

**T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**İKİ KATMANLI ENTEGRASYON MİMARİSİYLE HASTALIĞA ÖZEL
BİRLİKTELİK AĞI ÇIKARIMI**

MUSTAFA ÖZGÜR CİNGİZ

**DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI**

**DANIŞMAN
PROF. DR. BANU DİRİ**

İSTANBUL, 2018

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**İKİ KATMANLI ENTEGRASYON MİMARİSİYLE HASTALIĞA ÖZEL
BİRLİKTELİK AĞI ÇIKARIMI**

Mustafa Özgür CİNGİZ tarafından hazırlanan tez çalışması 27.03.2018 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri 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. Zeynep ORMAN
İstanbul Üniversitesi

Prof. Dr. Uğur SEZERMAN
Acıbadem Üniversitesi

Yrd. Doç. Dr. Göksel BİRİCİK
Yıldız Teknik Üniversitesi

Bu alıřma, TBİTAK BİDEB 2211 Yurt İi Doktora Burs Programı tarafından desteklenmiřtir.

ÖNSÖZ

Tez aşamasında benimle bilgi birikimini paylaşan, motive eden ve desteğini benden esirgemeyen sevgili danışman hocam Prof. Dr. Banu DİRİ'ye teşekkürlerimi sunarım. Tez izleme komitemde yer alan ve değerli yorumlarıyla katkıda bulunan Prof. Dr. Nizamettin AYDIN'a ve Doç. Dr. Zeynep ORMAN'a teşekkürlerimi ayrıca belirtmek isterim. Doktora aşamasına gelene kadar beni okutan ve yetiştiren annem ve babam Yasemin CİNGİZ, Yaşar CİNGİZ'e, beni manevi anlamda ayakta tutan değerli eşim Betül CİNGİZ'e ve çocuklarım Göktuğ, Ekin ve Öykü CİNGİZ'e de teşekkür ederim. Hepiniz iyi ki varsınız. BİDEB 2211 Yurt içi Doktora Bursu kapsamında tez sürecinde katkıda bulunan TÜBİTAK'a da ayrıca teşekkürlerimi iletmek isterim.

Mart, 2018

Mustafa Özgür CİNGİZ

İÇİNDEKİLER

	Sayfa
SİMGE LİSTESİ.....	vii
KISALTMA LİSTESİ.....	viii
ŞEKİL LİSTESİ.....	ix
ÇİZELGE LİSTESİ	xi
ÖZET	
ABSTRACT.....	xiv
BÖLÜM 1	
GİRİŞ.....	1
1.1 Literatür Özeti	1
1.2 Tezin Amacı	5
1.3 Hipotez	7
1.4 Gen Birliktelik Ağları.....	7
1.5 Biyolojik Veri Kümeleri.....	11
1.5.1 Mikrodizin Gen İfadesi Veri Kümesi	12
1.5.2 RNA- Seq Veri Kümesi.....	14
1.5.3 MikroRNA- Hedef Gen İlişki Veri Kümesi.....	16
BÖLÜM 2	
GEN AĞI ÇIKARIM ALGORİTMALARI	17
2.1 C3NET (Conservative Causal Core Networks)	17
2.2 ARACNE (Accurate Cellular Networks)	19
2.3 WGCNA (Weighted Gene Co-expression Network Analysis).....	21
BÖLÜM 3	
miRNA- HEDEF GEN VERİLERİNDE BİRLİKTELİK KURALI TABANLI YAKLAŞIMLA GBA ELDE EDİLMESİ (ARNetMiT)	23

BÖLÜM 4

İKİ KATMANLI GBA ENTEGRASYON MİMARİSİ.....	33
--	----

BÖLÜM 5

GBA'LARIN PERFORMANS DEĞERLENDİRME KRİTERLERİ.....	40
5.1 Literatür Verileri Kullanımı ile Gerçeklenen Örtüşme Analizi Değerlendirmesi.....	40
5.2 GBA'ların Topolojik Değerlendirilmesi.....	43
5.3 GBA'ların Gen Ontoloji Terimleri Kullanarak Biyolojik Değerlendirilmesi	44

BÖLÜM 6

DENEYSEL SONUÇLAR	49
6.1 miRNA- Hedef Gen Veri Kümesini Kullanarak ARNetMiT Yaklaşımının Performans Değerlendirilmesi	49
6.1.1 ARNetMiT Yaklaşımının Örtüşme Analiz Değerlendirmesi.....	50
6.1.2 ARNetMiT Yaklaşımının Topolojik Değerlendirmesi.....	54
6.1.3 ARNetMiT Yaklaşımının Biyolojik Olarak Değerlendirilmesi.....	55
6.2 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Performans Değerlendirilmesi.....	57
6.2.1 Entegrasyon Çalışmasında Kullanılan Prostat ve Göğüs Kanserine Ait Veri kümeleri	57
6.2.2 Entegrasyon Çalışmasında Kullanılan Prostat ve Göğüs Kanserine Ait Veri kümelerinin Normalizasyonu	58
6.2.3 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Örtüşme Analizi Değerlendirilmesi	61
6.2.4 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Topolojik Özelliklerinin Değerlendirilmesi	67
6.2.5 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Biyolojik Olarak Değerlendirilmesi	72

BÖLÜM 7

SONUÇ VE ÖNERİLER	77
KAYNAKLAR.....	81

EK-A

GEN BİRLİKTELİK AĞLARININ F-ÖLÇÜM DEĞERİ İLE BİYOLOJİK DEĞERLENDİRMESİNİN GERÇEKLENMESİ	88
A-1 ARNetMiT Tabanlı Gen Birliktelik Ağlarının F-Ölçüm Değeri İle Değerlendirilmesi.....	89
A-2 İki Katmanlı Entegrasyon Mimarisiyle Elde Edilen Gen Birliktelik Ağlarının F-Ölçüm Değeri İle Değerlendirilmesi	90
ÖZGEÇMİŞ	93

SİMGE LİSTESİ

A	Komşuluk matrisi
$A_g(j,r)$	g geninin j ve r örneklerindeki okuma sayıları kullanılarak hesaplanan mutlak ifade seviyesi değeri
B	İkili değer döndüren fonksiyon
C	Ölçekleme Sabiti
D	Derece matrisi
D_j	j örneğine ait okuma sayısı
f_j	j örneğinin okuma değerine ait düzeltme katsayısı
F_1	F_1 -Ölçüm Değeri
$H(x)$	x değişkeninin belirsizlik değeri
K	Çekirdek Fonksiyon
$K_{g,j}$	g genin j örneğindeki okuma sayısı
L	Laplace matrisi
$M_g(j,r)$	g geninin j ve r örneklerindeki okuma sayıları kullanılarak hesaplanan log örnek değişikliği değeri (M- değeri)
$MI(x)$	x değişkeninin belirsizlik değeri
$MI(x,y)$	x ve y değişkenlerin ortak bilgi değeri
minD	Minimum destek eşik değeri
minG	Minimum güven eşik değeri
N	Toplam miRNA sayısı
$n(x)$	x genini regüle eden miRNA sayısı
$n(x, y)$	x ve y genlerini ortak regüle eden miRNA sayısı
$O_{x,y}$	x ve y genilişkinin gen ağı çıkarım algoritmaları tarafından tahmin sayısı
p	Hipergeometrik dağılım ile hesaplanan olasılık değeri
$r_{x,y}$	x ve y değişkenlerinin arasındaki Pearson korelasyon katsayı değeri
R^2	Belirleme Katsayısı
w_g	g geninin normalize edilmiş okuma değeri
α	Üs parametresi
β	Difüzyon parametresi

KISALTMA LİSTESİ

BÇO	Basit Çoğunluk Oyu
BÇOGBA	Basit Çoğunluk Oyu Tabanlı Gen Birliktelik Ağı
BKT	Birliktelik Kuralı Tabanlı
BTGBA	Birleşim Tabanlı Gen Birliktelik Ağı
cDNA	Tamamlayıcı DNA
dGBA	Destek Değeri Tabanlı Gen Birliktelik Ağı
dgGBA	Destek-Güven Değeri Tabanlı Gen Birliktelik Ağı
DN	Doğru Negatif
DP	Doğru Pozitif
GAÇ	Gen Ağı Çıkarımı
GBA	Gen Birliktelik Ağları
GDA	Gen Düzenleyici Ağ
GDOS	Gözlemlenen Değerin Ortalama Sapması
GO	Gen Ontolojisi
GTDGS	Gözlemlerin Tahmini Değere Göre Sapması
KTGBA	Kesişim Tabanlı Gen Birliktelik Ağı
MDDO	M- Değerlerinin Düzenlenmiş Ortalaması
MGİ	mRNA Gen İfadesi
miRNA	MikroRNA
mRNA	Mesajcı RNA
NCBI	National Center for Biotechnology Information
PPI	Protein- Protein İlişkisi
RNA- Seq	RNA- Sequencing (RNA Dizileme)
TCGA	The Cancer Genome Atlas
YN	Yanlış Negatif
YP	Yanlış Pozitif

ŞEKİL LİSTESİ

	Sayfa
Şekil 1. 1	Prostat kanseri ile ilişkili Gen Birliktelik Ağı 8
Şekil 1. 2	Gen ilişkileri gösteren kuvvet yasası dağılımı 9
Şekil 1. 3	Mikrodizin gen ifadesi deney adımları 13
Şekil 1. 4	Mikrodizin gen ifadesi veri örneği 13
Şekil 1. 5	RNA dizilime analizi temel adımları 15
Şekil 2. 1	C3NET algoritması gen ağı çıkarım örneği 19
Şekil 2. 2	ARACNE algoritması gen ağı çıkarım örneği 20
Şekil 3. 1	ARNetMiT yaklaşımının genel bakış açısı 24
Şekil 3. 2	ARNetMiT yaklaşımının sözde kodu(Pseudocode) 29
Şekil 3. 3	ARNetMiT temel adımlarının bir örnek üzerinde gösterimi 30
Şekil 4. 1	İki katmanlı GBA entegrasyonu mimarisi 34
Şekil 4. 2	İkinci entegrasyon aşamasına ait adımların sözde kodu 37
Şekil 5. 1	Topolojik değerlendirme temel adımları 43
Şekil 5. 2	GO- terimleri ve terimlerle ilişkili genlere ait örnek 45
Şekil 5. 3	GBA'ların biyolojik değerlendirmesine ait temel adımları 45
Şekil 5. 4	K- çekirdek graf (K=3) 46
Şekil 6. 1	Dört farklı kanser türüyle ilişkili dGBA'ların kesinlik ve DP değerleri 50
Şekil 6. 2	Dört farklı kanser türüyle ilişkili dgGBA'ların kesinlik ve DP değerleri 51
Şekil 6. 3	Farklı GAÇ algoritmalarıyla elde edilmiş GBA'ların kesinlik ve DP değerleri 52
Şekil 6. 4	GAÇ algoritmaları ve ARNetMiT tabanlı GBA'lardaki ortak tahmin edilen ilişki değerleri 53
Şekil 6. 5	GAÇ algoritmaları ve ARNetMiT tabanlı GBA'ların R^2 değerleri 55
Şekil 6. 6	GAÇ algoritmaları ve ARNetMiT tabanlı GBA'ların hastalıkla örtüşen GO-terimleri kullanılarak biyolojik olarak değerlendirilmesi 56
Şekil 6. 7	Dağılım normalizasyonu 59
Şekil 6. 8	MDDO normalizasyon örneği 61
Şekil 6. 9	Prostat kanserine ait başlangıç ve birinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları 62
Şekil 6. 10	Göğüs kanserine ait başlangıç ve birinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları 64
Şekil 6. 11	Prostat kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları 65

Şekil 6. 12	Göğüs kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları	66
Şekil 6. 13	Prostat kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların R^2 değerleri.....	68
Şekil 6. 14	Göğüs kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların R^2 değerleri.....	69
Şekil 6. 15	Prostat kanserine ait tüm entegrasyon aşamalarında elde edilen . GBA'ların R^2 değerleri.....	70
Şekil 6. 16	Göğüs kanserine ait tüm entegrasyon aşamalarında elde edilen ... GBA'ların R^2 değerleri.....	71
Şekil 6. 17	Prostat kanserine ait tüm entegrasyon aşamalarında elde edilen . GBA'ların hastalıkla örtüşen GO- terimleri kullanarak biyolojik olarak değerlendirilmesi	73
Şekil 6. 18	Göğüs kanserine ait tüm entegrasyon aşamalarında elde edilen ... GBA'ların hastalıkla örtüşen GO- terimleri kullanarak biyolojik olarak değerlendirilmesi	75
Şekil Ek-A. 1	GAÇ algoritmaları ve ARNetMiT tabanlı GBA'ların F-ölçüm değeri ile biyolojik olarak değerlendirilmesi.....	89
Şekil Ek-A. 2	Prostat kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların F-ölçüm değeriyle biyolojik olarak değerlendirilmesi	90
Şekil Ek-A. 3	Göğüs kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların F-ölçüm değeriyle biyolojik olarak değerlendirilmesi	91

ÇİZELGE LİSTESİ

	Sayfa
Çizelge 3. 1 Örnek miRNA- hedef gen ilişki tablosu	26
Çizelge 5. 1 Hata matrisi	42
Çizelge 6. 1 Hastalıklar ilişkili miRNA- hedef gen ilişki sayısı	50

TÜM GENOMDA GEN AĞI KESTİRİMİ İÇİN ÇEŞİTLİ VERİ KÜMELERİ ENTEGRASYONU

Mustafa Özgür CİNGİZ

Bilgisayar Mühendisliği Anabilim Dalı
Doktora Tezi

Tez Danışmanı: Prof. Dr. Banu DİRİ

Hastalıklara neden olan biyolojik mekanizmaların moleküler seviyede keşfi son yıllarda üzerinde önemle durulan bir çalışma alanıdır. İnsan genom projesinin tamamlanması ve DNA dizileme tekniklerindeki ilerlemelerle hastalıklara neden olan süreçler ve bu süreçler altında yatan moleküler ilişkiler daha hızlı ve kolay bir şekilde belirlenmektedir. Gen birliktelik ağları farklı fenotiplere sahip örneklerde benzer örüntüler sergileyen genler arasındaki ilişkilerden oluşan moleküler ilişki ağlarıdır. Benzer örüntü sergileyen genlerin yapılan literatür çalışmalarında benzer biyolojik fonksiyonlara sahip olduğu ve benzer süreçlerde yer aldıkları anlaşılmıştır. Tez kapsamında mikrodizin gen ifadesi, RNA- Seq ve miRNA- hedef gen verileri üzerinde farklı gen ağı çıkarım algoritmaları kullanılarak gen birliktelik ağları elde edilmiştir. Literatürde farklı biyolojik veri kümeleri ve gen ağı çıkarım algoritmalarını aynı yapı içerisinde kullanarak hastalıklarla ilişkilendirilmiş kapsamlı ve doğruluklu gen birliktelik ağı çıkarımı gerçekleştiren çalışma sayısı kısıtlıdır. Tez kapsamında temel hedefimiz farklı gen ağı çıkarım algoritmaları ve biyolojik veri kümelerini birlikte kullanarak göğüs ve prostat kanseriyle ilgili yüksek doğruluklu ve kapsamlı gen birliktelik ağları oluşturmaktır. Bu hedefle kesişim, basit çoğunluk oyu ve birleşim gibi temel entegrasyon yöntemleriyle iki katmanlı bir yapı oluşturularak gen birliktelik ağlarının entegrasyonunu gerçekleştirilmiştir. İki katmanlı entegrasyon mimarisi ile aynı veri kümesi üzerinde farklı gen ağı çıkarım algoritmalarından elde edilen gen birliktelik ağlarının entegrasyonu birinci entegrasyon aşamasında, farklı biyolojik veri kümeleri kullanılarak elde edilen gen birliktelik ağlarının entegrasyonu ise ikinci entegrasyon

aşamasında gerçekleşmiştir. Elde edilen gen birliktelik ağlarının performansı biyolojik ve topolojik özelliklerine göre değerlendirilmiştir. Bu iki değerlendirme kriterine ek olarak literatür verileri ile örtüşme analizi yapılmıştır. Sonuç olarak, sadece gen ağı çıkarım algoritmaları kullanılarak yapılan gen birliktelik ağı entegrasyonunun performans arttırımına etkisi kısıtlı iken, farklı biyolojik veri kümelerinden elde edilen gen birliktelik ağlarının entegrasyonunun performans artışı sağladığı gözlemlenmiştir. Çalışmamızda aynı zamanda daha önce miRNA- hedef gen verileri üzerinde gen birliktelik ağı çıkarımında hiç kullanılmamış hash tabanlı birliktelik kuralı algoritmasını uygulayarak ilgili veri kümelerinden teorik olarak 152 hastalıkla ilişkili GBA'lar elde edilmiş ve bu algoritma ARNetMiT R paketi halinde kullanıcılara sunulmuştur.

Anahtar Kelimeler: Mikrodizin gen ifadesi verisi, RNA- Seq verisi, miRNA- hedef gen verisi, Gen birliktelik ağları, Gen ağı entegrasyonu, Hash tabanlı birliktelik kuralı algoritması

**INTEGRATION OF VARIOUS DATASETS FOR GENOME-WIDE GENE
NETWORK INFERENCE**

Mustafa Özgür CİNGİZ

Department of Computer Engineering

PhD. Thesis

Adviser: Prof. Dr. Banu DİRİ

Discovering of biological mechanisms, which causes diseases at the molecular level, has been an important field of study in recent years. After the completion of the human genome project and the progress of DNA sequencing techniques, the relationships of molecules, which are related to cancer related biological processes, are revealed more quickly and easily. Gene coexpression networks are networks of molecular relationships formed by relation between genes which present similar patterns in samples with different phenotypes. Genes, which shows similar patterns on biological datasets, were found to have similar biological functions and involve in similar biological processes according to the studies in literature. In the thesis, gene coexpression networks were obtained by using different gene network inference algorithms on microarray gene expression, RNA-Seq and miRNA-target gene expression data. In the literature, there are limited number of studies that use different biological data sets and gene network inference algorithms in the same structure to construct a comprehensive and accurate gene coexpression networks for diseases. Our main target in thesis is to build up high-precision and comprehensive gene networks for breast and prostate cancer using different gene network inference algorithms and biological data sets together. According to this purpose, integration of gene coexpression networks has been achieved by forming a two-tier structure with basic integration methods such as intersection, simple majority voting and union. Gene expression networks, which are inferred by different gene network inference

algorithms on the same data, are integrated in the first integration phase of the two-tier structure. Gene expression networks that are inferred from different biological datasets are integrated in the second integration phase of two-tier structure. The performance of the obtained gene association networks was evaluated according to their biological and topological properties. In addition to these two evaluation criteria, an overlap analysis was performed with the literature data. The usage of different gene network inference algorithms at integration phase slightly increases performance slightly. However, integration of gene coexpression networks that are inferred on different biological datasets enhances the performances. We also developed a hash based association rule algorithm that infers potential gene coexpression networks of 152 diseases from miRNA- target gene data. We also present our algorithm as ARNetMiT R package.

Keywords: Microarray gene expression data, RNA- Seq data, MiRNA- target gene data, Gene coexpression networks, Integration of gene networks, Hash based association rule algorithm

1.1 Literatür Özeti

Hücrede yer alan moleküllerin karmaşık yapısı nedeniyle moleküllerin işlevi ve birbiriyle ilişkisi henüz tam anlamıyla ortaya çıkarılamamıştır [1]. Protein sentezi sürecinde meydana gelen farklılaşmaların metabolik hastalıklar, kanser gibi kompleks hastalıklara neden olduğu farklı çalışmalarda belirtilmiştir [2], [3]. Protein sentezindeki değişimlerin anlaşılması için proteinleri oluşturan genlerin belirlenmesi ve bu genlerin diğer genlerle ilişkilerin ortaya çıkarılması gerekmektedir. DNA dizileme tekniklerindeki gelişmeler moleküller arasındaki ilişkilerin ortaya çıkarılmasında önemli rol oynamaktadır [1], [4], [5], [6]. Çalışmamız kapsamında bu bölümde farklı biyolojik veri kümeleri kullanılarak elde edilen gen birliktelik ağı (GBA) çıkarım çalışmaları ve gen birliktelik ağlarının entegrasyonu ile literatürde yapılmış çalışmalar kısaca özetlenmiştir.

GBA'ların çıkarımıyla ilgili ilk çalışmalar mikrodizin gen ifadesi verileri kullanılarak elde edilmiştir. Mikrodizin veri kümelerinin dışında RNA-Seq [7], [8], [9] miRNA gen ifadesi [10], [11], ChIP-on-chip [12], ChIP-chip [13], [14] verileri gibi farklı veri kümelerini birlikte kullanımıyla kansere neden olan genlerin ve modüllerin belirlenmesi, gen ağı çıkarımı gibi farklı çalışmalar gerçekleştirilmiştir.

