sahip hastalıkların araştırılmasında daha doğru sonuçlara ulaşılabilmesidir.

A-3 Gen Protein Eşlemesi

Çalışmamızda kullandığımız gen-protein eşlemesine ait bilgiler Çizelge A-2.1'de verilmiştir.

[1] Translasyon (translation), ribozomlarda gerçekleşen ve hücre çekirdeğinde sentezlenen mRNA'lardaki nükleotit dizilişlerine uygun olarak protein sentezlenmesi sürecini ifade etmektedir.

Çizelge A-2. 1 Gen Protein eşlemesi

	Protein Sembolü	Gen Sembolü		Protein Sembolü	Gen Sembolü
1	ARAF	ARAF	42	CD20	MS4A1
2	ACC1	ACACA	43	CD26	CD26
3	ACETYLATUBULINLYS40	TUBA1B	44	CD31	PECAM1
4	ACVRL1	ACVRL1	45	CD49B	ITGA2
5	ADAR1	ADAR	46	CDK1_pY15	CDK1
6	AMPKALPHA	PRKAA1	47	CHK1_pS296	CHEK1
7	ANNEXIN1	ANXA1	48	CHK2	CHEK2
8	ANNEXINVII	ANXA7	49	CIAP	BIRC2
9	AR	AR	50	CLAUDIN7	CLDN7
10	ARID1A	ARID1A	51	COG3	COG3
11	ASNS	ASNS	52	COLLAGENVI	COL6A1
12	ATM	ATM	53	CYCLINB1	CCNB1
13	AXL	AXL	54	CYCLIND1	CCND1
14	BRAF_pS445	BRAF	55	CYCLINE1	CCNE1
15	BAD_pS112	BAD	56	CYCLINE2	CCNE2
16	BAK	BAK1	57	DIRAS3	DIRAS3
17	BAP1C4	BAP1	58	DJ1	PARK7
18	BAX	BAX	59	DUSP4	DUSP4
19	BCLXL	BCL2L1	60	DVL3	DVL3
20	BCL2	BCL2	61	ECADHERIN	CDH1
21	BCL2A1	BCL2A1	62	EEF2	EEF2
22	BECLIN	BECN1	63	EEF2K	EEF2K
23	BETACATENIN	CTNNB1	64	EGFR	EGFR
24	BID	BID	65	EIF4E	EIF4E
25	BIM	BCL2L11	66	EIF4G	EIF4G1
26	BRCA2	BRCA2	67	ENY2	ENY2
27	BRD4	BRD4	68	EPPK1	EPPK1
28	CABL	ABL	69	ERALPHA_pS118	ESR1
29	CJUN_pS73	JUN	70	ERCC1	ERCC1
30	CKIT	KIT	71	ERCC5	ERCC5
31	CMET_pY1235	MET	72	ERK2	MAPK1
32	CMYC	MYC	73	ETS1	ETS1
33	CRAF_pS338	RAF1	74	FASN	FASN
34	CASPASE3	CASP3	75	FIBRONECTIN	FN1
35	CASPASE7CLEAVEDD198	CASP7	76	FOXM1	FOXM1
36	CASPASE8	CASP8	77	FOXO3A	FOXO3
37	CAVEOLIN1	CAV1	78	G6PD	G6PD
38	GAB2	GAB2	79	P27_pT198	CDKN1B
39	GAPDH	GAPDH	80	P38MAPK	MAPK14
40	GATA3	GATA3	81	P53	TP53
41	GATA6	GATA6	82	P62LCKLIGAND	SQSTM1

Çizelge A-2. 1 Gen Protein eşlemesi (Devamı)