İlk GBA çıkarım çalışmalarında genler arasında ilişkilerin çıkartılması ve GBA'ların oluşturulması için bilgi teorisi tabanlı, regresyon tabanlı ve Bayes ağları tabanlı yaklaşımlar gen ifadesi verileri üzerinde kullanılmıştır [15], [16], [17]. Bilgi teorisi tabanlı yaklaşımlar genler arasındaki ilişkileri Pearson korelasyon katsayısı, ortak bilgi gibi bilgi teorisi tabanlı metriklerle ölçmektedir. Genler arasında ilişki değerleri

belirlelendikten sonra algoritmalar farklı yaklaşımlarla belirledikleri eşik değerini kullanarak gen ağı çıkarımlarını gerçeklerler [17]. Regresyon tabanlı yaklaşımlar her bir geni hedef gen olarak belirleyip, hedef genin gen ifadesi değerinin diğer tüm genlerin gen ifadelerinin doğrusal bir kombinasyonu olarak düşünür [17]. Hedef geni düzenleyen her bir gen için çoklu regresyon analizi ile katsayılar belirlenir ve bu işlem tüm hedef genler için yapılır. Çoklu regresyondan elde edilen tüm katsayılar gözlemlenen veriye göre minimum hata prensibiyle yeniden düzenlenir. Elde edilen son katsayılar genler arasındaki ilişkiyi göstermektedir. Bu yöntem genler arasındaki ilişkilerin doğrusal olduğu varsayımına dayanmaktadır. Biyolojik olarak genler arasındaki ilişkiler doğrusal olmamakla birlikte [17] yöntemin işlemsel karmaşıklığı da genom düzeyinde genler arasında ilişki belirlemede fazla vakit alabilmektedir. GBA çıkarımında kullanılan bir diğer popüler algoritma Bayes ağları yaklaşımıdır. Bayes ağları yaklaşımında genler arasındaki ilişkilerde yön bilgisi belirlenerek düzenleyici ve düzenlenen genler ortaya çıkarılmış olur. Bu yön bilgisi çıkarımı özellikle gen düzenleyici ağ (GDA) çıkarımında önemlidir. Bayes ağları olasılık ve çizgi teoremi kavramlarını kullanan olasılıksal grafik modelleridir [1]. Yönlü graf olan bu modelde genler (düğümler) rassal değişken, genler arasındaki yönlü ilişkiler (kenarlar) ise genler arasındaki koşullu olasılık dağılımına (conditional probability distribution) göre belirlenen ilişkileri göstermektedir. Rassal değişkenler koşullu olasılık dağılımına göre $P(x|P_a(x))$ şeklinde seçilir. P_a her bir düğümün ana düğümlerini (parent nodes) göstermektedir. Düğümler, ana düğümlerine bağımlı olmakla birlikte diğer düğümlerden bağımsızdır. Popüler üç farklı yaklaşımla elde edilen GBA'lar hastalıklarla ilişkili modüllerin keşfinde, farklı ifade edilen genlerin (differentially expressed genes) bulunması, merkez düğümlerin (hub nodes) belirlenmesinde, birbirleriyle ilişkisi ölçülmek istenen düğümler arasındaki mesafelerin ağlar üzerinde belirlenmesi gibi pek çok alanda kullanılmaktadır [18].

Hasta örneklerinden elde edilen GBA'lardaki ilişki sayısı genom düzeyinde fazla olacağı için hastalıkla ilişkili alt ağların (modül) keşfi ve bu alt ağların yer aldığı biyolojik süreçlerin anlaşılması GBA çalışmalarında önem taşımaktadır. Bununla birlikte normal hücrelerden alınan verilerden oluşturulan GBA'lar ile hastalıkla ilişkili hücrelere ait veriler kullanılarak elde edilen GBA'lar arasındaki topolojik ve ilişkisel farklılıkların

belirlenmesi moleküler bazda deęişimler ile hastalıklar arasındaki ilişkinin ortaya çıkmasını sağlamaktadır. Gen ağlarında bazı genlerin dięer genlerle fazla sayıda ilişkisi olduęu ve biyolojik süreçlerde düzenleyici olarak görev alması nedeniyle gen ağlarını kullanarak merkez düęümlerin belirlenmesi de GBA çalışmalarının önemli konularından biridir. Hastalıklarla ilişkili genlerden oluşturulan gen listelerinin gen ağları (veya protein- protein ağları) üzerinde birbirlerine uzaklığı veya yakınlığı da genler arasındaki ilişkiyi araştırmacılara sunmaktadır.

GBA'ların entegrasyonu ile ilgili çalışmalar literatürde iki ana başlık altında incelenebilir. Bunlardan ilki aynı türde ilişkileri içeren (homojen ilişkiler) ağların entegrasyonu dięeri ise farklı türde biyolojik ilişkileri birleştiren heterojen ağların entegrasyonu ile ilgili çalışmalardır [1]. Örneğin aynı mikrodizin gen ifadesi verisinde farklı gen ağı çıkarım algoritmalarıyla oluşturulmuş GBA'ların entegrasyonu homojen entegrasyona örnektir. Transkripsiyon faktörü- gen ilişkileri ile gen- gen ilişkilerini birleştiren entegrasyon çalışmaları da heterojen entegrasyona örnek verilebilir.

Ha ve arkadaşları [19] normal ve kanserli hücrelerden aldıkları mRNA gen ifadesi, kopya sayısı deęişimi, DNA metilasyonu ve miRNA gen ifadesi veri kümelerinde sağlıklı ve kanserli hücrelerle ilişkili GBA'ları ayrı ayrı belirlemiştir. Daha sonra tüm veri kümeleri için normal ve kanserli hücrelere ait GBA'lardaki farkları bularak 4 veri kümesine ait kanserle ilişkilendirilmiş alt modüller çıkartılmış ve bu modüllerde geçen merkez genleri kullanarak PI3K yolağının (pathway) beyin tümörüyle ilişkili olduęu sonucuna varmışlardır. Farklı biyolojik veri kümelerinden alınan sonuçların entegrasyonu moleküler düzeyde daha güvenilir ilişkilerin çıkarımı açısından önem taşımaktadır. CMGRN [14] fare hücrelerinden alınan transkripsiyon faktör- gen, miRNA- gen ve gen- gen ilişkileri entegre edilerek heterojen bir ağ entegrasyonu yapılmıştır. Bu entegrasyonun bir bütün halinde gerçekleştirilmesinde moleküler ilişkilerde ortak olarak yer alan genlere ait ilişkiler bağlayıcı görevde yer almıştır. Göv ve Arga [20] ise miRNA- gen ve TF- gen ilişkilerini heterojen bir şekilde entegre ederek hiyerarşik bir regülasyon ağı elde etmişlerdir. Bir dięer çalışmada [21] farklı kanser türleriyle ilişkili 60 kanserli hücreye ait mikrodizin gen ifadesi verisi (NCI- 60) kullanılarak 64 tane kanserle ilişkilendirilmiş gen elde edilmiş ve bu genlerin kopya sayısı deęişimleri incelenmiştir. Çalışma sonunda gen kopya sayısındaki deęişimleri fazla olan bazı genler kanserle

ilişkilendirilmiş ve aynı genlerin gen ifadesi veri kümesindeki mRNA değerlerinin düşük olduğu gözlemlenmiştir. Vu ve Vohradsky [13] transkripsiyon faktörlerin genom üzerindeki bağlandığı alan bilgisi, ChIP-chip verisi ve mRNA gen ifadesi verilerini kullanarak maya (yeast) hücreleriyle ilgili transkripsiyon- gen ilişkilerini kapsayan moleküler ağ oluşturmuştur. Çalışma ilk olarak transkripsiyon faktör- gen ilişkilerini veren ağlar basit operatörlerle birleştirilmiş daha sonra mRNA gen ifadesinden elde edilen gen ağlarının ortak kesişimi alınarak yeni bir ağ oluşturulmuştur. Sun ve ekibi [22] mRNA gen ifadesi, DNA metilasyon verisi ve gen kopya sayısı değişimlerini birlikte kullanarak göğüs kanserine neden olan genleri çalışmalarında belirlemişlerdir.

Heterojen entegrasyon çalışmaları çoğunlukla moleküler bazda düzenleyici (regulator) ve düzenlenen (regulating) ilişkilerini belirlemek için kullanılmaktadır. Bu alandaki entegrasyon çalışmaları GBA çıkarımından öte GDA çıkarımına dayanmaktadır.

Farklı biyolojik veri kümelerden elde edilen aynı türdeki moleküler ilişkilerin entegrasyonunda en sık kullanılan yöntemler ağ tabanlı yöntemler ve çekirdek fonksiyon temelli makine öğrenmesi yöntemleridir [1]. Ağ tabanlı yaklaşımlarla birden fazla biyolojik veri kümesi kullanılarak moleküler ilişkiler elde edilebilir. Ağ modelleri farklı veri kümelerinden elde edilen aynı tür ilişkileri birleşim, kesişim gibi basit operatörler kullanılarak entegre gen ağlarını oluşturmaktadır. Aynı tür ilişkileri içeren GBA'ların entegrasyonunda her bir GBA'nın entegrasyona etkisi farklı entegrasyon aşamasında entegre edilerek her bir GBA için farklı ağırlık değerleri belirlenebilir. Predictionet [23] literatür verisi olarak kullanılan protein- protein ilişkileri ve mRNA gen ifadesi verisi kullanılarak elde edilen gen- gen ilişkilerini ağırlıklandırma yöntemiyle birleştiren çalışmalardan biridir. Homojen veri entegrasyonunda çekirdek fonksiyon tabanlı (kernel based methods) yaklaşımlar popüler olarak kullanılmaktadır. Entegrasyonla oluşturulan ağlardaki moleküler ilişkiler difüzyon çekirdek fonksiyonlar yardımıyla belirlenir. Temel olarak $K = e^{-\beta L}$ olarak gösterilen difüzyon çekirdek (diffusion kernel) fonksiyonlarda K_{ij} ağ üzerinde düğüm "i" ve düğüm "j" arasındaki yakınlık (closeness) değerini göstermektedir. β difüzyon parametresidir. L matrisi Laplacian matrisi olmakla birlikte $L=D-A$ denklemiyle elde edilmektedir. Burada D düğümlerin derecelerini gösteren köşegen matris (diagonal matrix), A ise komşu matrisi (adjacency matrix) sembolize etmektedir. Laplacian matrisini kullanarak elde

edilen düğümlerin ağ üzerindeki yakınlık değerlerine benzer değerler rassal yürüyüş (Random Walk) tabanlı farklı difüzyon çekirdek yöntemleri kullanılarak da çıkartılabilmektedir. En bilinen ve sıklıkla kullanılan rassal yürüyüş tabanlı difüzyon çekirdek fonksiyonlarından biri "gidiş- geliş zamanı" çekirdek fonksiyondur (Commute Time Kernel). Bu fonksiyon ile ağ üzerinde ortalama olarak bir düğümden diğer düğüme kaç iteratif adımda varıldığını ve tekrar başlangıç düğümüne geri dönüldüğü bilgisi elde edilir. Bu fonksiyona göre birbirlerine daha yakın olan düğümlerin ağ üzerindeki yakınlık değerleri daha yüksek olacaktır [1], [24]. Çekirdek tabanlı yaklaşımların en büyük dezavantajı çekirdek fonksiyonu seçimidir. Bu problemi ortadan kaldırmak için bir veri kümesi üzerine farklı benzerlik ölçümü yapan değişik çekirdek fonksiyonları uygulanır ve elde edilen sonuçlar doğrusal olarak birleştirilir [25]. Heriche ve ekibinin yaptığı çalışma çekirdek bazlı entegrasyon çalışmaları için başarılı bir örnektir [24]. Bu çalışmada protein- protein ilişki ağı, biyolojik süreçlerle (GO terimleri) ilgili ilişkileri içeren ağ, mRNA gen ifadesinden elde edilmiş gen- gen ilişkileri ve metin madenciliği ile belirlenmiş gen gen ilişkilerini kapsayan ağların herbiri için Laplacian matrisi oluşturulmuştur. Laplacian matrisini kullanarak moleküller arasındaki yakınlık değerlerini gösteren çekirdek difüzyon matrisleri her bir veri kümesi için çıkartılmıştır. Daha sonra ayrı ayrı elde edilen çekirdek matrisleri doğrusal olarak birleştirilerek elde edilen gen ağı merkez düğümlerin keşfinde kullanılmıştır.

Çalışmamızda mikrodizin gen ifadesi, RNA-Seq ve miRNA- hedef gen ilişki verileri üzerinde elde edilen GBA'lar kesişim, birleşim ve basit çoğunluk oylama (veya ortak karar, Simple Majority Voting) yöntemleri kullanarak geliştirdiğimiz iki katmanlı entegrasyon mimarisiyle birleştirilmiştir. Tez kapsamında basit operatörlerin kullanılmasındaki en önemli neden entegrasyon sürecindeki hesaplama karmaşıklığının az olmasıdır. Geliştirdiğimiz iki katmanlı birleştirme çatısının birinci ve ikinci entegrasyon aşamalarındaki tüm GBA'lar üzerinde birleşim, kesişim ve ortak karar operatörleri uygulanmıştır.

1.2 Tezin Amacı

Hastalıkları anlamak, hastalıklara neden olan moleküler düzeydeki ilişkileri keşfetmek ve elde edilen bilgilere göre tedavi yöntemleri geliştirmek tıp, moleküler biyoloji,

genetik, ilaç geliştirme gibi sağlıkla ilişkili birçok alanın temel çalışma alanıdır. Hücrelerde protein senteziyle görevli genlere ait DNA dizilimlerindeki değişim protein sentezindeki olağan akışı değiştirebilmektedir [26], [27]. Canlılarda protein sentezi sırasında meydana gelen farklılaşmalar kanser gibi farklı hastalıklara neden olmaktadır [2], [3]. Hastalıkların altında yatan mekanizmanın anlaşılmasında, olağan protein sentezi yapısı değişen genlerin belirlenmesi ve bu genlerin birbirleriyle ilişkilerinin ortaya çıkarılması önemli rol oynar [28], [29], [30]. Genler arasındaki ilişkilerin keşfiyle hastalıkla ilişkilendirilebilecek gen modülleri (ilişkili gen kümeleri) belirlenebilmektedir.

Çalışmamızda farklı biyolojik veri kümeleri ve gen ağı çıkarım algoritmaları kullanarak prostat ve göğüs kanseriyle ilişkili gen- gen ilişkilerinin oluşturduğu gen birliktelik ağları elde edilmiştir. Tezimizin temel amacı, farklı biyolojik veri kümeleri ve gen ağı çıkarım algoritmalarını birlikte kullanarak hastalıklarla ilişkili kapsamlı ve yüksek doğruluklu gen birliktelik ağlarını oluşturmaya çalışmaktır. Bu amaç için farklı biyolojik veri kümeleri ve farklı algoritmalarla elde edilen GBA'lar çalışmamızda entegre edilmiştir. Farklı veri kümeleri üzerinde ve farklı algoritmalar tarafından elde edilen GBA'ların entegrasyonu iki katmanlı bir mimari yaklaşımıyla gerçekleştirilmiş ve tüm entegrasyon aşamalarında yer alan GBA'ların performans sonuçları değerlendirilmiştir.

Çalışmamızda veri kümeleri olarak mikrodizin gen ifadesi (mRNA gen ifadesi, microarray gene expression), RNA-Seq (RNA- Dizileme, RNA- Sequencing) ve MikroRNA (miRNA)- hedef gen ilişki verileri kullanılmıştır. Mikrodizin gen ifadesi ve RNA- Seq verileri üzerinde gen birliktelik ağları gen ağı çıkarım algoritmalarıyla direkt oluşturulmaktadır. Aynı algoritmalar ile miRNA- hedef gen ilişki verileri üzerinde ön işlem adımları uygulanarak Gen Birliktelik Ağı çıkarımı gerçekleştirilebilir. Tez kapsamında miRNA- hedef gen ilişki verileri üzerinde literatürde henüz uygulanmamış hash tabanlı birliktelik kuralı algoritması uygulayarak ilgili veri kümelerinden teorik olarak 152 hastalıkla ilişkili GBA'lar elde edilmiştir. miRNA- hedef gen ilişki verileri üzerinde GBA çıkarımı ve hash tabanlı birliktelik kuralı algoritması uygulanabilmesi için bir R yazılım paketi de çalışmamız kapsamında geliştirilmiştir.

1.3 Hipotez

GBA'lar farklı biyolojik veri kümelerinde benzer örüntüler taşıyan ve ortak ifade edilen genler olarak adlandırılan genlerden oluşan gen ağlarıdır. GBA'daki genlerin hastalıkla ilişkili biyolojik süreçlerde birlikte yer aldığı varsayımı biyolojide kabul gören bir yaklaşımdır [30].

Tez kapsamında farklı hastalıklarla ilgili kapsamlı ve yüksek doğruluklu GBA'ların oluşturulması temel hedefimizdir. Bu amaç için farklı biyolojik veri kümelerinin ve farklı gen ağı çıkarım algoritmalarının birlikte kullanılarak hastalıklarla ilişkili kapsamlı ve doğruluklu gen birliktelik ağları çıkarımı yapılması öngörülmüştür. Farklı gen ağı çıkarım algoritmalarını kullanarak bir veri kümesi üzerinde oluşturulacak GBA'ların tek bir gen ağı çıkarım algoritmasına bağımlılığı giderilmiş olur. Benzer bir durum biyolojik veri kümeleri için de geçerlidir. Farklı biyolojik veri kümeleri üzerinde elde edilen GBA'ların entegrasyonu ile biyolojik olarak daha gerçekçi ve kapsamlı GBA'ların elde edileceği düşünülmüştür.

Gen Ağı Çıkarım algoritmaları ve farklı biyolojik veri kümelerini eş zamanlı kullanarak hastalıklarla ilişkili GBA çıkarımı için tez çalışmamızda İki Katmanlı Entegrasyon Mimarisi yaklaşımı geliştirilmiştir. Bu yapıda farklı veri kümeleri üzerinde farklı gen ağı çıkarım algoritmaları tarafından geliştirilen GBA'lar iki aşamalı olarak birbirlerine basit operatorler kullanarak entegre edilmiştir.

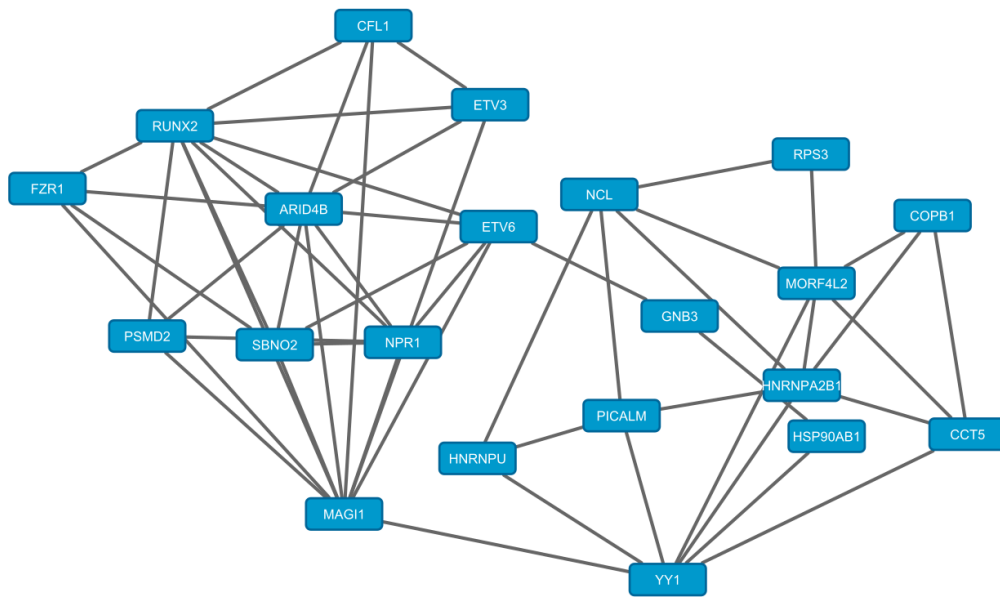
Çalışmamızda GBA entegrasyonunun yanı sıra daha önce miRNA- hedef gen verileri üzerinde GBA çıkarımında hiç kullanmamış hash tabanlı birliktelik kuralı algoritması geliştirilmiş ve R paketi haline getirilmiştir. Hash tabanlı birliktelik kuralı algoritması, "ortak olarak çok sayıda aynı miRNA tarafından regüle edilen genler ortak ifade edilen genlerdir" varsayımına dayanarak geliştirilmiştir.

1.4 Gen Birliktelik Ağları

Kromozom üzerinde farklı sayıda nükleotit dizisinden oluşan ve kalıtım sağlayan birimlere *gen* denir [30], [31]. İnsanda 20 bin civarında bulunan genlerin en temel işlevi proteinleri oluşturan amino asitlerin dizilimini belirlemektir [32]. Yapılan çalışmalarda özellikle şeker hastalığı, kanser, kalp hastalıkları gibi karmaşık hastalık mekanizmasına

sahip hastalıklara sadece bir genin değil, birden fazla genin yer aldığı biyolojik olay ve süreçlerin neden olduğu belirtilmiştir [33], [34], [35]. Bu nedenle genlerin ilişkilerinin keşfi hastalıkları anlamak açısından önemlidir.

Son yıllarda DNA dizileme tekniklerinde (DNA sequencing) yaşanan gelişmeler farklı yaklaşımlarla elde edilen biyolojik veri kümesi çeşitliliğini arttırmıştır. Biyolojik veri kümelerini oluşturan örneklerde (farklı deneysel koşullarda veya zamanlarda) benzer örüntü gösteren genler ortak ifade edilen genler (co-expressed genes) olarak adlandırılmaktadır [36]. Ortak ifade edilen genlerin ağ yapısında gösterimiyle Gen Birliktelik Ağları meydana gelmektedir. GBA'larda genler düğüm (node), genler arasındaki ilişkiler ise kenar (edge) olarak gösterilmektedir. GBA'larda gen- gen ilişkileri arasında bir düzenleyici- düzenlenen ilişkisi gibi bir yön bilgisi bulunmadığı için GBA'lar yönsüz çizgelerdir (undirected graph). Şekil 1.1'de mikrodizin verileri kullanılarak elde edilen prostat kanser ile ilişkili GBA yer almaktadır. Şekil 1.1'deki dikdörtgenler genleri (düğümleri), dikdörtgenler arasındaki bağlantılar ise genlerin birbirleriyle olan ilişkilerini göstermektedir.



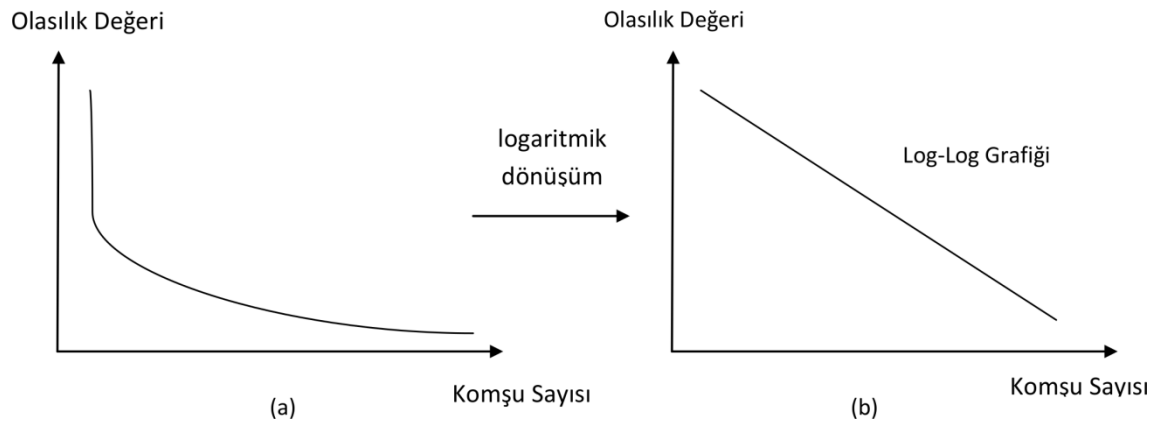
Şekil 1. 1 Prostat kanseri ile ilişkili Gen Birliktelik Ağı

GBA'ların biyolojik olarak önemli olmasının en büyük sebebi GBA'ları oluşturan genlerin farklı hastalıklara neden olan biyolojik süreçlerde, olaylarda ortak olarak yer aldığı varsayımına dayanmasıdır [33], [34], [35]. Bunun dışında GBA'da yer alan genlerin protein üretim seviyelerini (gen ifadesi değerleri) düzenleyen transkripsiyon faktör olan

ilişkilerinin belirlenmesi yani Gen Düzenleyici Ağların (GDA) çıkarılması da biyolojide moleküler seviyede hastalıkların mekanizmasını anlamak açısından önem göstermektedir [37], [38].

Gen ağları olan GBA ve GDA'lar genellikle ölçeksiz ağlardır (scale free network) [39], [40], [41]. Ölçeksiz ağlarda yer alan düğümlerin derece değerleri kuvvet yasası dağılımı (power law distribution) özelliğini gösterir [40], [42]. Ölçeksiz ağlarda yer alan az sayıda düğümün çok sayıda düğümlle ilişkisi bulunurken çoğunlukta olan diğer düğümlerin az düğümlle ilişkisi bulunmaktadır. Biyolojik ağlarda çok sayıda düğümlle ilişkisi olan düğümlere merkez düğüm (hub node) denilmektedir. Ölçeksiz ağ üzerinden rassal olarak seçilmiş bir düğümün x tane düğümlle ilişkisinin (kenarı) bulunma olasılığı kuvvet yasası fonksiyonuna göre $p(x) = Cx^{-\alpha}$ denklemiyle belirlenmektedir. Burada C kuvvet yasasının ölçeklenmesi için kullanılan sabiti, α ise üs parametresi olarak gösterilmiştir. Chung [43] ölçeksiz ağ modeline göre üs parametresinin değeri 2 ile 3 arasında bir sayı olarak belirtmiştir. Üs parametresi gen birliktelik ağlarındaki düğüm derecelerinin kuvvet yasası dağılımına göre uygun olarak belirlenmesini sağlamaktadır.

Gen ağlarında az sayıda genin çok sayıda ilişkisinin olması, geri kalan genlerin ise az sayıda ilişkisinin olması genlerin derecelerini gösteren kuvvet yasası dağılımının sağa çarpık bir dağılım olmasına neden olur. Şekil 1.2'de genlerin derece değerlerine ait kuvvet yasası dağılımı örnek olarak gösterilmiştir.



Şekil 1. 2 Gen ilişkileri gösteren kuvvet yasası dağılımı

Sağa çarpık dağılım, ilişki sayısı az olan genlerin yoğunlukta olduğunu göstermektedir. Rasgele seçilmiş bir genin Şekil 1.2(a)'da gösterildiği gibi az sayıda komşusunun olma olasılığı fazladır. Şekil 1.2(a)'da gösterilmiş grafik genlerin komşuluk sayısının kuvvet

yasası dağılımını göstermesi nedeniyle doğrusal değildir. Log-log graflar sayesinde komşu sayısı dağılımının doğal logaritması alınarak genler arasındaki komşuluk sayısı doğrusal bir şekilde gösterilebilir. Buna göre kuvvet yasası fonksiyonu $\log(p(x)) = \log(C) - \alpha \log(x)$ halini almış olur.

Logaritmik dönüşümden sonra Şekil 1.2(b)'de gösterilen ilişki sayısı ve ilişki sayılarının olasılık değeri arasındaki dağılım doğrusal olarak modellenenebilir bir yapıya dönüştürülmüştür. Literatürde yapılan farklı çalışmalarda veriler üzerinde regresyon analizi uygulanarak doğrusal model oluşturulmuş ve bu modelin başarısı uygunluk iyiliği (goodness of fit) değerlendirilmesiyle ölçülmüştür [44]. Regresyon analizi ile bağımsız bir (veya birden fazla) değişkenle bağımlı değişken arasındaki ilişki belirlenmektedir. Gen ağlarında yer alan genlerin ilişkilerine ait olasılık dağılımlarını tahmin edebilmek için doğrusal model logaritmik dönüşümle oluşturulmuş olur. Buna göre bağımlı değişken olan olasılık değerleri Y olarak, bağımsız değişken olan ilişki sayıları X olarak gösterildiğinde iki değişken arasındaki ilişki denklemiyle gösterilen bir doğru denklemi olarak tanımlanır. β_0 , $X=0$ olduğundan bağımlı değişkenin alacağı değerleri gösterirken, ϵ değeri hata terimi, β_1 regresyon katsayısı olarak gösterilmektedir.

Elde edilen basit doğrusal modelin (simple regression model) uygunluk iyiliği, belirleme katsayısı (R-kare- R^2 , coefficient of determination) istatistiği ile ölçülebilir. R^2 değeriyle model aracılığıyla belirlenen tahmini değerlerin gözlenen değerlerle uyumluluğu değerlendirilir. Güvenilir bir regresyon modelinde model tarafından tahmin edilen değerlerin gözlenen değerlere yakın çıkması beklenir. En küçük kareler yöntemiyle (least squares) tahmini değerler ve gözlenen değerler arasındaki farkın karesinin az olması modelin başarılı olduğunu göstermektedir. Modelin başarılı olması ise en küçük kareler yöntemiyle bağımsız değişkene ait katsayıların başarılı kestirilmesi gerekir (β_0 , β_1). Denklem 1.1'de en küçük kareler yöntemiyle β_0 , β_1 değerlerini belirleyen formül gösterilmiştir.

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \quad (1.1)$$

$X=0$ iken, elde edilen β_0 değeri ve eğimi veren β_1 değeri kullanılarak modeli tanımlayan doğru denklemi belirlenmiş olur. R^2 istatistiği en küçük kareler yöntemiyle elde edilen

doğru denkleminin gerçek verileri tahmin gücünü ölçmektedir. R^2 değeri 0 ile 1 arasında değer almaktadır. Bu değer 1'e yakın olması model ile tahmin edilen tahmini verilerin gözlemlenen gerçek verileri iyi temsil ettiğini göstermektedir. Yüksek R^2 değeri bağımlı değişkenin bağımsız değişken tarafından iyi tahmin edildiği bilgisini de vermektedir. Değerin sıfıra yaklaşması doğru denklemden elde edilen veriler ile gerçek verilerin tahmininin zayıf belirlendiğini gösterir. Buna göre R^2 istatistiği denklem 1.2 ile hesaplanır.

$$R^2 = 1 - \frac{GDOS}{GTDGS} \quad (1.2)$$

- GDOS (Gözlemlenen Değerin Ortalama Sapması): Gözlemlenen değerlerin ortalama sapmasını göstermektedir. $GDOS = \sum_i (y_i - \bar{y})^2$ olarak hesaplanır. Buna göre $\bar{y}$ gözlemlenen değerlerin ortalamasını, y_i ise y'nin tüm gözlemlenen (gerçek) değerlerini göstermektedir.
- GTDGS (Gözlemlerin Tahmini Değere Göre Sapması): Gözlemlenen değerlerin tahmini değerlere göre sapmasını göstermektedir. $GTDGS = \sum_i (y_i - \hat{y}_i)^2$ olarak hesaplanır. Bu denklemde $\hat{y}_i$ doğrusal model tarafından i'inci örneğin tahmini değerini göstermektedir.

R^2 değeri ortalama değer kullanmak yerine regresyon denkleminiyle tahmin edilen değerleri kullanarak oluşturulan modelin başarısını ölçmektedir. Çalışmamızda yüksek R^2 değerleri GBA'larının ölçeksiz ağ tanımını iyi yansıttığını yani gen ağı üzerindeki genleri temsil eden düğümlerin komşuluk değerlerinin kuvvet yasasına uygunluğunu göstermektedir. Gelecek bölümlerde R^2 ölçümünü kullanarak gen ağı çıkarımı yapan algoritmalar da çalışmamızda yer verilmiştir. R^2 değeri aynı zamanda literatürde yapılan çalışmalarda da [44] GBA'ların performans değerlendirilmesinde de kullanılmıştır. Tez kapsamında GBA'ların performansının ölçümü için R^2 istatistiğinden yararlanılmıştır.

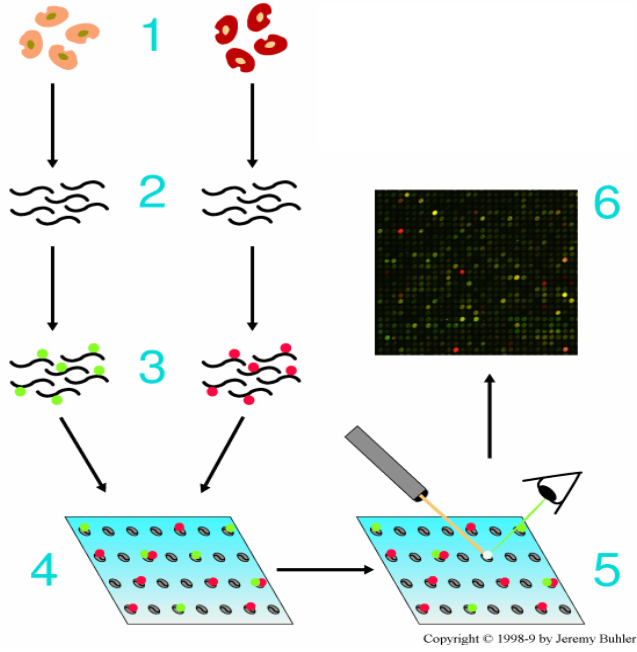
1.5 Biyolojik Veri Kümeleri

Çalışmamız kapsamında GBA'ların çıkarımı için üç farklı biyolojik veri kümesi kullanılmıştır. Bu veri kümeleri mikrodizin gen ifadesi, RNA-Seq ve miRNA- hedef gen verileridir. Gen ağlarının çıkarımı ve literatürde gen ağlarıyla ilgili çalışmaların daha iyi anlaşılması için bu bölümde üç farklı biyolojik veri kümesi detaylıca anlatılmıştır.

1.5.1 Mikrodizin Gen İfadesi Veri Kümesi

Mikrodizin (microarray) gen ifadesi (mRNA gen ifadesi, MGİ) verileri GBA ve GDA çıkarımında uzun süredir sıklıkla kullanılan veri kümeleridir [45], [46], [47]. MGİ verileri genlerin gen ifadesi profillemesinde (gene expression profiling) kullanılarak binlerce genin protein sentezi için anlık aktivitesini gösterebilmektedir [47]. Genlerin farklı hücrelerdeki (normal veya kanserli gibi) protein sentezi aktivitesinin ölçülmesiyle iki fenotip arasında farklılık bulunan genler pek çok çalışmada hastalıklarla ilişkilendirilmektedir [48], [49], [50]. Mikrodizin veri kümesi analizinde kanserli hücre ve normal hücre gibi farklı tip hücrelerden elde edilen mesajcı RNA'lar (mRNA) ayrı ayrı RNA izolasyonu ile hücreden alınır. Tersine transkripsiyon (reverse transcription) ile tek sarmallı mRNA'nın kopyaları DNA benzeri çift sarmallı yapıya çevrilir. Burada amaç mRNA'nın üretildiği gen dizilimlerini içeren DNA formatına çevrilerek genlerin ifade değerlerinin belirlenmesidir [51]. Tersine transkripsiyon ile tamamlayıcı DNA'lar her bir farklı hücre için elde edildikten sonra (complementary DNA, cDNA) bir fenotip bir kimyasal renk verici maddeyle (örnek olarak yeşil renk normal hücreler için seçilmiş) bir diğer fenotip ise başka bir kimyasal renk verici madde ile (örnek olarak kırmızı renk kanserli hücreler için seçilmiş) işaretlenir ve her bir örnekten alınan cDNA'lar renklerle etiketlenmiş olur. Farklı tip hücrelerden elde edilen cDNA'lar birleştirilerek (hybridization) mikrodizin çiplerinin üzerine uygulanır. İki farklı fenotipin takibi için farklı tip hücrelerden alınan örneklere ait cDNA'lar aynı çiplerin üzerine uygulanmaktadır.

Mikroçipler üzerinde farklı renklerle işaretlenmiş bir tarayıcıdan geçirilerek mikroçip üzerindeki her bir gen parçasına (probe) ait alanın renk yoğunluğu (intensity) değerleri veri olarak bu deney tipinde sunulmaktadır. Bir gen parçasının renk değerinin kırmızı olması kanserli ilişkili hücrelerde aktif olduğunu, yeşil olması ise kanserle ilişkili olmayan hücrelerde aktif olduğunu, sarı olduğunda da her iki farklı durumda aktif olduğunu göstermektedir. Temel adımları yukarıda anlatılan Mikrodizin gen ifadesi analizine ait işlem adımları Şekil 1.3'te gösterilmiştir.



Şekil 1. 3 Mikrodizin gen ifadesi deney adımları [52]

Gen parçası değerlerinin yoğunluk değerlerini kullanarak genler arasındaki ilişki çıkarımı, hastalık gibi farklı fenotiple ilişkilendirilebilecek genlerin keşfi, tek nükleotit polimorfizmi gibi farklı amaçlar için kullanılmaktadır. Şekil 1.4'de deney sonucunda elde edilmiş gen parçalarının yoğunluk değerleri gösterilmiştir. Şekilde satırlar gen parçalarını, sütunlar ise farklı örnekleri veya zaman serilerini göstermektedir. Mikrodizin veri kümesindeki değerler ise genlerin yoğunluk değerleridir.

coli_v4_Build_6:allProbes	dinI_U_N0025_r1	dinI_U_N0025_r2	dinI_U_N0025_r3	dinP_U_N0025_r1	dinP_U_N0025_r2	dinP_U_N0025_r3	lexA_U_N0025_r1	lexA_U_N0025_r2	lexA_U_N0025_r3	lon_U_N0025											
aaaD_b4634_3	9.05367	9.07827	9.10114	9.19489	9.22786	9.25386	9.23052	9.98833	9.19785	9.03439	9.12375	9.18492	9.04063	9.10267	8.92786	9.06723	9.27864	8.90923	9.19974	9.32354	9.09219
aaaA_b3241_14	7.20924	7.08695	7.07071	7.21464	7.12724	7.24448	7.03049	6.96986	7.15323	6.99117	7.23193	7.1495	7.02843	7.20205	7.33154	7.33489	7.16526	7.12292	7.11731	7.3624	7.10905
aaaB_b3240_15	7.21535	7.14312	7.19478	7.30189	7.51344	7.1945	7.31944	7.29981	7.23088	7.10381	7.28395	7.17703	7.10019	7.41297	7.43705	7.52368	7.12164	7.08363	7.39178	7.39168	7.3184
aaaR_b3243_15	7.9327	7.86983	7.90374	8.02228	7.88712	7.9801	7.89692	7.91303	7.73329	7.90787	7.95218	7.9709	7.90234	7.71327	7.88102	7.94579	7.91599	7.90248	7.80187	7.74775	7.77202
aaaX_b3242_12	6.86667	6.75021	6.78882	6.6893	6.89494	6.9454	7.10521	6.57845	6.70136	6.79866	6.97202	6.89901	6.81862	6.65609	6.80885	6.69436	6.70972	6.82944	6.87576	6.9831	6.70891
aaaS_b2836_14	9.07763	9.04851	9.21648	9.12263	9.14818	9.1291	9.1207	9.23898	9.14278	9.06859	9.19071	9.30914	9.09163	8.99722	9.24814	9.10847	8.97032	8.74454	9.13941	9.13932	9.23066
aat_b0885_14	9.20777	9.02169	8.9419	8.96712	8.87269	8.91636	8.98629	9.11163	8.85853	9.01781	8.9083	9.08188	9.05198	8.92606	9.05962	8.95219	8.91123	8.91701	9.00326	8.93538	8.76614
abgA_b1338_14	8.12799	8.01818	8.06149	8.1553	8.2973	7.95071	8.12016	8.16646	8.198	8.04075	8.07261	8.0114	8.23125	8.16428	8.12018	7.98084	8.1128	7.93974	7.9962	8.12524	8.12265
abgB_b1337_15	7.15056	7.16019	7.24002	7.33097	7.42327	7.16392	7.44575	7.50291	7.32466	7.05091	7.32685	7.17605	7.28411	7.44618	7.21557	7.08942	7.45388	7.09266	7.2307	7.41096	7.33228
abgR_b1339_15	7.26254	7.15513	6.96854	7.30765	7.34121	7.10777	7.21983	7.38239	7.28996	7.06932	7.19596	7.10023	7.36936	7.36394	7.30789	7.08776	6.96589	7.12941	7.13876	7.20285	7.18163
abgT_b1336_15	6.62351	6.41542	6.42616	6.59251	6.6761	6.32518	6.61029	6.53633	6.79119	6.58328	6.62812	6.60595	6.58225	6.63619	6.63512	6.57634	6.53203	6.3309	6.55034	6.73807	6.46953
abxR_b0715_15	8.10194	8.13186	8.12541	8.11332	7.95794	8.20392	8.21608	8.03594	8.02516	8.00686	8.08453	7.97701	8.15335	7.98077	8.07252	8.07061	7.93277	7.99823	8.19065	8.1062	8.07024
accA_b0185_15	10.1539	10.1922	10.1517	10.3141	10.408	10.4816	10.2298	10.3993	10.3913	10.3372	10.3024	10.1073	10.2445	10.0376	10.3728	10.2696	10.389	10.4823	10.3176	10.308	10.2709
accB_b3255_15	11.6726	11.8348	11.7188	11.8119	11.8117	11.8038	11.7549	11.7918	11.7288	11.7381	11.8344	11.6565	11.7215	11.8026	11.8439	11.3897	11.4889	11.6803	11.8889	11.8847	11.8517
accC_b3256_15	10.153	10.435	10.2373	10.1045	10.4774	10.3365	10.2623	10.1791	10.3979	10.3156	10.5706	10.2372	10.3409	10.5126	10.3611	10.2413	10.3364	10.1035	10.35	10.2024	10.6213
accD_b2316_14	10.2082	10.2521	10.2624	10.2534	10.4145	10.1413	10.3837	10.365	10.4033	10.3832	10.3636	10.2175	10.1267	10.3601	10.2098	10.2936	10.5353	10.5349	10.191	10.2905	10.2342
aceA_b4015_15	8.09666	8.08428	8.1626	7.79923	7.78849	8.10021	8.13938	8.08887	8.0485	8.03961	7.84616	8.12924	8.04705	8.36998	8.05708	8.42852	8.26682	8.08212	7.85933	7.79905	8.01578
aceB_b4014_15	8.89153	8.97226	8.72618	8.5562	8.46798	8.83046	9.09192	9.00062	9.00896	9.0089	8.62123	9.04459	8.9609	8.96147	8.9349	9.26684	9.09429	9.07036	8.90447	8.88927	8.94458
aceE_b0114_15	12.7735	12.742	12.6837	12.6944	12.8127	12.8789	12.9083	12.9293	12.9183	12.7219	13.0436	12.3916	12.8569	12.7569	13.0815	12.7837	12.7582	12.9366	13.0875	12.6642	12.8904
aceF_b0115_15	12.3123	12.3109	12.1912	12.1929	12.3299	12.367	12.3364	12.4616	12.6689	12.3109	12.5401	11.8973	12.5167	12.5826	12.6828	12.321	12.1772	12.3063	12.5126	12.3525	12.6217
aceK_b4016_15	7.4924	7.71401	7.56302	7.70624	7.62002	7.63776	7.66458	7.78263	7.74853	7.46224	7.63297	7.68469	7.60705	7.71412	7.7077	8.59386	8.41017	8.21084	7.75971	7.86206	7.68277
ackA_b2296_15	12.6552	12.5835	12.6195	12.7294	12.7632	12.6833	12.5872	12.7314	12.5427	12.4727	12.7685	12.2742	12.7742	12.861	12.8957	12.2417	12.349	12.5493	12.9745	13.0087	12.8235
acnA_b1276_15	8.93919	8.80787	8.79764	8.74927	9.00591	8.71075	8.84355	8.70136	8.94933	8.88614	8.74539	8.77307	8.90466	8.84678	8.70443	8.86805	8.73831	8.65248	8.75591	8.77455	8.87354
acnB_b0118_15	9.79234	9.81436	9.96646	9.85312	10.1696	10.0926	10.002	10.2859	10.3087	10.2869	10.2878	10.1146	10.0314	9.76352	10.2142	10.0443	9.89844	10.0137	10.2667	10.0114	10.2255
acpH_b0404_14	7.47842	7.4249	7.40762	7.30794	7.46942	7.32686	7.39047	7.39204	7.48709	7.2125	7.43665	7.41915	7.35831	7.57948	7.30734	7.3966	7.46232	7.05853	7.62391	7.4573	7.42289
acpI_b1094_15	13.2492	13.1935	13.1438	13.0994	13.1187	13.1793	13.0751	13.0574	13.0978	13.1219	13.0612	13.191	13.0978	13.0767	13.1371	13.1195	13.1194	13.0712	13.0934	13.0494	13.1178
acpS_b2563_15	9.31929	9.27136	9.4016	9.44906	9.39099	9.25032	9.28077	9.32832	9.25534	9.32275	9.26998	9.43073	9.18669	9.28377	9.26839	9.05626	9.14367	9.19382	9.26132	9.33947	9.2234
acpT_b3475_15	9.29445	9.17089	9.15995	9.08804	9.20287	9.06666	9.02668	9.07899	9.00868	8.84388	9.1725	9.0681	9.13035	9.14207	9.01817	9.20936	9.15944	9.18116	9.12452	9.23247	8.99276