	Protein Sembolü	Gen Sembolü		Protein Sembolü	Gen Sembolü
83	GCN5L2	KAT2A	124	P70S6K_pT389	RPS6KB1
84	HER2	ERBB2	125	P90RSK	RPS6KA1
85	HER3	ERBB3	126	PAI1	SERPINE1
86	HEREGULIN	NRG1	127	PARP1	PARP1
87	HSP70	HSP70	128	PAXILLIN	PXN
88	IGF1R_pY1135Y1136	IGF1R	129	PCNA	PCNA
89	IGFBP2	IGFBP2	130	PDL1	CD274
90	INPP4B	INPP4B	131	PDCD4	PDCD4
91	IRF1	IRF1	132	PDK1	PDPK1
92	IRS1	IRS1	133	PEA15	PEA15
93	JAB1	COPS5	134	PI3KP110ALPHA	PIK3CA
94	JAK2	JAK2	135	PI3KP85	PIK3R1
95	JNK2	MAPK9	136	PKCALPHA	PRKCA
96	KU80	XRCC5	137	PKCDELTA_pS664	PRKCD
97	LCK	LCK	138	PKCPANBETAII_pS660	PRKCB
98	LKB1	STK11	139	PR	PGR
99	MEK1	MAP2K1	140	PRAS40_pT246	AKT1S1
100	MIG6	ERRFI1	141	PRDX1	PRDX1
101	MRE11	MRE11A	142	PREX1	PREX1
102	MSH2	MSH2	143	PTEN	PTEN
103	MSH6	MSH6	144	RAB25	RAB25
104	MTOR	MTOR	145	RAD50	RAD50
105	MYH11	MYH11	146	RAD51	RAD51
106	MYOSINIIA_pS1943	MYH9	147	RAPTOR	RPTOR
107	NCADHERIN	CDH2	148	RB_pS807S811	RB1
108	NRAS	NRAS	149	RBM15	RBM15
109	NDRG1_pT346	NDRG1	150	RICTOR_pT1135	RICTOR
110	NFKBP65_pS536	RELA	151	S6	RPS6
111	NF2	NF2	152	SETD2	SETD2
112	NOTCH1	NOTCH1	153	SF2	SRSF1
113	PCADHERIN	CDH3	154	SHC_pY317	SHC1
114	P16INK4A	CDKN2A	155	SHP2_pY542	PTPN11
115	P21	CDKN1A	156	SLC1A5	SLC1A5
116	SMAC	DIABLO	157	TRANSGLUTAMINASE	TGM2
117	SMAD1	SMAD1	158	TSC1	TSC1
118	SMAD3	SMAD3	159	TUBERIN_pT1462	TSC2
119	SMAD4	SMAD4	160	VEGFR2	KDR
120	SNAIL	SNAI1	161	X1433BETA	YWHAB
121	SRC	SRC	162	X1433EPSILON	YWHAE
122	STAT3_pY705	STAT3	163	X1433ZETA	YWHAZ
123	STAT5ALPHA	STAT5A	164	X4EBP1	EIF4EBP1

Çizelge A-2. 1 Gen Protein eşlemesi (Devamı)

	Protein Sembolü	Gen Sembolü		Protein Sembolü	Gen Sembolü
165	STATHMIN	STMN1	170	X53BP1	TP53BP1
166	SYK	SYK	171	XBP1	XBP1
167	TAZ	TAZ	172	XRCC1	XRCC1
168	TFRC	TFRC	173	YAP	YAP1
169	TIGAR	C12ORF5	174	YB1	YBX1

EK-B

GELİŞTİRİLEN WEB UYGULAMASI

Bölüm 3'te GAÇ algoritmalarının başarımını değerlendirmek için kullandığımız verilerin diğer araştırmacıların kullanımına da sunulması için geliştirilen web uygulaması bu bölümde tanıtılmıştır.