Şekil 1. 4 Mikrodizin gen ifadesi veri örneği

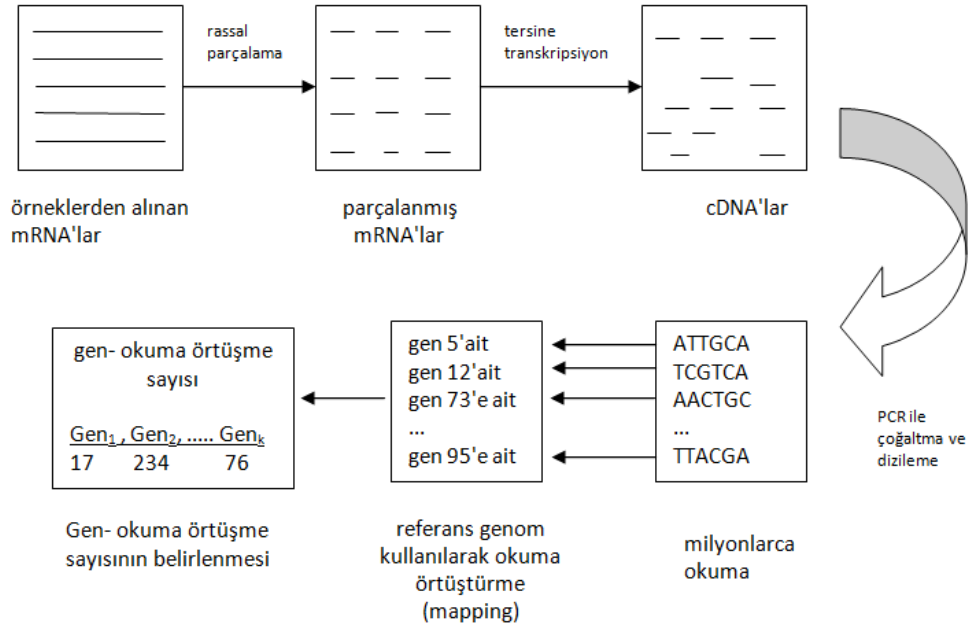
1.5.2 RNA- Seq Veri Kümesi

Son yıllarda GBA çıkarımında kullanılan bir diğer biyolojik veri kümesi de RNA dizileme (RNA sequencing, RNA- Seq) veri kümesidir. Gelecek nesil dizeleme teknikleri kullanarak elde edilen RNA- Seq verileri belirli bir zamanda bir biyolojik örnekteki RNA (mesajcı RNA, taşıyıcı RNA, ribozomal RNA, mirna-siRNA gibi mikro RNA'lar) miktarını göstermektedir [53], [54], [55]. RNA-Seq analizi transkriptoma kapsamlı bir bakış açısı sağlamanın yanı sıra RNA'yı üreten DNA alanlarının belirlenmesi ve bu alanların genlerle ilişkilendirilmesinde de sıklıkla kullanılmaktadır. RNA-Seq analizi, RNA moleküllerinden DNA dizilimi parçalarını elde etmeye yarayan deneysel adımları içerir. RNA-Seq ile genom üzerinde genlere ait daha keşfedilmemiş gen izoformlarının belirlenmesi mümkündür. RNA-Seq analiziyle tek nükleotit polimorfizmi, mutasyon, yeni gen izoformların keşfi ve gen birleşimi (genlere ait dizilimlerin birbirleriyle belirli oranlarda birleşmesi) gibi pek çok çalışma alanında kullanılmaktadır [54].

RNA- Seq analizinden önce aşağıdaki işlemlerin gerçekleştirilmesi gerekmektedir:

- RNA Çıkarımı (RNA Extraction): Tüm transkriptomun hücre içerisinden çıkartılması gerekir.
- RNA İzolasyonu (RNA Isolation): Üzerinde analiz yapılacak RNA türünün belirlenmesi ve türe ait örneklerin alınması aşamasıdır. RNA-Seq analizi çalışmalarında genellikle mesajcı RNA (mRNA) örnekleri ile çalışılmaktadır.

RNA çıkarımı ve izolasyonu adımımdan sonra tüm transkriptomu elde etmemizi sağlayan, yeni gen varyasyonlarını açığa çıkaran ve kısa mRNA okumalarını (reads) kullanarak genlerin gen ifadesi verilerini elde etmemize olanak sunan RNA- Seq analizi gerçekleştirilebilir. Şekil 1.5'te RNA- Seq analizindeki temel adımlar sıralanmıştır.



Şekil 1. 5 RNA dizilime analizi temel adımları

Şekil 1.5'e göre RNA-Seq analizinin temel adımları sırasıyla aşağıda maddeler halinde gösterilmiştir.

- mRNA'yı rassal olarak parçalarına ayırma (RNA Fragmentation)
- Tersine transkripsiyon ile mRNA kullanılarak cDNA'ların oluşturulması
- Polimeraz Zincir Reaksiyonu (PCR, Polymerase Chain Reaction) ile cDNA'ların on/yüz milyonlarca küçük DNA dizilimleri/parçaları olan okumalarının (reads) oluşturulması. Okumalar ortalama olarak 25- 125 bp (baz çifti) uzunluğundadır.
- On/yüz milyonlarca okumanın referans genom üzerine örtüştürülmesi (mapping).
- Referans genomda her bir gene karşılık gelen okuma sayısının çıkartılarak genlerin gen ifade değerlerinin belirlenmesi (gene expression quantification).

MGİ deneyindeki gibi RNA-Seq deneyinin sonucunda da genlerin gen ifadesi değerleri elde edilmiş olur. Gen ifadesi değerlerini kullanarak RNA-Seq verileri üzerinde GBA çıkarımı yapan çalışmalar son yıllarda oldukça popüler hale gelmiştir [44], [7], [8], [9]. MGİ verileri üzerine uygulanan temel gen ağı çıkarım adımları RNA-Seq verileri üzerine de uygulanarak GBA çıkarımı gerçekleştirilir. İlk adımda gen ifadesi verileri normalize

edilir, daha sonra genler arasındaki ilişkiler farklı gen ağı çıkarım algoritmaları kullanılarak belirlenir ve GBA çıkarımı RNA-Seq verileri kullanılarak oluşturulmuş olur.

1.5.3 MikroRNA- Hedef Gen İlişki Veri Kümesi

MikroRNA (miRNA) ortalama 21-23 nükleotit uzunluğunda olan hayvanlarda, bitkilerde ve bazı virüslerde bulunabilen RNA türüdür. miRNA'lar her ne kadar protein sentezi doğrudan katılmasalar da protein sentezini etkileyebilmektedirler [56], [57]. DNA'dan *RNA polimeraz II* enzimiyle ilk olarak çift sarmallı ortalama 100-120 baz çifti uzunluğunda RNA (pri-miRNA) elde edilir. *pri-miRNA*, *Drosha* ve *Pasha* enzimleriyle *pre-miRNA* yapısına çevrilir ve *exporting 5* enzimiyle çekirdekten sitoplazmaya aktarılır. *pre-miRNA*, *Dicer* enzimiyle sitoplazmada tekrar parçalanan ve miRNA- miRNA ikili (duplex) formuna dönüştürülür. miRNA-miRNA ikili formu iki parçaya ayrıldıktan sonra *Argonaute* enzimine bağlanır ve *RISC (RNA-induced silencing complex)* kompleksi oluşmuş olur. *RICS* kompleksi hedef genin ürettiği mRNA'ya bağlanarak onun yıkımına neden olur ve protein sentezi durdurulmuş olur [58].

miRNA'ların, gen ifadesi değerlerini değiştirmesiyle hücrede meydana gelen protein sentezi değişimleri kanser gibi pek çok hastalıkla ilişkilendirilmiştir [58], [59]. Bu nedenle miRNA- hedef gen ilişkilerinin keşfi hastalıkların anlaşılması için önem taşımaktadır.

Çalışmamız kapsamında miRNA- hedef gen ilişkilerini kullanarak GBA çıkarımı gerçekleştirilmiştir. MGİ ve RNA-Seq verisinden farklı olarak miRNA -hedef gen verilerinde ilişkiler 0 veya 1 olarak tutulmaktadır. Bir miRNA ile gen arasından ilişki varsa bu ilişkinin değeri 1 yoksa ise 0 olarak belirlenmiştir. MGİ ve RNA-Seq'te kullanılan gen ağı çıkarım algoritmalarının yanı sıra hash tabanlı birliktelik kuralı algoritması tez kapsamında geliştirilerek miRNA- hedef gen verileri üzerinde GBA çıkarımı gerçekleştirilmiştir.

GEN AĞI ÇIKARIM ALGORİTMALARI

Çalışmamızda bilgi teorisi tabanlı Gen Ağı Çıkarım (GAÇ) algoritmaları olan C3NET, ARACNE ve WGCNA algoritmaları kullanılmıştır. MiG, RNA-Seq ve miRNA -hedef gen verileri üzerinde GBA çıkarımında kullanılan algoritmaların içeriklerine bu bölümde değinilmiştir.

2.1 C3NET (Conservative Causal Core Networks)

C3NET, GBA çıkarım çalışmalarında özellikle basit ve hızlı olması nedeniyle tercih edilen gen ağı çıkarım algoritmalarındandır. Üç temel adımdan oluşan C3NET'in ilk adımı, genler arasındaki ilişkiler ortak bilgi hesaplamasıyla belirlenir.

$$\begin{aligned} MI(X, Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \quad (2.1)$$

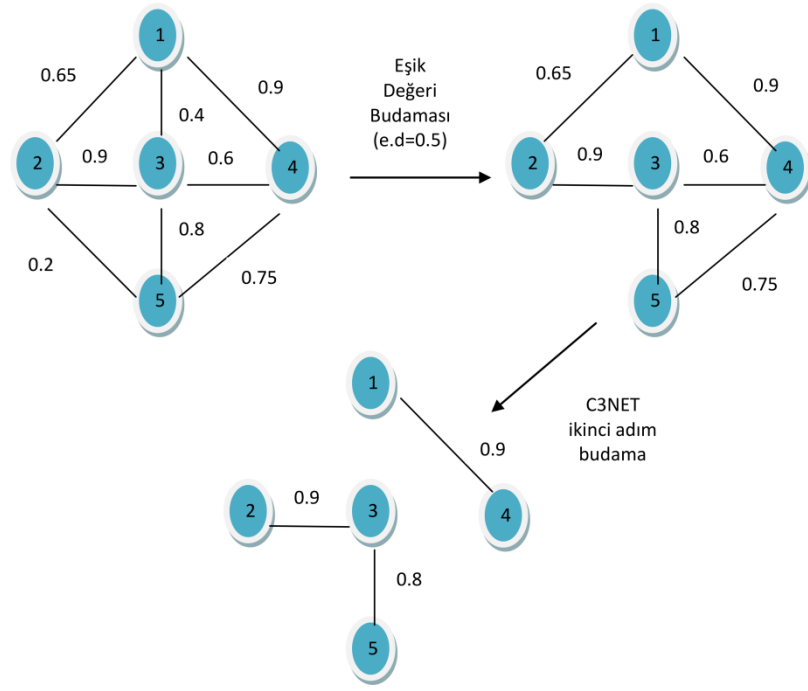
$$MI(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Denklem 2.1'de yer alan MI rassal x ve y değişkenlerinin arasındaki ortak bilgi değerini göstermektedir. Ortak bilgi değeri bir rassal değişkenin başka bir rassal değişken kullanarak ne ölçüde bilgi edinilebileceğini göstermektedir. Değişken dağılımında rassal bir değişkenin bilgi değerinin başka bir değişkene bağımlılığını ölçmek için bilgi teorisinde koşullu entropi metriği kullanılmaktadır. Entropi değişkenlerin belirsizlik değerini veren istatistiksel bir yaklaşımdır (bilgi eksikliği, lack of information). $H(X)$, x değişkeni ile ilgili belirsizlik değerini verir, $H(X|Y)$ değeri ise y değişkeni bilindiğinde x

değişkenine ait belirsiz değerini verecektir. Bu iki değer birbirlerine eşit ise y değişkeni verildiğinde x değişkeni hakkında herhangi bir bilgi kazancı olmadığı için bu iki rassal değişken birbirlerinden bağımsız değişkenler olarak adlandırılır. Denklem 2.1'de x ve y değişkenlerinin tüm örneklerde aldıkları değerler hesaplanarak iki değişken arasındaki ortak bilgi değeri belirlenir.

C3NET algoritmasının ikinci adımı genler arasındaki ilişkileri budamak için kullanılan eşik değeri belirleme adımıdır. Buna göre değişkenler arasındaki ortak bilgi değerleri eşik değerinin altında kalırsa genler arasındaki ilişkiler GBA çıkarımında kullanılmamak üzere ilişki listesinden çıkarılmaktadır. Eşik değerinin belirlenmesinde tekrar örnekleme (resampling) metodu kullanılmaktadır. Genler arasındaki ilişkileri gösteren ortak bilgi matrisi n defa rasgele karıştırıldıktan (shuffle) sonra tüm gen- gen arasındaki ortak bilgi değerleri bir dizide küçükten büyüğe sıralanır ve kullanıcı tarafından girilen α değerine göre (dizinin uzunluğu * $(1 - \alpha)$) eşik değeri belirlenir. Örneğin 20 tane gen- gen ilişkisini tutan bir ortak bilgi matrisi 4 kere (orijinal hali kullanılarak toplamda beş matris) tekrar örnekleme ile tüm verileri karıştırılır ve toplamda 100 tane gen- gen arasındaki ilişki bir dizide tutularak küçükten büyüğe sıralanır. α değeri bu örnekte 0.1 alındığında varsayıldığında $100 * (1 - 0.1) = 90$ olarak belirlenir ve 90. sıradaki (en yüksek 90. ortak bilgi değeri) gen- gen arasındaki ortak bilgi değeri eşik değeri olarak belirlenir. Bu değer altında ortak bilgi değerine sahip tüm gen- gen ilişkileri ilişki listesinden çıkartılır ve GBA oluşturulması için geri kalan eşik değerinin üzerinde ortak bilgi değerine sahip gen-gen ilişkileri kullanılır.

C3NET algoritmasının üçüncü ve son adımında ise bir genin eşik değeri budamasından sonraki tüm ilişkiler içerisinde en yüksek ortak bilgi değerine sahip genin belirlenmesidir. C3NET algoritmasına göre bir genin sadece bir genle ilişkisi olabilir. Şekil 2.1'de küçük bir örnekle C3NET'in tüm adımları anlatılmıştır.



Şekil 2. 1 C3NET algoritması gen ağı çıkarım örneği

Şekil 2.1'deki örnekte ilk adımda genler arasındaki ortak bilgi değerleri çıkartılmıştır. İkinci adımda belirli bir eşik değerinin altında kalan gen- gen ilişkileri GBA oluşturma için kullanılacak gen ilişki listesinden çıkartılmıştır. Son aşamada ise ikinci aşamada diğer genlerle ilişkisi belirlenen her genin sadece en yüksek ortak bilgi değerine sahip gen ile ilişkisi GBA oluşumunda kullanılacak gen ilişki listesinde tutulmuş ve diğer gen ilişkileri elenmiştir. Çalışmamızda C3NET kullanmamızın en büyük nedeni farklı çalışmalarda diğer GAÇ algoritmalarına göre daha hızlı olduğunun belirtilmesidir [60] . C3NET algoritmasını uygulamak için *C3NET* R paketinden [60] yararlanılmıştır. R paketine ait daha önceden tanımlanmış (default) parametreler kullanılarak C3NET tabanlı GBA çıkarımı çalışmamızda gerçekleştirilmiştir.

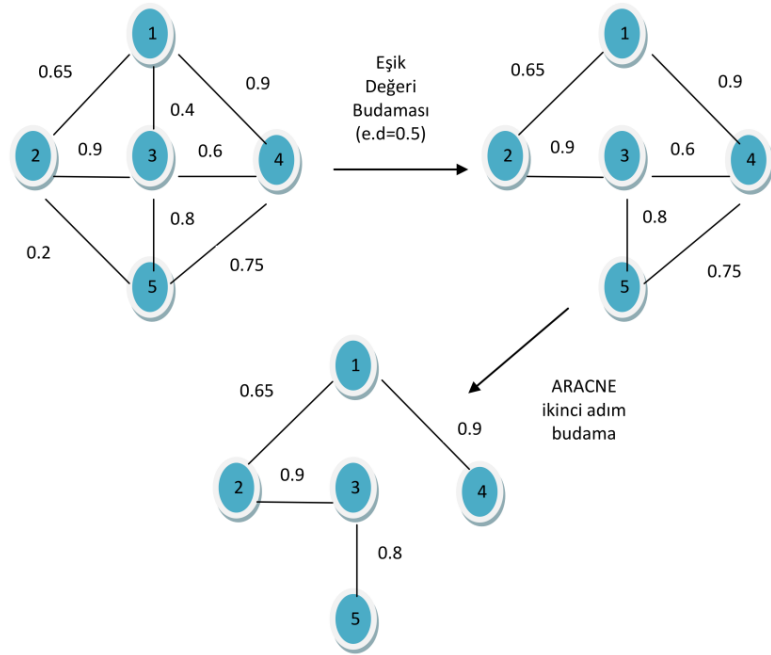
2.2 ARACNE (Accurate Cellular Networks)

ARACNE, GBA ve GDA çıkarımı çalışmalarında oldukça sık kullanılan bir gen ağı çıkarım algoritmasıdır. Bilgi teorisi yaklaşımı algoritmalarından olan ARACNE algoritması C3NET algoritması gibi üç temel adımdan oluşmaktadır. C3NET algoritmasının ilk iki adımı olan genler arasında ortak bilgi değerlerinin hesaplanması ve eşik değerine göre budama işlemi ARACNE algoritması tarafından da gerçekleştirilir. ARACNE algoritması üçüncü

adımda dolaylı (indirect) ilişkileri gen-gen ilişki listesinden eleyerek GBA çıkarımını gerçekleştirmektedir [61] .

$$MI(X, Z) \leq \argmin \{MI(X, Y), MI(Y, Z)\} \quad (2.2)$$

X geni ile Z geni arasındaki ilişkinin dolaylı bir ilişki olup olmadığının kontrolü için denklem 2.2'deki hesaplamadan yararlanılmıştır. ARACNE, X ve Z geni arasındaki ilişkinin iki genin Y geni ile olan ilişkisi nedeniyle meydana gelip gelmediğinin kontrolünü yapmaktadır. X - Y ve Y - Z genleri arasındaki iki ortak bilgi değerlerinden en küçük olanı, X - Z genleri arasındaki ortak bilgi değerinden büyükse X ile Z arasındaki ilişki dolaylı bir ilişki olarak belirlenir ve GBA çıkarımı için kullanılan gen ilişki listesinden çıkartılır. Şekil 2.2'de C3NET örneğinde verdiğimiz örnek üzerinden ARACNE algoritmasının adımları gösterilmiştir.



Şekil 2. 2 ARACNE algoritması gen ağı çıkarım örneği

Buna göre C3NET'teki ilk iki adım ARACNE algoritması içinde geçerlidir. Ortak bilgi değerleri çıkarılır ve bir eşik değerine göre ilişki budama işlemi gerçekleştirilir. Son budama adımında ise dolaylı ilişkiler elenmiştir. Buna göre eşik değeri budamasından sonra gen 3, 4 ve 5 arasında gen- gen ilişkileri olduğu gözlemlenmiştir.

- $MI(3,4) \leq \text{argmin} \{MI(3,5), MI(4,5)\}$; $0.6 \leq \text{argmin} \{0.8, 0.75\}$ doğruluğu sağlandığı için gen 3 ile gen 4 arasındaki ilişki dolaylı ilişki olarak belirlenmiştir.
- $MI(3,5) \leq \text{argmin} \{MI(3,4), MI(4,5)\}$; $0.8 \leq \text{argmin} \{0.6, 0.75\}$ doğruluğu sağlanmamaktadır. Bu nedenle gen 3 ile gen 5 arasındaki ilişki GBA çıkarımında kullanılacaktır.
- $MI(4,5) \leq \text{argmin} \{MI(3,4), MI(3,5)\}$; $0.75 \leq \text{argmin} \{0.6, 0.8\}$ doğruluğu sağlandığı için gen 4 ile gen 5 arasındaki ilişki dolaylı ilişki olarak belirlenmiştir.

Buna göre sadece gen 3 ve gen 5 arasındaki ilişki gen ağı çıkarımında kullanılmıştır. Örnekte buna benzer başka üçlü ilişki olmadığı için sadece bu üçlüye ait dolaylı ilişkiler çıkartılarak GBA oluşturulmuştur.

ARACNE algoritmasının gen ağı çıkarım çalışmalarında farklı metriklere göre yapılan değerlendirmelerde en başarılı GAÇ algoritmalarından biri olduğu gözlemlenmiştir [61], [62]. ARACNE algoritması uygulaması için *minet* R paketinden yararlanılmıştır [63].

2.3 WGCNA (Weighted Gene Co-expression Network Analysis)

WGCNA [64] özellikle birbirleriyle ilişkili genlerin ve bu genlerin oluşturduğu modüllerin keşfi için kullanılan bir algoritmadır. Bilgi tabanlı bir algoritma olan WGCNA genler arasında ilişki belirlenmesi için ortak bilgi değeri yerine (C3NET ve ARACNE'nin kullandığı ilişki ölçümü) Pearson İlişki Katsayısı yaklaşımını kullanmaktadır.

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}} \quad (2.3)$$

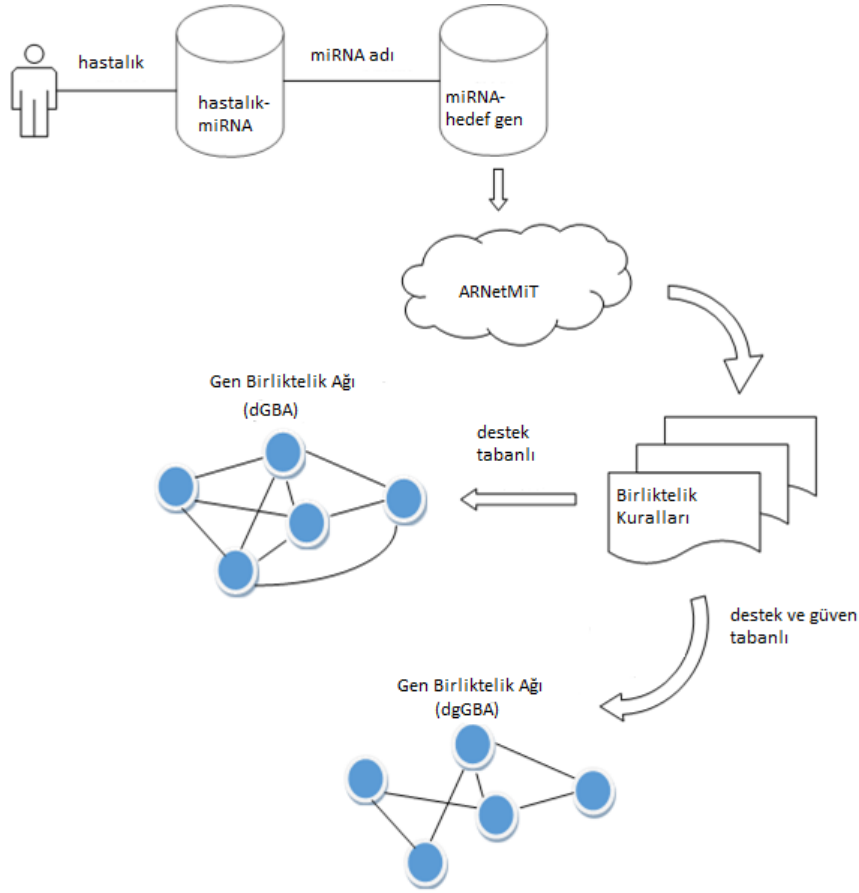
Denklem 2.3'te r_{xy} , x ve y genlerinin arasındaki Pearson korelasyon değerini veren değerdir. Buna göre n örnek sayısı, x geninin tüm örneklerdeki ortalama değeri $\bar{x}$, y geninin tüm örneklerdeki ortalama değeri $\bar{y}$ ile denklemde gösterilmiştir.

WGCNA, genler arasındaki ilişki katsayı değerlerini belirledikten sonra eşik değerini bulmak için belirleme katsayısı (R^2 , coefficient of determination) istatistiğini kullanır. Belirleme katsayısını eşik değeri olarak kullanarak ölçeksiz ağ özelliğini başarılı şekilde sağlayan GBA'lar elde edilmiş olur. GBA'ların istenen belirleme katsayısını sağlaması için gereken gen-gen arasındaki korelasyon eşik değeri WGCNA paketi tarafından

kullanıcılara katı eşik değeri (hard threshold) olarak sunulmaktadır. WGCNA uygulaması için çalışmamızda *WGCNA* R paketi kullanılmıştır [64].

miRNA- HEDEF GEN VERİLERİNDE BİRLİKTELİK KURALI TABANLI YAKLAŞIMLA GBA ELDE EDİLMESİ (ARNetMiT)

Mikrodizin gen ifadesi, RNA- Seq verisi ve miRNA- hedef gen verileri üzerinde GBA çıkarımında kullanılabilen C3NET, ARACNE ve WGCNA algoritmaları Bölüm 2'de anlatılmıştır. Tez kapsamında literatürde daha önce GBA çıkarımı için hiç kullanılmamış hash tabanlı birliktelik kuralı algoritması tasarlanmış ve miRNA- hedef gen verileri üzerinde GBA çıkarımı için kullanılmıştır [65]. Birliktelik kuralı algoritması tabanlı miRNA- hedef gen ilişkilerine ait gen ağı (Association Rules Based Gene Network of miRNA Targets, ARNetMiT) tez kapsamında bir R yazılım paketi haline getirilmiştir [65]. Şekil 3.1'de ARNetMiT yaklaşımının genel bakış açısı sunulmuştur.



Şekil 3. 1 ARNetMiT yaklaşımının genel bakış açısı

ARNetMiT yaklaşımı hastalık- miRNA ve miRNA- hedef gen verilerini kullanarak hedef hastalıkla ilişkili GBA çıkarımı gerçeklemektedir. Hastalık- miRNA verisi hastalıklarla ilişkilendirilmiş miRNA'ları barındıran veri kümesidir. Bağlandıkları genin ifade değerini değiştiren miRNA'lar ve miRNA'ların bağlandıkları genlere ait ilişkiler ise miRNA- hedef gen veri kümesinde bulunmaktadır. Çalışmamızda hastalık- miRNA hedef gen ilişkileri için *miR2Disease*[66] veri kümesi kullanılmıştır. *miR2Disease* veri kümesinde tekrarsız 2429 tane hastalık- miRNA ilişkisi bulunmaktadır. Bu veriler 477 miRNA ve 152 hastalık arasındaki ilişkileri içermektedir. miRNA- hedef gen ilişkileri iki farklı veri kümesi kullanılarak elde edilmiştir. Bunlardan ilki 322,389 deneysel olarak belirlenmiş tekrarsız ilişkiyi içeren *miRTarBase* [67], [68] bir diğeri ise 637 tane tekrarsız miRNA- hedef gen ilişkisini içeren *miR2Disease* [66] veri kümeleridir. Toplamda iki veri kümesinden 2801 miRNA ve 15,059 hedef gen arasında tekrarsız 322,904 miRNA- hedef gen ilişkisi elde edilmiştir.

Çalışmamızda kullanıcı tarafından belirlenen herhangi bir hastalıkla (152 hastalık listesinde yer alan hastalıklardan biri) ilişkili GBA'lar ARNetMiT ile elde edilmektedir. Kullanıcı bir hastalık seçtikten sonra o hastalıkla ilişkili miRNA'lar belirlenir ve belirlenen miRNA'ların ilişkili olduğu hedef genler kullanılarak bu veriler üzerinde gen ağı çıkarımı gerçekleştirilmiş olur. Hastalık- miRNA ve miRNA- hedef gen verileri ortak olarak her iki veri kümesinde de yer alan miRNA'lar kullanılarak birbirleriyle ilişkilendirilmektedir.

ARNetMiT varsayımında iki genin gen ifade değeri çok sayıda ortak miRNA tarafından değiştiriliyorsa bu genler ilişkili genlerdir (ortak ifade edilen genler, co-expressed genes) ve hücrede benzer biyolojik süreçlerde, fonksiyonlarda ve/veya bölümlerde yer alır. Buna göre iki geni etkileyen ortak miRNA sayısının fazla olması ilgili genlerin ilişki değerlerinin kuvvetli olduğunu gösterir. Genler arasındaki gizli ilişkilerin keşfi için ARNetMiT, hash tabanlı birliktelik kuralı algoritması kullanmaktadır. Bu algoritma tasarımı genlerin ifade değerlerini düzenleyen miRNA'lar işlem (transaction), miRNA'ların bağladıkları hedef genler ise parçalar (items) olarak belirlenmiştir. Farklı farklı işlemlerde yer alan ortak parçalara ait ilişki keşfi finansal işlemler, hava tahmini, coğrafi bilgi sistemleri, bankacılık işlemleri [69] gibi farklı alanlarda uzun süredir kullanılmakla birlikte aynı yaklaşımla miRNA- hedef gen verileri üzerinden GBA çıkarımı daha önce literatürde kullanılmamıştır.

Birliktelik kuralı tabanlı algoritmalar özellikle büyük verilerde gizli örüntülerin keşfinde sıklıkla kullanılmaktadır. Apriori [70] ve Eclat [71] farklı alanlarda sıklıkla kullanılan popüler birliktelik kuralı tabanlı algoritmalarıdır. Apriori ve Eclat algoritmaları işlemler içerisinde belirli eşik değerinin üzerinde sıklıkla geçen parça kümelerini (frequent item sets) belirlemeye çalışır.

- İşlem (T), parça (I) olarak sembolize edildiğinde, herhangi bir k işlemi çok sayıda parçayı kapsayabilmektedir : $T_k = \{I_1, I_2, I_3, \dots, I_n\}$ ve $T = \{T_1, T_2, T_3, \dots, T_k, \dots, T_z\}$.
- ARNetMiT miRNA'ları işlem, hedef genleri ise parça olarak belirleyerek parçalar arasındaki (yani genler arasındaki) gizli ilişkileri belirlemeye çalışmaktadır.

Çizelge 3. 1 Örnek miRNA- hedef gen ilişki tablosu

İşlem (miRNA)	Parça (hedef gen)
miRNA ₁	Gen _A , Gen _C , Gen _D
miRNA ₂	Gen _B , Gen _C , Gen _E
miRNA ₃	Gen _A , Gen _B , Gen _C , Gen _E
miRNA ₄	Gen _B , Gen _E

Çizelge 3.1'de örnek bir miRNA- hedef gen ilişki tablosu gösterilmiştir. Bu tabloya göre dört tane miRNA ve her bir miRNA'nın ilişkili olduğu hedef genler verilmiştir. Her bir işlem, parça kümesinin bir alt kümesidir. Birliktelik kuralı algoritmalarında temel kullanılan kavramlar aşağıda kısaca özetlenmiştir.

- Parça Kümesi (Itemset): Bir veya birden fazla parçadan oluşan parça koleksiyonudur. Örneğin {Gen_B, Gen_E, Gen_C }.
- Destek Sayısı (Support Count): Her bir parça kümesinin geçme sıklığını (frekansını) veren bilgidir. Örneğin, Çizelge 3.1'e göre { Gen_B, Gen_E } için bu değer üçtür. İki genin ifade değeri üç ortak miRNA tarafından düzelenmiştir (regüle edilmiştir).
- Destek (Support): Her bir parça kümesinin geçme sıklığının tüm işlem sayısına oranıdır. { Gen_B, Gen_E } örneğine göre bu değer 3/4'tür. Buradaki 3 { Gen_B, Gen_E}'nin destek sayısını vermekte, 4 ise tüm işlem sayısını göstermektedir.
- Sık Geçen Parça Kümesi (Frequent Itemset): Belirli bir minimum destek değerinden yüksek değerde destek değeri olan parça kümelerine denir.
- Birliktelik Kuralı (Association Rule): Birliktelik kuralı öncül (antecedent) ve takipçi (consequent) diye iki parçadan oluşur. Öncül veri kümesinden olan bir parça, takipçi ise öncülle ilişkili olan bir diğer parçayı göstermektedir. Öncül X,

takipçi Y olarak alındığında aralarındaki ilişki $X \rightarrow Y$ notasyonu ile gösterilir. $X \cap Y = \{\}$ ve $\{X, Y\} \subseteq I$.

- Kural Değerlendirme Metrikleri: Destek (support), güven (confidence), kaldırma (lift) değeri, Gini katsayısı, J- ölçümü gibi farklı metriklerle birliktelik kuralları değerlendirilir. Çalışmamızda en çok kullanılan destek ve güven metrikleri kullanılarak birliktelik kuralları değerlendirmesi yapılmıştır.

İlk olarak miRNA- hedef gen veri kümesinde belirli bir eşik değeri üzerindeki miRNA tarafından regüle edilen genler seçilir.

$$destek(x) = \frac{n(x)}{N} \quad (3.1)$$

$$\forall x \in G : \{destek(x) > minDD_1 \rightarrow x \in F_1\}$$

Denklem 3.1'de bir x geninin kaç tane miRNA tarafından regüle edildiği $n(x)$ fonksiyonuyla, toplam miRNA sayısı ise N ile gösterilmiştir. Buna göre bir gene ait destek değeri $n(x)/N$ olarak belirlenmiştir. Genin destek değeri, belirlenen minimum destek değerinden ($minDD_1$) az ise, bu gen GBA oluşumunda kullanılacak gen listesinden çıkartılır. Eşik değerinin üzerindeki GBA oluşturulması için kullanılacak genler Denklem 3.1'de gösterildiği gibi sıklıkla geçen ilk parçalar kümesi F_1 'e eklenir.

İlk aşamada belirlenen sıklıkla regüle edilen genler arasındaki gizli ilişki bilgisi ortaya çıkarılarak ortak ifade edilen genler (co-expressed genes) belirlenmiş olur. Genler arasındaki ilişkiler Denklem 3.2'de gösterildiği gibi birlikte regüle edildikleri ortak miRNA sayısına göre belirlenmiştir.

$$\forall \{x, y\} \in F_1 : \left\{ \frac{n(x, y)}{N} > minDD_2 \rightarrow \{x, y\} \in F_d \right\} \quad (3.2)$$

Denklem 3.2'de gen x ve gen y 'yi ortak regülen eden miRNA sayısı $n(x, y)$ şeklinde pay olarak, tüm miRNA sayısı ise N payda olarak gösterilmiştir. Elde edilen değer ikinci bir destek değerinden büyük ise bu genler ortak ifade edilen genler olarak belirlenir ve GBA oluşumu için kullanılır. GBA oluşumunda kullanılacak ortak ifade edilen genlerin arasındaki ilişkiler F_d kümesinde tutulmaktadır. İki gene ait destek değerinin yüksek olması, iki geni ortak olarak regüle eden miRNA sayısının fazla olduğunu gösterir ve bu genler varsayımımıza göre kuvvetli ilişkisi olan ortak ifade edilen genlerdir.

ARNetMiT'te minimum destek değerleri kullanıcı tarafından belirlenen zorunlu değerlerdir. Buna göre $\min DD_1$ ve $\min DD_2$ değerleri örüntülerin dağılımlarından elde edilen kutu grafiğindeki ilk çeyrek, medyan, üçüncü çeyrek değerleri olarak *ARNetMiT* R paketi tarafından kullanıcılara seçenek olarak sunulmuştur. Bununla beraber örüntü değerleri dağılımındaki maksimum sayı kullanıcı tarafından belirlenen bir sayı (rank yani sıralama değeri) ile çarpılarak minimum destek değeri belirlenebilir (sıralama parametresi kullanımı). Örneğin, 50 miRNA içeren bir miRNA- hedef gen veri kümesinde bir genin maksimum 20 miRNA tarafından regüle edildiği varsayıldığında ve sıralama parametresi 0.5 olarak kullanıcı tarafından girildiğinde minimum destek değeri $20 \times 0.5 = 10$ olarak belirlenir. Buna göre 10'dan fazla sayıda miRNA tarafından regüle edilen genler F_1 gen listesine eklenir. Aynı şekilde, ortak ifade edilen genleri temsil eden F_d ilişki listesi de aynı filtreleme yöntemiyle elde edilir.

ARNetMiT R paketi, sadece destek değerine bağlı GBA çıkarımının yanı sıra destek ve güven değerini birlikte kullanılmasıyla elde edilen GBA çıkarımına da olanak sağlamaktadır. Güven değeri destek değeri gibi zayıf gen- gen ilişkilerini elemektedir. Güven değeri ortak miRNA'lar tarafından regüle edilen genler arasındaki ilişkileri koşullu bağımlılık değerine göre hesaplar. Denklem 3.3'te güven değeri belirleme denklemi gösterilmiştir.

$$güven(x \Rightarrow y) = P(y | x) = \frac{n(x, y)}{n(x)} \quad (3.3)$$

$$\forall x, y \in F_1 \wedge \forall \{x, y\} \in F_d : \{güven(x \Rightarrow y) > \min G \rightarrow \{x, y\} \in F_g\}$$

Denklem 3.3'e göre iki gen arasındaki güven değerlerinin hesaplanması için genlerin F_1 listesinde olması ve iki genin de ortak ifade edilen genlerin gösterildiği F_d listesinde yer alması gerekir. Bir X geni verildiğinde, Y geninin ortak miRNA'lar tarafından regüle edilme olasılığı $P(Y/X)$ koşullu olasılığı olarak belirlenmektedir. X ve Y genini ortak regüle eden miRNA sayısı $n(x, y)$ olarak pay, X genini regüle eden miRNA sayısı ise payda olarak denklemde gösterilmiştir. Elde edilen güven değeri minimum güven değerinden ($\min G$) büyük ise elde edilen gen ilişkisi GBA oluşumunda kullanılmaktadır. Güven değerinin bire yaklaşması X ve Y genlerinin kuvvetli birer ortak ifade gen çifti olduğunu gösterir. F_g listesinde destek ve güven değerine göre belirlenmiş ortak ifade edilen gen çiftleri bulunmaktadır. Çalışmamızda F_d listesi kullanılarak oluşturulan GBA'lar destek

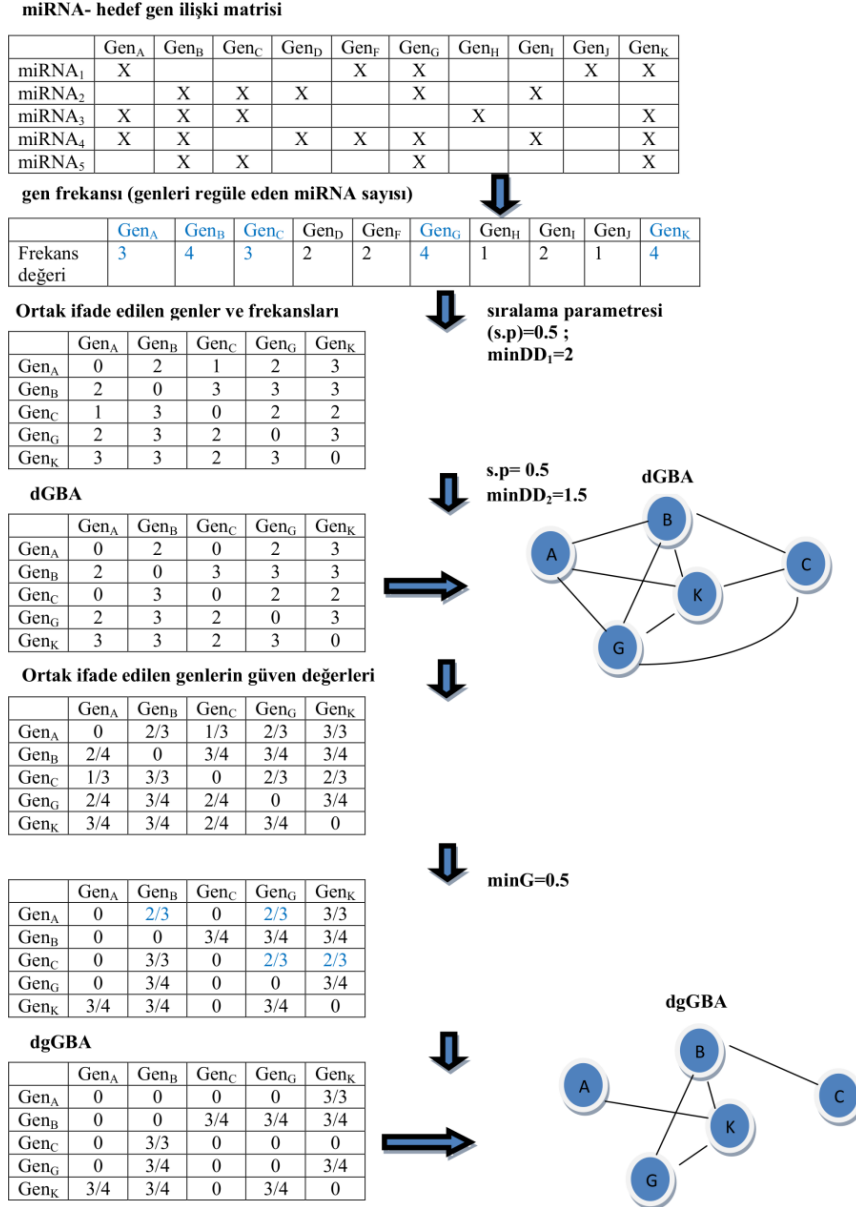
değeri tabanlı GBA'lar (dGBA), F_g listesi kullanılarak oluşturulan GBA'lar ise destek-güven değeri tabanlı GBA'lar (dgGBA) olarak adlandırılmıştır.