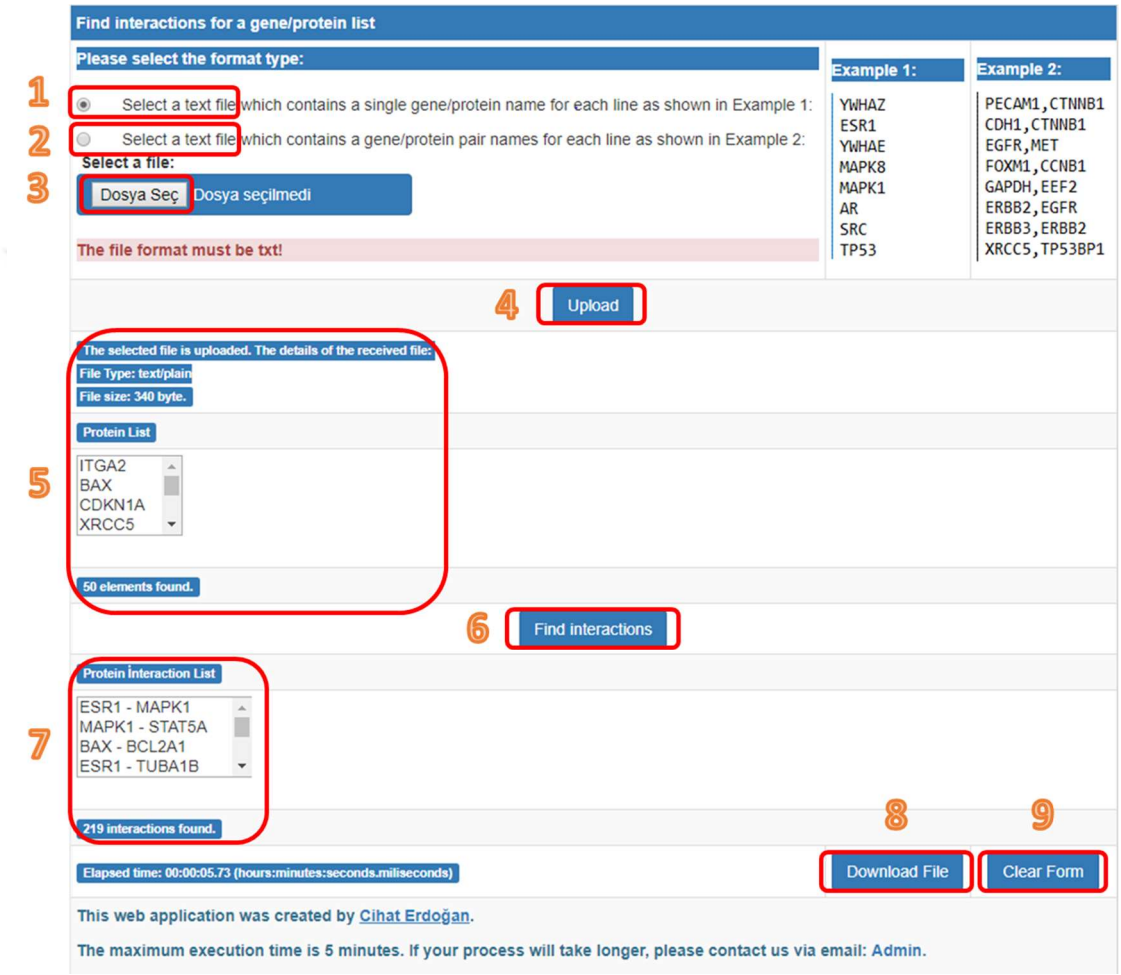
B-1 Geliştirilen Uygulama

Ağ çıkarım algoritmaları tarafından bulunan protein-protein ilişkilerinin literatürde yer alan veritabanlarına göre değerlendirilmesi için ASP.Net (R.Net [101] paketiyle) ve R dili kullanılarak Web tabanlı bir uygulama geliştirilmiştir ve uygulamanın ara yüzü Şekil B-1.1'de verilmiştir.

Literatürdeki ilişkileri bulmak için PC [50]'den indirilen BioPAX [51] dosyası içinde yer alan protein-protein ilişkileri Babur vd. [61] tarafından geliştirilen Java uygulaması kullanılarak tespit edilmiştir. Tespit edilen bir buçuk milyondan fazla protein-protein ilişkisi üzerinde diğer araştırmacılarında arama yapabilmesi için bu web uygulaması tasarlanmıştır.

Uygulamanın temelde iki farklı işlevi bulunmaktadır. Bunlardan birincisi çalışılmak istenilen gen/protein sembolleri, her satıra bir sembol gelmek koşuluyla bir text dosyasına yazılır. İlgi dosya Şekil B-1.1'de 1'nci adımda gösterilen seçenek işaretlenerek 3'ncü adımda gösterilen buton ile seçilir ve 4'ncü adımda ki butona basılarak sunucuya yüklenir. Dosya içindeki her bir satır okunarak ilgili gen/protein 5'nci adımda görüleceği üzere listbox'a eklenir ve dosyadan kaç adet gen/protein okunduğu bilgisi listbox'ın altında belirtilir. Son olarak 6'ncı adımda bulunan "Find interactions" butonu ile yüklediğimiz dosyada bulunan gen/protein'lerin literatürde olan ilişkileri (interactions)

bulunur ve 7'nci adımda gösterildiği üzere listbox içerisinde listelenir. İstenildiği takdirde bulunan ilişkiler 8'nci adımdaki "Download File" butonuyla indirilebilir. Yeni bir işlem yapmak için 9'ncu adımdaki "Clear Form" butonu ile web form temizlenebilir. Bu işlev ile çalışılan gen/protein'lerle ilgili doğrulanmış ağ (true network) oluşturulmak üzere belirlenen ilişkiler kullanıla bilinir. Örnek olarak Şekil B-1.1'de rastgele seçilen 50 protein için literatürdeki 219 protein-protein ilişkisi bulunmuştur.