ARNetMiT'in 322,904 tane miRNA- hedef gen verisi içerisinde gizli gen-gen ilişkilerini çıkarması hesaplama süreci açısından problematiktir. Bu problemi aşmak için uzun zaman alıcı ARNetMiT genler ve her geni regüle eden miRNA'ların indislerini bir hash veri yapısında tutar. Genler ve bu genleri regüle eden miRNA'lara ulaşım hash tablo yapısıyla çok daha hızlı yapılmaktadır. Sözde kodta yer verildiği gibi hash tablosunda genler (gen isimleri) anahtar (key), genleri hedef alan miRNA'lar ise değerler (values) olarak kullanılmıştır. Genleri anahtar olarak kullanarak genlerin ilişkili olduğu miRNA'ların hash tablosunda hızlı ulaşımı sağlanır. Hash fonksiyonları hızlı erişim için anahtar kelimeyi (ARNetMiT'te gen isimleri) aldıktan sonra sabit bir tablo için yeni bir indeks değeri oluşturur. Özetleme işleminde (hashing) anahtar kelimeye ait karakterlerin ASCII karakter karşılığını kullanarak "Anahtar mod N" dönüşümü ile her bir anahtarla ilgili hash indeksi oluşturulması en sık kullanılan özetleme fonksiyonu yaklaşımlarındandır. Burada gen isimlerinde yer alan karakterlerin ASCII karşılıkları toplanarak asal bir sayıya (örneğin 37) göre modu alınır ve hash indeks değeri hesaplanmış olur. Çalışmamızda *hash* R paketi [72] ve bu paketdeki hash fonksiyonu ile özetleme (hashing) işlemi gerçekleştirilmiştir.

```
BEGIN
genes= unique( diseaseRelatedmiRNA[,2]) ; // ilk sütun hastalık, ikinci sütun miRNA'ları gösteren tablo oluşturma
h= hash(); // hash yapısını oluştur
Declare frequencyFirstPattern[length(genes)];
FOR i= 1 to length (genes)
    frequencyFirstPattern [i] = getFrequency(genes[i]) ;
    set (h, genes[i], frequencyFirstPattern [i]); // genler ve genlerin sıklık değerleri hash yapısında tutulur
END FOR
support_val1 = max(frequencyFirstPattern)* rankvalue ;
frequentGenes= eliminate (frequencyFirstPattern, support_val1); // minDD1(support_val1) değerinden fazla sayıda miRNA
    tarafından regüle edilen genlerin oluşturduğu gen listesini (F1) elde etme
FOR j= 1 to length (frequentGenes)
    miRNAIndices[j]=get(h,frequentGenes[j]); // F1'de yer alan j genini regüle eden miRNA'ların indisleri elde edilir.
FOR k= 1 to length (frequentGenes)
    IF k != j
        miRNAIndices[k]=get(h, frequentGenes[k]); // F1'de yer alan k genini regüle eden miRNA'ların indisleri elde edilir.
        frequencySecondPattern[j,k]= intersect(miRNAIndices[j], miRNAIndices[k]); // k ve j genlerini ortak olarak regüle
            eden miRNA'ların beirlenmesi
    END IF
END FOR
support_val2 = max(frequencySecondPattern)* rankvalue ;
coexpressedGenes= eliminate(frequencySecondPattern, support_val2); //minDD2(support_val2) değerinden fazla sayıda
    ortak miRNA tarafından regüle edilen gen çiftlerinin oluşturduğu gen ilişki listesi (F2) elde etme. dGBA'ların oluşturulması.
confidenceTable= getFrequency(coexpressedGenes) / getFrequency(frequentGenes); //güven değerlerinin oluşturulması
directedCoexpressedGenes= eliminate (confidenceTable, confidence_val); // minimum güven değerine göre eleme
    işleminin yapılması ve dgGBA'ların elde edilmesi.
END
```

Şekil 3. 2 ARNetMiT yaklaşımının sözde kodu(Pseudocode)

Detaylarını anlattığımız ARNetMiT yaklaşımının sözde kodu Şekil 3.2'de gösterilmiştir. Şekil 3.2'ye göre genler ve genleri regüle eden miRNA indisleri (indis sayısına göre sıklık değeri veya frekans değeri belirlenir) hash tablosuna kaydedilmiştir. Daha sonra genlerin sıklık değerlerine göre ilk minimum destek değerinden az sayıda miRNA tarafından regüle edilen genler GBA oluşumu için kullanılacak gen listesinden çıkarılmıştır.



Şekil 3.3 ARNetMiT temel adımlarının bir örnek üzerinde gösterimi

Eleme aşamasından sonra kalan tüm genlerin miRNA indis değerlerine hash yapısındaki anahtar ile ulaşım sağlanarak genleri ortak regüle eden miRNA'lar bulunmuştur. Ortak miRNA'ların sayısı ikinci minimum destek değerinden büyük ise dGBA için gen- gen listeleri oluşturulmuş olur. Ortak ifade edilen genlerin sıklık değerlerinin, genlerin sıklık değerine bölünmesiyle güven değerleri elde edilir. Bir gen- gen ilişkisine ait güven değeri minimum güven değerinden büyükse sahip gen- gen ilişkisi dgGBA oluşumu için kullanılır. Şekil 3.3'te ARNetMIT'in temel adımları bir örnek üzerinde gösterilmiştir. Şekil 3.3'te kullanıcı hastalık seçimi yaptıktan sonra hastalıkla ilişkili miRNA'lar ve bu miRNA'ların regüle ettiği genler miRNA- hedef gen ilişki matrisinde gösterilmiştir. En fazla bir gen dört miRNA tarafından regüle edilmiş ve sıralama parametresi örnekte 0.5 olarak verilmiştir. İlk minimum destek değeri $4 \times 0.5 = 2$ olarak belirlenmiştir. İki ve ikinin altında miRNA tarafından regüle edilen genler GBA oluşumu için kullanılacak gen listesinden çıkartılmıştır. Örnekte on genden oluşan gen listesi sayısı ilk budama işleminden sonra beşe düşmüştür. İkinci adımda sıkça regüle edilen beş geni ortak regüle eden miRNA'lar belirlenmiş ve frekans değerleri elde edilmiştir. Bu değerlerle Şekil 3.3'de yer alan birlikte ortak ifade edilen genler ve frekansları tablosunda gösterilmiştir. Bu tablodaki değerler genleri ortak regüle eden miRNA sayılarını göstermektedir. İlgili tabloda gen çiftini regüle eden değer maksimum 3 olarak gösterilmiştir (örneğin $Gen_A - Gen_K$). Sıralama parametresi kullanıcı tarafından 0.5 olarak girildiği için $minDD_2$ değeri 1.5 olarak belirlenir. Buna göre 1.5'dan daha fazla sayıda ortak miRNA tarafından regüle edilen genler ortak ifade edilen genler olarak belirlenmiştir. İkinci destek değeri filtrelemesinden sonra Şekil 3.3'te gösterildiği gibi dGBA'lar elde edilmiştir. Ortak ifade edilen genlerin frekans değerlerinin, genlerin frekans değerlerine bölümüyle ortak ifade edilen genler için güven değerleri belirlenmiş olur. Bu değerler ortak ifade edilen genlerin güven değeri tablosunda gösterilmiştir. Kullanıcı tarafından girilen minimum güven değerinden ($minG$) küçük güven değerine sahip gen- gen ilişkileri elenir ve dgGBA elde edilmiş olur.

ARNetMIT R paketi farklı gen ağı çıkarım algoritmalarının miRNA- hedef gen ilişkileri kullanarak GBA çıkarımı gerçeklemesi için miRNA- hedef gen ilişkilerini ikili tabloya (boolean tablo) dönüşümünü de gerçekleştirmektedir. *buildmiRNATargetTable* fonksiyonunu kullanarak ARACNE, C3NET, WGCNA gibi ağ çıkarım algoritmaları miRNA-

hedef genler arasındaki ilişkileri 0-1 olarak gösteren tablolara erişerek GBA çıkarımı yapabilmektedir.

ARNetMiT R paketi, *hash* [72], *GAnet* [73] ve *RedeR*[74] R paketlerini kullanmıştır. *hash* R paketi özetleme (hashing) işleminde, *GAnet* R paketi basit ilişki operatörlerinin kullanımında, *RedeR* R paketi ise elde edilen GBA'ların görselleştirilmesinde kullanılmıştır. *ARNetMiT* R paketi ve pakete ait dokümantasyona <https://sites.google.com/site/arnetmit> web adresinden ulaşılabilir.

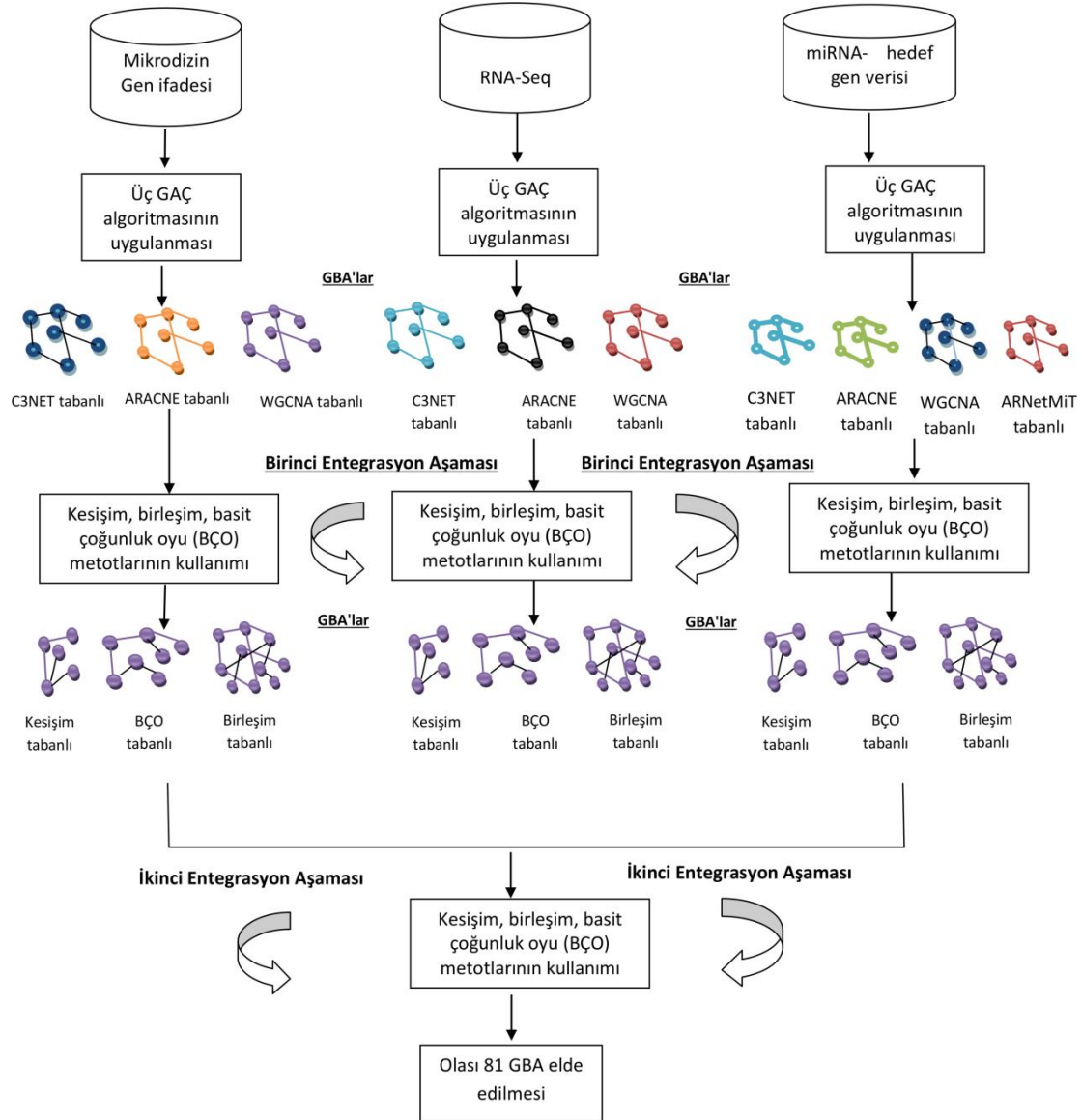
İKİ KATMANLI GBA ENTEGRASYON MİMARİSİ

Farklı yaklaşımların, farklı biyolojik veri kümelerinin bir arada kullanılarak elde edilen sonuçların entegrasyonu araştırmacıların biyolojik sistemleri daha doğruluklu ve kapsamlı bir şekilde anlamasını sağlamaktadır. Tez kapsamında gen birliktelik ağlarının entegrasyonu için farklı biyolojik veri kümeleri ve gen ağı çıkarım algoritmaları kullanılmıştır. Şekil 4.1'de iki katmanlı GBA entegrasyon mimarisinin genel adımları ve katmanları verilmiştir.

Şekil 4.1'de gösterildiği üzere iki katmanlı GBA entegrasyon mimarisinde mikrodizin gen ifadesi, RNA- Seq ve miRNA- hedef gen verileri kullanılmıştır. GBA çıkarımında her üç veri kümesi üzerinde ARACNE, C3NET ve WGCNA gen ağı çıkarım algoritmaları uygulanarak GBA çıkarımı başlangıç seviyesinde (initial phase) gerçekleştirilmiştir. Başlangıç seviyesinde her bir veri kümesine ait 3 farklı GBA çıkarımı 3 GAÇ algoritması kullanılarak oluşturulmuştur. Aynı veri kümesi üzerinde farklı GAÇ algoritmaları uygulanarak GAÇ algoritmalarına ait kestirim farklılıkları kullanılması hedeflenmiştir. Bu sayede aynı veri kümesinden çeşitli gen- gen ilişkileri taşıyan 3 GBA çıkarılmış olur. Başlangıç seviyesinde 3 farklı veri kümesine ait 9 GBA elde edilmiştir.

Tez kapsamında geliştirdiğimiz ARNetMiT yaklaşımını kullanarak aynı zamanda miRNA- hedef gen verileri üzerinde dGBA ve dgGBA'lar elde edilmiştir. Özetle başlangıç seviyesinde üç gen ağı çıkarım algoritması mikrodizin gen ifadesi ve RNA- Seq verileri üzerinde üçer farklı GBA çıkarımı gerçekleştirilmiştir. MiRNA- hedef gen veri kümesinde ise bu sayı ARNetMiT tabanlı GBA'lar da eklendiğinde beşe çıkmıştır. Çalışmamızda

ARNetMiT tabanlı algoritmaları kullanmadan oluşturulan (miRNA- hedef gen verileri üzerinde sadece ARACNE, C3NET ve WGCNA algoritmaları kullanımıyla oluşturulmuş) GBA entegrasyon sonuçları ile ARNetMiT tabanlı GBA'lar kullanılarak oluşturulmuş GBA'ların entegrasyonu karşılaştırılmıştır. Buradaki temel amaç ARNetMiT'in entegrasyon çalışmasına katkısını gözlemlemektir.



Şekil 4. 1 İki katmanlı GBA entegrasyonu mimarisi

Her bir veri kümesi kullanılarak başlangıç seviyesinde elde edilen GBA'ların entegrasyonu için üç basit metot kullanılmıştır. Bu metotlar kesişim, basit çoğunluk oyu (BÇO) ve birleşim metotlarıdır. Buna göre elde edilen GBA'ların en az birinde yer alan

gen- gen ilişkilerinin kullanılmasıyla birleşim tabanlı GBA (BTGBA) elde edilmektedir. Tüm GBA'larda ortak olarak belirlenen gen- gen ilişkilerinin kullanılmasıyla kesişim tabanlı GBA'lar (KTGBA) elde edilmektedir. BÇO metodu, algoritma sayısının yarısından fazla sayıda GBA tarafından tahmin edilen gen- gen ilişkilerini kullanarak basit çoğunluk oyu tabanlı GBA (BÇOGBA) çıkarımı yapmaktadır. Mikrodizin gen ifadesi ve RNA- Seq veri kümesinde 3 gen ağı çıkarım kullanıldığı için BÇOGBA oluşturmak için en az 2 gen ağı çıkarım algoritması tarafından belirlenen gen- gen ilişkileri kullanılır. Bu değer miRNA- hedef gen veri kümesinde üçtür. Beş gen ağı çıkarım algoritması içerisinde en az üç ($> 5/2$ şartı sağlanmalı) tane GAÇ algoritması tarafından tahmin edilen gen- gen ilişkileri BÇOGBA oluşturmada kullanılır.

K = gen ağı çıkarımı yapan algoritma sayısı,

$GAÇ_A$ = gen ağı çıkarım algoritması,

I = genler arası ilişki,

O = gen ilişkilerini tahmin sayısı,

N = gen listesi,

B = Boolean fonksiyonu

$$O_{x,y} = \sum_{k=1}^K B \left(GAÇ_A_k(I_{x,y}) \right), \forall x,y \in N, x \neq y \quad (4.1)$$

Denklem 4.1'de B ikili (boolean) fonksiyon olmak üzere, bir veri kümesine ait gen listesindeki tüm genler arasındaki ilişkilerin gen ağı çıkarım algoritması tarafından tahmin edilip edilmediğini kontrol eder. Bu fonksiyon 0 veya 1 döndüren bir fonksiyondur. İkili fonksiyon GAÇ algoritması birbirlerin farklı olan x ve y genleri arasında ilişki çıkarımı yapmışsa 1 değerini, çıkarım yapamamış ise 0 değerini döndürmektedir. Denklem 4.1'de $GAÇ_A$ gen ağı çıkarım algoritmalarını, I genler arasındaki ilişkileri, N gen listesini sembolize etmektedir. K gen ağı çıkarımı yapan algoritma sayısı olmak üzere $O_{x,y}$, x ve y gen ikilisinin (olası ortak ifade edilen genler) kaç gen ağı çıkarım algoritması tarafından tahmin edildiğini gösterir. Bu değer kesişim tabanlı GBA'lardaki her bir gen- gen ilişkisi için gen ağı çıkarım algoritması sayısına, birleşim tabanlı GBA'lardaki her bir gen- gen ilişkisi için en az bire, basit çoğunluk oyu

tabanlı GBA'lardaki her bir gen- gen ilişkisi için en az gen ağı çıkarım algoritmasının yarısından büyük bir doğal sayıya eşit olması gerekmektedir.

Birinci entegrasyon aşamasında amaç aynı veri kümeleri üzerinde farklı gen ağı çıkarım algoritmalarıyla elde edilmiş GBA'ların entegrasyonudur. Birinci entegrasyon aşamasında gen ağı çıkarımlarındaki algoritma tabanlı farklılıkların GBA entegrasyonunda kullanılması hedeflenmiştir. Birinci entegrasyon aşaması aynı tür veri üzerinden elde edilen GBA'ların entegrasyonuna dayalı olduğu için çalışmamızda grup içi entegrasyon olarak adlandırılmıştır. Birinci entegrasyon aşamasında farklı 3 veri kümesine ait 9 GBA çıkarımı yapılmıştır. Bu duruma göre 3 veri kümesi kullanılarak her bir veri kümesi için KTGBA, BÇOGBA ve BTGBA elde edilmiştir. Basit entegrasyon metotlarının yapısı gereği KTGBA'daki gen- gen ilişkilerinin tümü BÇOGBA'larda, BÇOGBA'lardaki gen- gen ilişkilerinin de tümü BTGBA'lar tarafından kapsamaktadır.

İkinci entegrasyon aşamasındaki temel amaç farklı veri kümelerinden elde edilen GBA'ların entegrasyonudur. İlk aşamada aynı veri kümesi üzerindeki farklı gen ağı çıkarım algoritmaları tarafından üretilen GBA'ların entegrasyonu ile algoritmaların tahmin farklılıklarından (diversity) yararlanılmıştır ve bu nedenle grup içi entegrasyon olarak tanımlanmıştır. İkinci aşamada ise farklı biyolojik veri kümelerinden elde edilen GBA entegrasyonu gerçekleştirildiği için bu aşamadaki entegrasyon gruplar arası entegrasyon olarak adlandırılmıştır. Gruplar arası entegrasyon ile farklı biyolojik veri kümelerinden elde edilen GBA'lar entegre edilmeye çalışılmaktadır. Farklı veri kümelerinden elde edilen GBA'ların içerdiği gen- gen ilişkisi çeşitliliği ikinci entegrasyon aşamasında kullanılmış ve hastalıklarla ilgili daha kapsamlı, doğruluklu GBA'ların elde edilmesi hedeflenmiştir.

Input :

OP= [kesişim, bco, birleşim] /* kullanılan metot dizisi */

GE =[KTGBA, BCOGBA, BTGBA] /* Mikrodizin gen ifadesi verileri kullanarak birinci entegrasyon aşamasından sonra elde edilen GBA'ları tutan dizi */

RS =[KTGBA, BCOGBA, BTGBA] /* RNA- Seq verisi kullanarak birinci entegrasyon aşamasından sonra elde edilen GBA'ları tutan dizi */

MR =[KTGBA, BCOGBA, BTGBA] /* miRNA- hedef gen verileri kullanarak birinci entegrasyon aşamasından sonra elde edilen GBA'ları tutan dizi */

Output:

finalGCNs /* İkinci entegrasyon aşamasından sonra elde edilen tüm GBA'ları içeren dizi */

Begin :

1. finalGCNs = array[length(OP)* length(GE)* length(RS)* length(MR)] /* create array */
2. counter= 0
3. for i=1 to length(OP)
4. for j=1 to length(GE)
5. for k=1 to length(RS)
6. for z=1 to length(MR)
7. finalGCNs[counter]= $OP_i(GE_j, RS_k, MR_z)$
8. counter =counter + 1
9. end
10. end
11. end
12. end
13. return finalGCNs

End:

Şekil 4. 2 İkinci entegrasyon aşamasına ait adımların sözde kodu

Şekil 4.2'de ikinci entegrasyon aşamasının adımlarını içeren sözde kodları gösterilmiştir. İkinci entegrasyon aşamasında birinci entegrasyon aşamasında elde edilmiş dokuz GBA'nın olası tüm kombinasyonları birleşim, basit çoğunluk oyu ve kesişim operatörleri kullanılarak entegre edilmiştir. Birinci ve ikinci entegrasyon aşamasından önce GBA'larda yer alan tekrarlı ilişkiler belirlenmiş ve tek bir ilişki gibi değerlendirilerek entegrasyon aşamasında kullanılmıştır. Şekil 4.2'de gösterildiği üzere ikinci entegrasyon aşaması iç içe dört tane döngü içermektedir. Bu döngülerin ilki üç entegrasyon

operatörüne, ikinci döngü mikrodizin gen ifadesi verileri kullanılarak birinci entegrasyon aşamasında elde edilmiş üç GBA'ya, üçüncü döngü RNA- Seq verileri kullanılarak birinci entegrasyon aşamasında elde edilmiş üç GBA'ya ve sonuncu döngü miRNA- hedef gen verileri kullanılarak birinci entegrasyon aşamasında elde edilmiş elde edilmiş üç GBA'ya erişim için kullanılmaktadır. Her bir döngünün indis değeri üç olması nedeniyle $3*3*3*3= 81$ olası GBA ikinci entegrasyon sonunda elde edilebilir. Şekil 4.2'de yer alan dört döngüye ait indis değerleri (i, j, k ve z) aşağıda verilerek ikinci entegrasyon aşamasında elde edilen GBA'lar iki farklı örnek üzerinden gösterilmiştir.

- $i=2, j=1, k=3, z=2$ olarak döngülerde belirtildiğinde $OR_2(GE_1, RS_3, MR_2)$ işlemi sonucunda elde edilecek GBA belirlenmiş olur. $OR_2(GE_1, RS_3, MR_2)$, gen ifadesi verileri üzerinde birinci aşamada elde edilmiş KTGBA, RNA- Seq verisi üzerinde birinci aşamada elde edilmiş BTGBA ve miRNA- hedef gen verileri üzerinde birinci aşamada elde edilmiş BÇOGBA'nın entegrasyonunun basit çoğunluk oyu yaklaşımıyla gerçekleştirildiği göstermektedir.
- $i=3, j=3, k=2, z=1$ olarak döngülerde belirtildiğinde $OR_3(GE_3, RS_2, MR_1)$ işlemi sonucunda elde edilecek GBA belirlenmiş olur. $OR_3(GE_3, RS_2, MR_1)$, gen ifadesi verileri üzerinde birinci aşamada elde edilmiş BTGBA, RNA- Seq verisi üzerinde birinci aşamada elde edilmiş BÇOGBA ve miRNA- hedef gen verileri üzerinde birinci aşamada elde edilmiş KTGBA'nın entegrasyonunun birleşim yaklaşımıyla gerçekleştirildiği göstermektedir.

Bu şekilde olası 81 GBA çıkarımı ikinci entegrasyon aşamasında oluştulmaktadır. Özellikle kesişim operatörünün kullanıldığı bazı entegrasyon çalışmalarında üç GBA'da ortak olarak tahmin edilen gen- gen ilişkisi olmaması nedeniyle genellikle kesişim tabanlı GBA'lar ikinci entegrasyon sürecinde çoğunlukla elde edilememiştir. Özellikle birinci entegrasyon aşamasında kesişim metoduyla oluşturulmuş ve az sayıda gen- gen ilişkisini bulunduran GBA'ların ikinci aşama entegrasyonunda kullanılmasıyla ikinci aşamada bazı GBA'lar oluşturulamamıştır.

İki katmanlı GBA entegrasyon mimarisiyle birinci entegrasyon aşamasında aynı veri kümeleri üzerinde GAÇ algoritmalarının kestirim gücü, ikinci entegrasyon aşamasında ise biyolojik veri kümelerinin çeşitliliği kullanılarak oluşturulmuş GBA'ların

entegrasyonu hedeflenmiştir. Buna göre hastalıklarla ilişkili tahmin edilen gen- gen ilişkilerinin GAÇ algoritmasına ve biyolojik veri kümesine bağımlılığını azaltarak kapsamlı ve doğruluklu GBA'ların tez kapsamında belirlenmesi amaçlanmıştır.

GBA'LARIN PERFORMANS DEĞERLENDİRME KRİTERLERİ

Literatürde GBA'ların performansının değerlendirilmesi için pek çok yaklaşım geliştirilmiştir. Yönsüz ilişkileri içeren literatür verilerinin kullanımı (protein- protein ilişkileri), tersine mühendislik yöntemleriyle elde edilmiş doğrulama verilerinin kullanımı (DREAM project), GBA'ların topolojik özelliklerini kullanarak yapılan yapısal değerlendirmeler, GBA'larda yer alan merkez genleri (hub genes) bulan ve bu genlerin hastalıklarla ilişkisinin incelenmesi, GBA'ları alt ağlara ayırarak ilgili alt ağlarda yer alan genlerin yer aldığı biyolojik süreçlerin hastalıklarla ilişkisinin incelenmesi gibi pek çok farklı GBA performans değerlendirme kriteri çalışmalarda yer almaktadır. Tez kapsamında GBA performans değerlendirmesi için ilk olarak literatür verilerini kullanarak örtüşme analizi gerçekleştirilmiştir. İkinci değerlendirme kriteri ise GBA'ların yapısal değerlendirmesine olanak sağlayan topolojik değerlendirmedir. Topolojik değerlendirmede GBA'ların ölçeksiz ağ yapısına uygunluğu belirlenmektedir. Son değerlendirme biyolojik değerlendirmedir. Biyolojik değerlendirmede GBA'larda dayanıklı ilişkiler içeren alt ağları bularak bu alt ağlardaki genlerin hastalıklara ait biyolojik süreçlerle ilişkisi incelenmiştir.

5.1 Literatür Verileri Kullanımı ile Gerçeklenen Örtüşme Analizi Değerlendirmesi

GBA'larda yer alan gen- gen ilişkileri arasında yön bilgisi bulunmamaktadır. Bu nedenle GBA'larda genler arasında düzenleyici (regulator) ve düzenlenen (regulating) ilişkisi yer almaz. Gen- gen ilişkileri gibi aralarında yön bilgisi bulundurmayan protein- protein ilişkileri (PPI, literatür verileri) GBA performanslarını değerlendirmede sıklıkla kullanılan

veri kümelerindendir [7], [73]. Genlerin ürünleri olan proteinlerin protein ağları üzerindeki ilişkileri, genlerin dolaylı ilişkilerini yansıttığı varsayımına dayanarak PPI verileri GBA'lardaki gen-gen ilişkilerinin doğrulama kümesi olarak kullanılabilir.

Çalışmamızda *ganet* R paketinde [73] yer alan 1,559,366 adet tekrarsız ve deneysel olarak doğrulanmış PPI verisi örtüşme analizinde kullanılmıştır. Örüntü tanıma, bilgi edinimi, sınıflandırma gibi alanlarda performans analizi için kullanılan kesinlik (precision) ve doğru pozitif (DP veya doğru artık) metrikleri kullanılarak örtüşme analizi gerçekleştirilmiştir. Örtüşme analizinde GBA'da yer alan bir gen- gen ilişkisi PPI veri kümesinde de yer alıyorsa bu ilişki DP olarak adlandırılır. GBA'da bulunan ama PPI veri kümesinde yer almayan ilişkiler yanlış pozitif (YP veya yanlış artı) olarak belirlenir. Bir GBA'da yer alan ve test kümesinde de yer alan ilişkilerin (DP), tüm GBA tarafından tahmin edilen ilişkilere oranı (DP + YP) kesinlik değerini vermektedir. Anma (recall) değeri ise DP değerinin test verisindeki tüm tahminlere (DP + YN) oranıdır. Anma ve kesinlik değerlerinin harmonik ortalaması, $(2 * \text{Kesinlik} * \text{Anma}) / (\text{Kesinlik} + \text{Anma})$, ise F-ölçüm değerini (F-ölçeği değeri veya kesinlik-anma değerlerinin katsayıları harmonik ortalama da aynı olması nedeniyle F_1 ölçeği) vermektedir. Performans değerlendirme metriği olarak sıklıkla kullanılan anma (recall) ve F-ölçüm değeri (F-measure) çalışmamızda literatür verileri ile yapılan örtüşme analizinde kullanılmamıştır. Anma değeri hesaplanırken GBA tarafından tahmin edilmemiş ama PPI veri kümesinde yer alan ilişkileri gösteren yanlış negatif (YN veya yanlış eksi) değeri kullanılmaktadır. Anma değeri GBA tarafından doğru tahmin edilen gen ilişkilerinin (DP), PPI veri kümesinde yer alan tüm gen ilişkilerine (DP + YN) oranıdır. PPI veri kümesinin 1,5 milyon verinin üzerinde olmasından dolayı elde edilecek anma değerleri çok küçük olacağı için çalışmamızda sadece performans ölçütü olarak kesinlik ve DP değerleri kullanılmıştır. Anma değerleri çok küçük olduğu için F-ölçüm değerleri de çalışmamızda performans değerlendirme kriteri olarak yer almamıştır.

Tez kapsamında GBA tarafından tahmin edilen ve PPI veri kümesinde yer alarak örtüşen gen- gen ilişkilerinin şans faktörüne bağlı oluşma olasılığı değerlendirilmiştir. Bir diğer deyişle örtüşen ilişkilerin istatistiksel olarak anlamlılığının kontrolü de çalışmamızda belirlenmiştir. Bunun için *ganet* R paketinin sunduğu istatistiksel metot olan Fisher Kesinlik Testi (Fisher's Exact Test) kullanılmıştır. Bu metot örtüşen gen- gen

ilişkilerinin hipergeometrik dağılım bilgisini kullanmaktadır. Hipergeometrik dağılım Binom dağılımına benzemektedir. Binom dağılımında ana kütlede her seferde yerine konarak oluşturulan örneklem (nesnelerin örneklemde çekilmesi) hipergeometrik dağılımda ana kütlede tekrar yerleştirme yapılmadan elde edilmektedir. Bu durumda ana kütledeki örnek sayısı değişeceği için n. defa tekrarlanan örneklem işleminin sonuçları bir önceki örneklemeye bağımlı olur. Örneklemeler (deneyler) birbirinden bağımsız olmadığı için Binom dağılımı yerine hipergeometrik dağılım kullanılmıştır. Çizelge 5.1'de hata matrisinde (confusion matrix) yer alan performans değerleri hipergeometrik dağılımın olasılık değerini gösteren hesaplamada kullanılmıştır. Hata matrisinde yer alan doğru negatif (DN veya doğru eksi) genlerin olası tüm ilişkilerinden PPI verisinde olan ve GBA tarafından tahmin edilen gen- gen ilişkileri çıkartıldığında elde edilmektedir.

Çizelge 5. 1 Hata matrisi

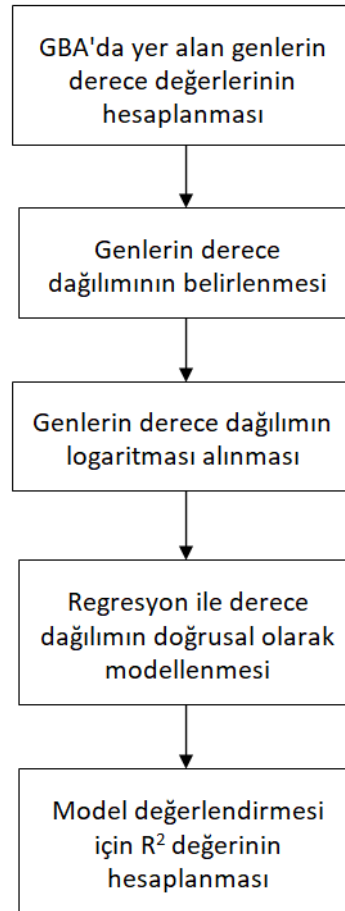
	PPI verisinde olan ilişkiler	PPI verisinde olmayan ilişkiler
GBA tarafından tahmin edilen ilişkiler	DP	YP
GBA tarafından tahmin edilmemiş ilişkiler	YN	DN

$$p = \frac{\binom{DP+YN}{DP} \binom{YP+DN}{YP}}{\binom{DP+YN+YP+DN}{DP+YP}} \quad (5.1)$$

Fisher Kesinlik Testi, PPI verisiyle örtüşen gen- gen ilişkilerinin şans eseri elde edilme olasılığı ile GBA tarafından tahmin edilme olasılığını sıfır hipotezinde (H_0) eşit olarak değerlendirmektedir. Elde edilen p-değeri belirli bir eşik değerinin altında ise sıfır hipotezi reddedilir ve örtüşen gen- gen ilişkilerinin şansa bağımlı olarak elde edilmediği sonucu ortaya çıkar. Çalışmamızda GBA'ların istatistiksel anlamlılığının kontrolü için p değeri eşik değeri olan 0.05'in altında çıkan GBA'lar değerlendirme dışı tutulmuştur.

5.2 GBA'ların Topolojik Değerlendirilmesi

Gen ağları az sayıda genin çok sayıda ilişkisinin olduğu, geri kalan çok sayıda genin de az sayıda ilişkisinin olduğu ölçeksiz ağlardır. Bu ağlardaki genlerin (düğümlerin) derece dağılımları kuvvet yasası dağılımını takip eder. Bölüm 1.1'de genlerin derece değerlerinin kuvvet yasası dağılımına sahip olduğu gösterilmiştir. Log-log graflar ile genlerin derece değerlerini gösteren kuvvet yasası dağılımı doğrusal bir şekilde gösterilmektedir. Gen dereceleri ve bu derecelerin olasılıkları arasındaki ilişkiyi veren dağılımın logaritması alındıktan sonra doğrusal modellenabilir yapıya dönüşmektedir. Çalışmamızda regresyon analizi uygulanarak doğrusal model oluşturulmuş ve bu modelin başarısı uygunluk iyiliği, (goodness of fit) belirleme katsayısı (R^2 , coefficient of determination) istatistiği ile ölçülmüştür. R^2 değeri ile doğrusal modelin, gözlenen değerleri kestirim başarısı ölçülmektedir. R^2 değerinin hesaplanmasıyla ilgili detaylar Bölüm 1.1'de anlatılmıştır.



Şekil 5. 1 Topolojik değerlendirme temel adımları

Çalışmamızda yüksek R^2 değerleri GBA'larının ölçeksiz ağ tanımını iyi yansıttığı yani gen ağı üzerindeki genleri temsil eden düğümlerin komşuluk değerlerinin kuvvet yasasına uygun olduğunu göstermektedir. Değerlendirme aşamasında tüm GBA'ların R^2 değeri hesaplanmıştır. Buna göre R^2 değeri bire yakın çıkan GBA'lar ölçeksiz ağ (biyolojik ağlara) tanımına yakın olarak belirlenmiştir. R^2 değeri sıfıra yakın olan GBA'lar ise rassal ağlara yakın topolojik özellik göstermektedir. Şekil 5.1'de yukarıdaki topolojik değerlendirmeye ait temel adımlar gösterilmiştir.

5.3 GBA'ların Gen Ontoloji Terimleri Kullanarak Biyolojik Değerlendirilmesi

Ortak ifade edilen genler tarafından üretilen proteinler hücre içerisinde benzer biyolojik görevlerde, süreçlerde ve benzer hücre alanlarında yer alma eğilimindedir. Gen Ontolojisi (GO), aynı alan (domain) içerisinde terimlerin bir veya birden çok terimle ilişkili olduğu bir yönlü graftır. Gen Ontolojisi proteinlerin moleküler fonksiyon (molecular function), yer aldıkları biyolojik süreç (biological process) ve protein hücre içerisindeki konum bilgisi (cellular component) özelliklerini sunan hiyerarşik bir sözlük yapısındadır (vocabulary). Bu sözlükte her bir terim (GO- terimi, GO- terms) proteinlere ait bir özelliği göstermektedir. GO- terimleri üç ana alanda (domain) sınıflandırılabilir.

- Hücresel Birleşen (Cellular Component): Hücrenin parçaları veya hücre dışı çevresi hakkında bilgi verir.
- Moleküler Fonksiyon (Molecular Function): Kataliz, bağlanma gibi moleküler seviyede gen ürünleri olan proteinlerin temel aktivitelerini gösterir.
- Biyolojik Süreç (Biological Process): Belirli bir başlangıcı ve bitişi olan süreçler biyolojik olaylardan (events) oluşmakla birlikte, bu olaylar genellikle birden çok moleküler fonksiyonun başarıyla gerçekleşmesiyle elde edilir.

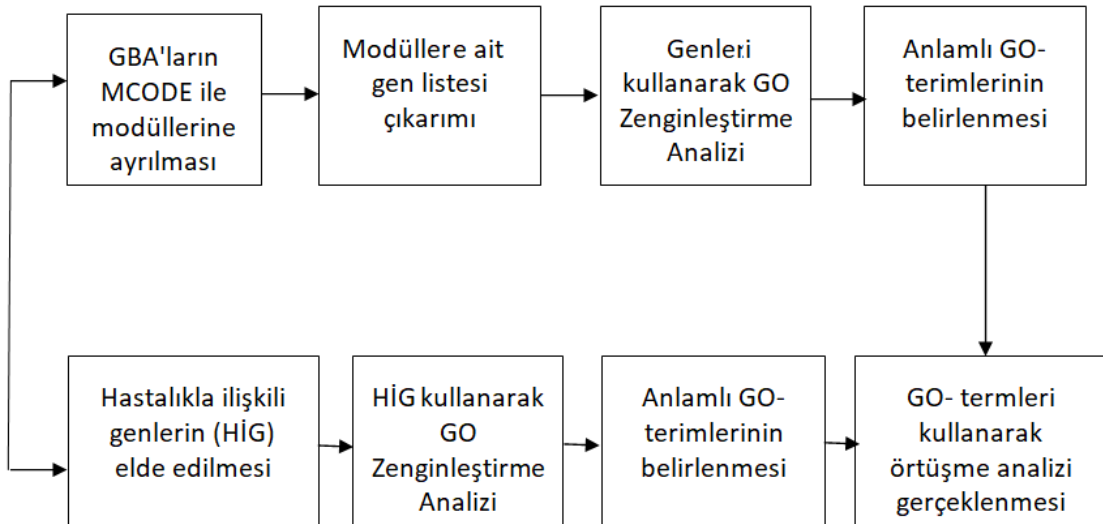
Her bir GO- terimi yukarıda sıralanan üç ana özellikten biriyle ilgili bilgi vermektedir. Şekil 5.2'de alanlarına göre bazı GO- terimleri ve o terimlerle ilgili gen listeleri örnek olarak gösterilmiştir. Şekil 5.2'de yer alan GO- terimleriyle ilgili örneğe bakıldığında, hücresel birleşen alanında yer alan GO -terime örnek "extracellular space", moleküler fonksiyon alanında yer alan GO- terime örnek "receptor binding" ve biyolojik süreç

alanında yer alan GO- terimine örnek multicellular "organismal development" verilebilir.