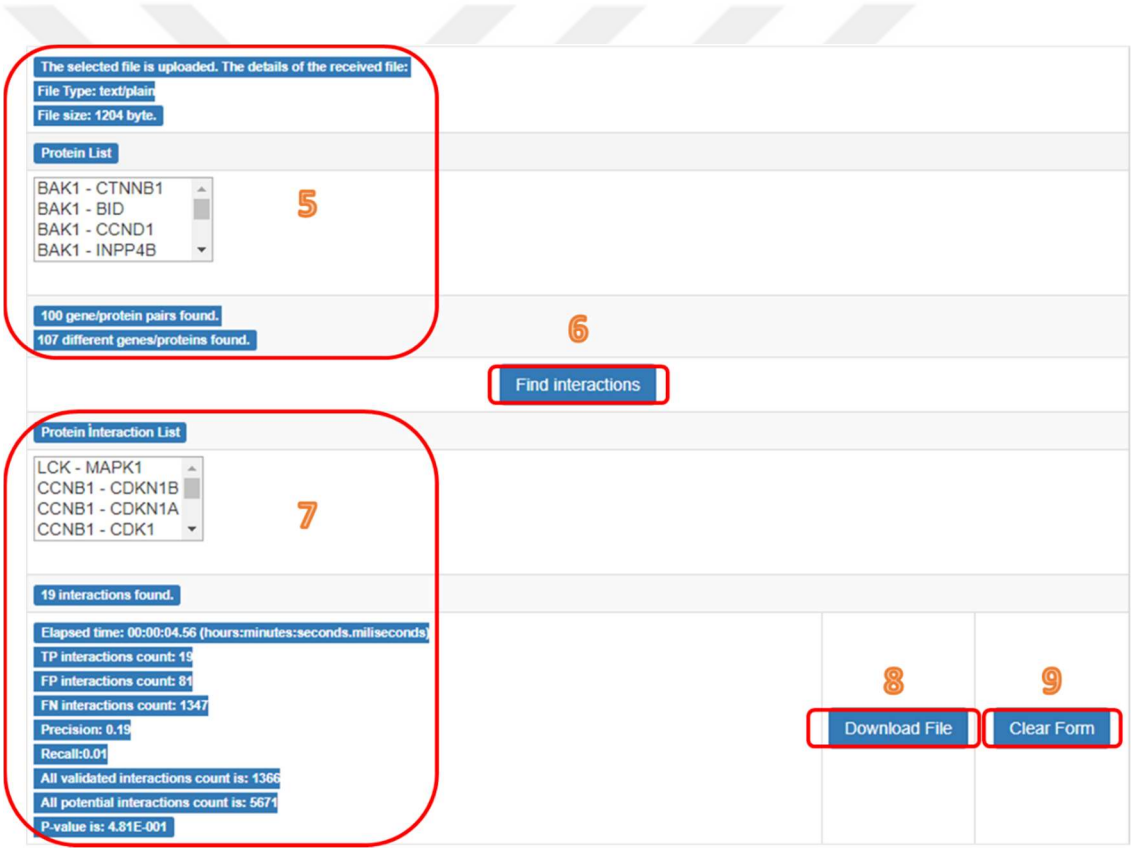


Şekil B-1. 1 Web Uygulama Ara yüzü - 1

Uygulamanın ikinci işlevi ise ağ çıkarım algoritmaları tarafından tahmin edilen ilişkilerin literatürdeki veritabanlarında olup olmadığını belirlemektir. Bunun için öncelikle ilişkili gen/protein çiftleri aralarına virgül konularak, her satıra bir çift protein veya gen sembolü gelecek şekilde bir text doyasına yazılır. Şekil B-1.1'de gösterilen adımlardan 2'nci adım seçilerek diğer adımlar ilk işlevde olduğu gibi sırasıyla gerçekleştirilir. Sonuç olarak Şekil B-1.2'de verilen 7'nci adım elde edilir. Verilen gen/protein çiftleri için literatürde olanlar belirlenerek 7'nci adımda yer alan listbox'a eklenir. Sisteme yüklenen

dosyadaki ilişkiler için doğru pozitif (TP), yanlış pozitif (FP), yanlış negatif (FN), kesinlik (precision) ve hassasiyet (recall) değerleri hesaplanarak gösterilir. Örnek olarak Şekil B-1.2'de sisteme yüklenen 100 adet protein-protein ilişkisinden TP olanların sayısı 19, FP olanların sayısı 81 ve FN olanların sayısı 1347 olmuştur. Bu değerlere göre kesinlik skoru 0,19 ve hassasiyet skoru 0,01 olarak bulunmuştur. Ayrıca, sisteme yüklenen 100 ilişki de toplam 107 farklı protein bulunmaktadır ve bu proteinlerin literatürde toplam 1.366 adet TP ilişkiye sahip olduğu bulunmuştur. Yüklenen dosyadaki toplam 107 farklı proteinin kendi aralarında olabilecek potansiyel ilişkilerinin sayısı da Bağıntı B-1.1 ile 5.761 olarak hesaplanmıştır. Burada n sayısı 107 olarak alınmıştır.

$$Tüm\ olası\ ilişkilerin\ sayısı = \frac{n \times (n-1)}{2} \qquad \text{(B-1.1)}$$



Şekil B-1. 2 Web Uygulama Ara yüzü - 2

Ayrıca, bulunan ilişkilerin istatistiksel olarak anlamlılığını değerlendirmek için FKT ile hesaplanan p-değeri de verilmiştir.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Cihat ERDOĞAN

Doğum Tarihi ve Yeri : 29.06.1982, Antalya

Yabancı Dili : İngilizce

E-posta : cerdogan@nku.edu.tr, cihaterdogan@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Müh.	Trakya Üniversitesi	2012
Lisans	Bilgisayar Müh.	Kırgızistan-Türkiye Manas Üniversitesi	2006

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2009	N.K.Ü. Çorlu Mühendislik Fak.	Araştırma Görevlisi

YAYINLARI

Makale

1. Erdoğan, C., Kurt, Z. ve Diri, B., (2017). "Estimation of the proteomic cancer co-expression sub networks by using association estimators", PLoS One, 12: e0188016.

2. Uzun, E. Erdoğan, C. ve Saygılı, A., (2016). "Hiyerarşik Kümeleme Modeli Kullanan Web Tabanlı Bir Ödev Değerlendirme Sistemi", Electronic Journal of Vocational Colleges, 6: 87-98.

3. Buluş, N., Erdoğan, C. ve Biri, B., (2016). "Metin Verilerde Dizgi Eşleme ve Sıkıştırılmış Dizgi Eşleme İşlemleri Arasındaki Performans Farklarının İncelenmesi", Electronic Journal of Vocational Colleges, 6: 60-76.

Bildiri

1. Erdoğan, C. Kurt, Z. ve Diri, B., (2017). “Ağ Çıkarım Algoritmalarındaki İlişki Tahmincilerinin Meme Kanseri Proteomik Verileri Üzerinde İncelenmesi”, 25th Signal Processing and Communications Applications Conference, 15-18 Mayıs 2017, Antalya.

2. Erdoğan, C. Kurt, Z. ve Diri, B., (2017). “An Analysis of the Association Estimators in Biological Network Inference Algorithms on Glioblastoma Multiforme Proteomic Data”, International Conference on Applied Analysis and Mathematical Modeling, 03-07 Temmuz 2017, İstanbul.

3. Uzun E., Buluş H. N., Erdoğan C., Kaya H., (2016). “İlişkisel Veritabanlarında Denormalizasyon Etkisi: Bir Anket Uygulaması”, Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı, 20-23 Ekim 2016, Tekirdağ.

4. Erdoğan C., Diri B., Buluş H. N., (2013). “Analyzing the performance differences between pattern matching and compressed pattern matching on texts”, International Conference on Electronics Computer and Computation, 07-09 Kasım 2013, Ankara.

Proje

1. Akıllı Ders Yönetim Sistemi ile Programlama Ödevleri için İntihal Tespiti Uygulaması, Yükseköğretim Kurumları Tarafından Destekli Bilimsel Araştırma Projesi, Araştırmacı, 01.01.2014-11.08.2015.