GO-ID	Description	p-val	corr p-val	cluster freq	total freq	genes
50789	regulation of biological process	3.1002E-12	3.4413E-10	143/244 58...	6548/17774	RAPH1 PVR ELK3 FOXQ1 ZFP36 NAMPT SESN2 PIM1 COL11 JUNC PHLD2A1 TRIM22 IER3 EPHA4 HIBN
5615	extracellular space	1.9897E-10	1.1043E-8	35/244 14.3%	747/17774	COL12A1 PDGFA TGFA CXCL1 CXCL3 PVR CXCL2 LRBP CXCL5 PTHLH SFN LGALS8 IL13RA2 LDLR IL1
5576	extracellular region	1.3144E-8	4.8634E-7	59/244 24.1%	2026/17774	COL12A1 PIK3CD CXCL1 CXCL3 PVR CXCL2 CXCL5 ADAMTSS C19ORF10 CASP1 TIMP1 LGALS8 IL13
5515	protein binding	2.5981E-8	6.9680E-7	154/244 63...	8117/17774	PVR ELK3 ZFP36 NAMPT PIM1 CYP1B1 SKP2 TNF4 COL11 IL13RA2 JUNC HMG13 SH2D5 PHLD2A1 TR
7275	multicellular organismal development	3.7481E-8	6.9680E-7	75/244 30.7%	2970/17774	COL12A1 CXCL1 HK2 ELK3 FOXQ1 PIM1 TIMP1 PHLD2A2 JUNC HRAS EPHA4 IL11 SL2A14 ARG2 TNF
6950	response to stress	3.7665E-8	6.9680E-7	53/244 21.7%	1773/17774	CXCL1 CXCL3 CXCL2 HK2 ELK3 NFKB1Z PIM1 CASP1 KDM6B IL11 SERPINB2 DUSP1 ITGA2 TPM1 PLA
5102	receptor binding	4.4424E-8	7.0444E-7	35/244 14.3%	921/17774	5... LAMA3 PDGFA TGFA CXCL1 CXCL3 CXCL2 DLL1 CXCL5 PTHLH EFNB1 ADAMTSS SOCS1 C19ORF10 NA
9605	response to external stimulus	1.7475E-7	2.4246E-6	25/244 10.2%	551/17774	3... PDGFA CXCL1 PTGS2 CXCL3 CXCL2 ARNTL2 CXCL5 CYP27B1 JUNC HRAS TGFBR2 DUSP1 CCL20 ITGA2
9653	anatomical structure morphogenesis	2.2355E-7	2.7129E-6	40/244 16.3%	1218/17774	... PDGFA ADRB2 DLL1 LRBP ELK3 FOXQ1 EFNB1 SOCS1 CHST11 PLXNA2 CASP1 PHLD2A2 JUNC HRAS IE
7165	signal transduction	2.4441E-7	2.7129E-6	53/244 21.7%	1878/17774	... RAPH1 PIK3CD CXCL1 CXCL5 ELK3 ZFP36 NAMPT CASP1 HRAS DUSP4 EPHA4 FGFR1 PLAU CRK1
30154	cell differentiation	1.4392E-6	1.4523E-5	47/244 19.2%	1668/17774	... LAMA3 ADRB2 DLL1 HK2 SLC9A1 EFNB1 SOCS1 CHST11 CASP1 SFN FLNB TIMP1 JUNC HRAS MCL1 L
7610	behavior	7.2257E-6	6.6838E-5	20/244 8.1%	466/17774	2... EPHA4 EGR1 TGFBR2 CCL20 PLAU PDGFA CXCL1 PPBP FOS PTGS2 CXCL3 CXCL2 CXCL5 VEGFA FOSL1
9719	response to endogenous stimulus	5.3700E-6	4.5851E-4	19/244 7.7%	493/17774	2... LYN EGR1 TGFBR2 DUSP1 ITGA2 PDGFA FOS PTGS2 EREG FOSL1 LAT51 CYP27B1 IL1A MMP14 SOCS1
8283	cell proliferation	3.4412E-4	2.7284E-3	16/244 6.5%	434/17774	2... CD274 TCF7L2 TGFBR2 ITGA2 TGFA CXCL1 EREG FERMT1 IL1A MMP14 IL23A PIM1 SFN SKP2 HMO1
9607	response to biotic stimulus	4.7524E-4	3.5167E-3	16/244 6.5%	447/17774	2... PPP1R15A CCL20 ODC1 PPBP FOS PTGS2 FOSL1 CYP27B1 IL1A SOCS1 IL23A IL1B CASP1 PTX3 TR1
40007	growth	6.1944E-4	4.0858E-3	10/244 4.0%	208/17774	1... TGFBR2 BHP2 CHST11 PLAU GAS1 PTGS2 FGFR2 EREG VEGFA SLC9A1
8219	cell death	6.5987E-4	4.0858E-3	21/244 8.6%	698/17774	3... PPP1R15A TGFBR2 ITA1 TNFRSF12A TNFRSF10B TNFRSF10A IL1A ARHGEF9 IL1B CASP1 GAS1 CASP
9790	embryonic development	6.6256E-4	4.0858E-3	19/244 7.7%	601/17774	3... TCF7L2 TGFBR2 LIF PDGFA DLL1 TBK3 VEGFA EFNB1 BHP2 VANGL2 CHST11 GJB3 GJB5 PLXNA2 GAS1
16265	death	7.2285E-4	4.2230E-3	21/244 8.6%	703/17774	3... PPP1R15A TGFBR2 ITA1 TNFRSF12A TNFRSF10B TNFRSF10A IL1A ARHGEF9 IL1B CASP1 GAS1 CASP
3	reproduction	8.0927E-4	4.4914E-3	23/244 9.4%	810/17774	4... SLC2A14 TCF7L2 TGFBR2 LIF FOS PTGS2 SERPINB5 HK2 PTHLH TBK3 EREG VEGFA PIAS1 FOSL1 CYP2
5578	proteinaceous extracellular matrix	1.2767E-3	6.7481E-3	12/244 4.9%	311/17774	1... ADAMTSS COL27A1 FLRT3 COL12A1 TFP12 LAMA3 MFAP2 FN1 TIMP1 ANGIPT4 P13 VEGFA
9628	response to abiotic stimulus	1.4457E-3	7.2941E-3	14/244 5.7%	404/17774	2... DUSP1 ITGA2 FOS ADRB2 VEGFA SLC9A1 FOSL1 IL1A MMP14 BHP2 IL1B JUNC HRAS STX1A
5488	binding	3.3602E-3	1.6217E-2	189/244 77...	12354/17774	... GLTP PVR RRP9 ELK3 FOXQ1 ZFP36 ZNF608 NAMPT PIM1 CYP1B1 LGALS8 SKP2 TNF4 COL11 MTPAP
5737	cytoplasm	3.7887E-3	1.7523E-2	126/244 51...	7641/17774	... GLTP RAPH1 PVR ELK3 ZFP36 FAM110C NAMPT SESN2 PIM1 CYP1B1 LRAT LGALS8 TNF4 MTPAP PH
5886	plasma membrane	4.0140E-3	1.7593E-2	69/244 28.2%	3732/17774	... ITPRIP RAPH1 PIK3CD PVR HK2 PIM1 TNF4 HRAS SLC16A3 SH3D19 EPHA4 FGFR1 TNFRSF12A MF
7154	cell communication	4.1208E-3	1.7593E-2	21/244 8.6%	812/17774	4... IL11 FGFR1 TGFBR2 GDF15 CCL20 PDGFA FOS DLL1 CXCL5 PTHLH TBK3 EREG FOSL1 EFNB1 BHP2 GJ
8150	biological process	4.7855E-3	1.9674E-2	212/244 86...	14289/17774	... GLTP COL12A1 RAPH1 RRP9 ELK3 NIPAL1 NAMPT CYP1B1 LRAT SKP2 TNF4 COL11 JUNC TRIM22 FE
16043	cellular component organization	5.7448E-3	2.2768E-2	50/244 20.4%	2561/17774	... COL12A1 CXCL1 HK2 PIM1 HIST1H1E CCN1 HRAS HMG13 TRIM22 HIST1H2AC SH3D19 KDM6B EPI
3700	transcription factor activity	5.9484E-3	2.2768E-2	23/244 9.4%	949/17774	5... EGR1 IRX2 TCF7L2 TSC22D1 ZNF192 FOS GRHL2 ARNTL2 TBK3 ELK3 RUNX1 FOSL1 FOXQ1 MMP14 K
30234	enzyme regulator activity	7.7915E-3	2.8829E-2	21/244 8.6%	860/17774	4... SERPINB3 SERPINB1 SERPINB2 TFP12 TNFRSF10B TNFRSF10A CXCL1 ARHGAP27 ARHGAP25 SERP
7049	cell cycle	1.3558E-2	4.8547E-2	19/244 7.7%	794/17774	4... PPP1R15A PRNP TCF7L2 TGFBR2 DUSP1 NOLC1 GUS2 HK2 EREG LAT51 CYP27B1 CCN1 IL2 SESN2 PI

Şekil 5. 2 GO- terimleri ve terimlerle ilişkili genlere ait örnek

Şekil 5.3'te GO- terimleri kullanarak GBA'ların biyolojik değerlendirmesine ait adımları gösterilmiştir.



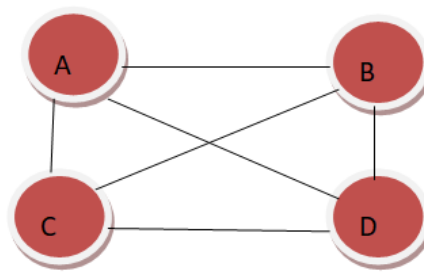
Şekil 5. 3 GBA'ların biyolojik değerlendirmesine ait temel adımları

Genom tabanlı ağ çıkarımında elde edilen GBA'lar çok sayıda gen- gen ilişkisi içermektedir. Literatürde GBA'ların biyolojik değerlendirmesi için GBA'larda yer alan ve birbirleriyle iyi derecede bağlı (highly coupled) gen-gen ilişkilerinin oluşturduğu modüller kullanılmaktadır [64], [75], [76]. Tez kapsamında GBA'ları biyolojik olarak değerlendirmek için GBA'larda yer alan modüller kullanılmıştır.

GBA'ları alt ağlara (modüllerine) ayırmak için çalışmamızda MCODE algoritması kullanılmıştır [77]. MCODE algoritması üç temel adımdan oluşur :

1. Düğüm Puanlama (Node Scoring): Ağ üzerinde komşuları birbirlerine sıkıca bağlı düğümlerin keşfedilmesi ve bu düğümlere yüksek puan ataması.
2. Küme Keşfi (Clustering Finding): En yüksek puanı olan düğümün belirlenmesi ve belirlenen düğümün komşularından eşik değerinin üzerinde puanı olan düğümlerin modül listesine eklenerek modüllerin oluşturulması. Bu işlem tüm ağda K-çekirdek (K- core) ağların tümü bulunana kadar devam eder.
3. Art İşlem (Post processing): Kümelere sadece bir düğümlle bağlanmış düğümlerin kümelerden çıkarılması (haircut)

Düğüm Puanlama: En yüksek K- çekirdek graflarda (the highest k-core graphs) yer alan tüm düğümlerin ağ üzerinde yer alan diğer düğümlerle en az k adet bağlantısı bulunur. Örneğin aşağıdaki alt ağdaki K değeri üçtür.



Şekil 5. 4 K- çekirdek graf ($K=3$)

GBA'da yer alan tüm düğümlerin (genlere) puan atama işlemi, düğümü içeren modülün (graf veya ağ) yoğunluk değeri ve en yüksek K değerini kullanarak hesaplanmaktadır. Bir modülün yoğunluk değeri bir modülde yer alan tüm ilişki sayısının, olası tüm ilişki sayısına oranı olarak belirlenir. Olası ilişki sayısı döngü olmayan ağlarda (düğüm sayısı * düğüm sayısı -1) /2 olarak hesaplanır. Elde edilen yoğunluk

değeri ile düğümün ait olduğu graftaki en yüksek K değeri çarpılarak düğüme ait puanlama hesabı tamamlanmış olur.

Şekil 5.4'te, A düğümünün puan hesaplaması, A düğümünü içeren modülde 6 tane ilişki varken olası ilişki sayısı da yine 6'dır. A düğümü $K=3$ çekirdek grafında yer aldığı için A düğümünün düğüm puanı $(6/6)*3= 3$ 'tür. Aynı yaklaşımla MCODE algoritması ilk aşamadasında GBA üzerindeki tüm genler için puan hesabı yapılmaktadır.

Küme Keşfi: Düğümler içerisinde puanı en yüksek düğüm tohum (seed) düğüm olarak belirlenir. Tohum düğümün komşularından eşik değerinden yüksek puanı olanlar K -çekirdek grafi oluşturmak için küme listesine eklenir. Eşik değerini hesaplamak için MCODE algoritması için denklem 5.2'de yer alan yaklaşım kullanılır.

$$\text{Eşik Değeri} = (1 - \text{kullanıcı parametresi}) * \text{tohum düğümü puan değeri} \quad (5.2)$$

Kullanıcı tarafından belirlenen parametre ve tohum düğüm değeri kullanılarak eşik değeri belirlenmiş olur.

Tohum düğümü K - çekirdek grafını eklemek için kullanılan ilk düğümdür. Tohum düğümüyle ilişkili ve düğüm puanı eşik puanından fazla olan düğümler düğüm listesine eklenir ve K - çekirdek grafi yani modülü oluşturulmuş olur. Oluşturulan ilk K - çekirdek grafında yer alan düğümler dışındaki en yüksek puana sahip düğüm tohum düğüm olarak belirlenir. Tohum düğümüyle aynı K - çekirdek grafında bulunan ve eşik değerinden fazla puana sahip diğer düğümlerde alt ağ oluşumunda kullanılacak düğümler olarak belirlenir.

Art İşlem: Oluşturulan modüllerdeki düğümlerle tek bir bağlantısı olan düğümlerin varsa bu düğümlerin elenmesiyle (haircut) art işlem gerçekleştirilmiş olur.

MCODE algoritması için *ProNet* R paketi [78] standart parametre değerleriyle çalışmamızda kullanılmıştır. Standart parametre değerlerinde graf yoğunluk değeri 0.8, düğüm minimum puan değeri 0.5 olarak paket tarafından belirlenmiş ve budama olarak tek bağlantı elemesi (haircut) kullanılmıştır.

GBA'ların biyolojik değerlendirmesi için MCODE algoritması ile modüller belirlenmiştir. Şekil 5.3'de gösterildiği üzere modüllerdeki genlerin ilişkili olduğu GO-terimleri *gProfiler* R paketi [79] kullanılarak GO zenginleştirme analiziyle elde edilmiş olur. GO

zenginleştirme analizi, modüllerdeki genlerin ilişkili oldukları GO- terimlerini kullanıcılara sunmaktadır. Bu işlemlerden sonra bir GBA'dan elde edilen modüllerin ilişkili olduğu GO- terimleri çıkartılır. GO- terimlerinin istatistiksel olarak anlamlı olup olmadığının kontrolü bölüm 5.1'de anlatıldığı gibi ilk olarak p-değeri hesabıyla gerçekleşir. Fisher kesinlik testiyle p- değerleri hesaplandıktan sonra istatistiksel olarak anlamlı GO- terimleri GBA'ya ait modüller için belirlenmiş olur.

Modüllerle ilişkili GO- terimlerinin hastalıklarla ilişkisinin incelenmesi için ilk olarak hastalıklarla ilişkili genler OMIM hastalık veri setinden [80] elde edilmiştir. Daha sonra hastalıkla ilişkili genler *gProfileR* R paketi kullanılarak GO zenginleştirme analiziyle hastalıklarla ilişkili GO- terimleri elde edilmiştir. Modüllere ait GO- terimleri ile hastalıkla ilişkili GO- terimlerinin örtüşme analizi gerçekleştirilmiştir. Bu örtüşme analizinde Fisher Kesinlik testine göre p- değeri yine 0.05'ten küçük olan modüller kullanılarak örtüşme analizi yapılmıştır. Biyolojik değerlendirmede GBA'daki modüllerden elde edilen GO- terimlerinin kaç tanesinin aynı zamanda hastalıklarla ilişkili GO- terimleri olduğu belirlenmiştir.

DENEYSEL SONUÇLAR

Bu bölümde çift katlı entegrasyon mimarisiyle farklı aşamalarda oluşturulmuş farklı kanser tiplerine ait GBA'lar örtüşme analizi, topolojik ve biyolojik değerlendirme yöntemlerine göre performansları değerlendirilmiştir. Deneysel sonuçlar bölümünde aynı zamanda daha önce literatürde miRNA- hedef gen verileri üzerinde kullanılmamış geliştirdiğimiz hash tabanlı birliktelik kuralı algoritmasının (ARNetMiT) sonuçları da aynı veri kümesi üzerinde popüler gen ağı çıkarım algoritmalarıyla karşılaştırılmıştır. ARNetMiT performans değerlendirmesinde de Bölüm 5'te yer alan değerlendirme kriterleri kullanılmıştır.

6.1 miRNA- Hedef Gen Veri Kümesini Kullanarak ARNetMiT Yaklaşımının Performans Değerlendirilmesi

Performans değerlendirmesi için miRNA- hedef gen verileri üzerinde dGBA ve dgGBA'nın performans sonuçları, aynı veriler üzerinde ARACNE, C3NET ve WGCNA algoritmaları kullanılarak elde edilmiş GBA'lar ile karşılaştırılmıştır. Bunun için performans kriterleri olan örtüşme analizi, topolojik ve biyolojik değerlendirmeden yararlanılmıştır.

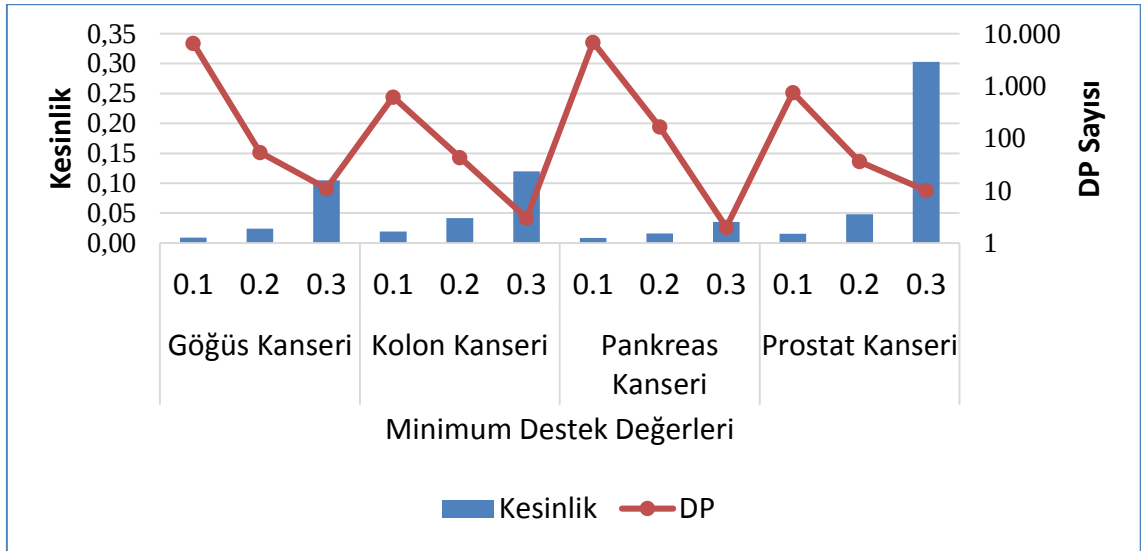
ARNetMiT R paketi 152 hastalıkla ilgili olası miRNA- hedef gen ilişkilerini kullanıcılara sunmaktadır. ARNetMiT performans analizinde göğüs, kolon, pankreas ve prostat kanserleriyle ilişkili miRNA- hedef gen verileri kullanılarak GBA çıkarımı gerçekleştirilmiştir. Çizelge 6.1'de bu dört hastalıkla ilişkili miRNA- hedef gen ilişki sayısı verilmiştir.

Çizelge 6. 1 Hastalıklar ilişkili miRNA- hedef gen ilişki sayısı

Hastalık	miRNA- hedef Gen Sayısı
Göğüs Kanseri	2179
Kolon Kanseri	2844
Pankreas Kanseri	2076
Prostat Kanseri	3232

6.1.1 ARNetMiT Yaklaşımının Örtüşme Analiz Değerlendirmesi

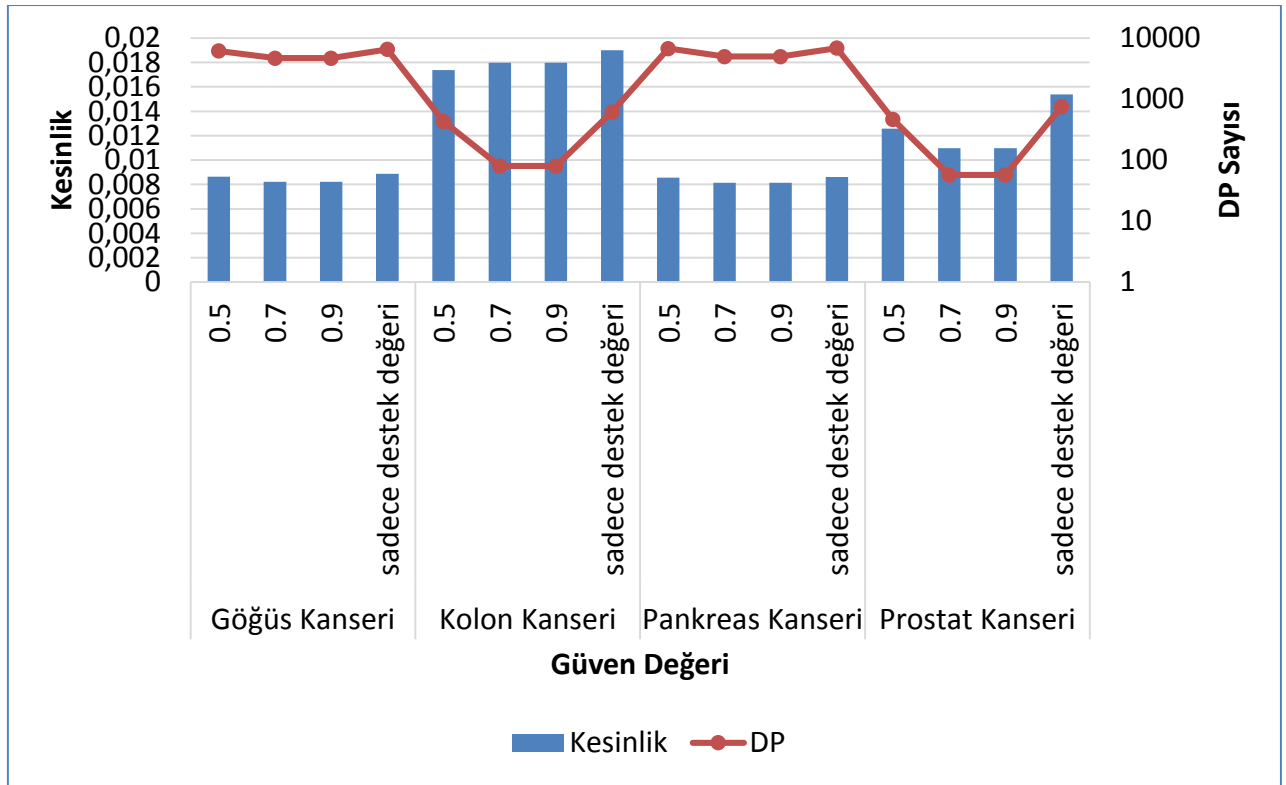
ARNetMiT kullanarak elde edilen dGBA ve dgGBA'ların performanslarının diğer algoritmalarla karşılaştırılması için en uygun destek ve güven değerleri ilk olarak belirlenmeye çalışılmıştır. Şekil 6.1'de göğüs, prostat, kolon ve pankreas kanserleriyle ilişkili gen- miRNA verileri üzerinde üç farklı minimum destek (0.1, 0.2, 0.3) değerine ait sonuçlar karşılaştırılmıştır. Şekil 6.1'de dört farklı kanser türüne ait farklı destek değerleri kullanılarak oluşturulmuş GBA'ların kesinlik ve doğru pozitif (DP) değerleri gösterilmiştir.



Şekil 6. 1 Dört farklı kanser türüyle ilişkili dGBA'ların kesinlik ve DP değerleri

Buna göre destek değeri arttıkça GBA'ların literatür verisiyle örtüşen gen- gen ilişkisi sayısının azaldığı gözlemlenmektedir. DP sayısında azalma olmasına rağmen destek değerleri arttıkça kesinlik değerleri de artmıştır. Her dört kanser türüyle ilişkili GBA'lar incelendiğinde destek değeri 0.1 alındığında kesinlik değerleri yüzde beşin altında çıkmış olmasına rağmen DP sayısı 10 binlere yakın çıkmıştır. Destek değeri arttıkça elenen gen-gen sayısı arttığı için DP sayısında azalma olmuştur. Entegrasyon çalışmamızda farklı veri kümelerine ait değişken gen- gen ilişki tahminlerinden yararlanılması hedeflenmiştir. Bu nedenle çift katmanlı entegrasyon mimarisinde kullanılan gGBA'ların tahmin gücünden yararlanmak için destek değerleri 0.1 olarak alınmıştır.

Destek değerlerine göre gen- gen ilişki budaması gerçekleştirildikten sonra destek- güven tabanlı GBA'lar için uygun güven değerleri üç farklı güven değeri kullanılarak Şekil 6.2'de gösterildiği gibi elde edilmiştir.

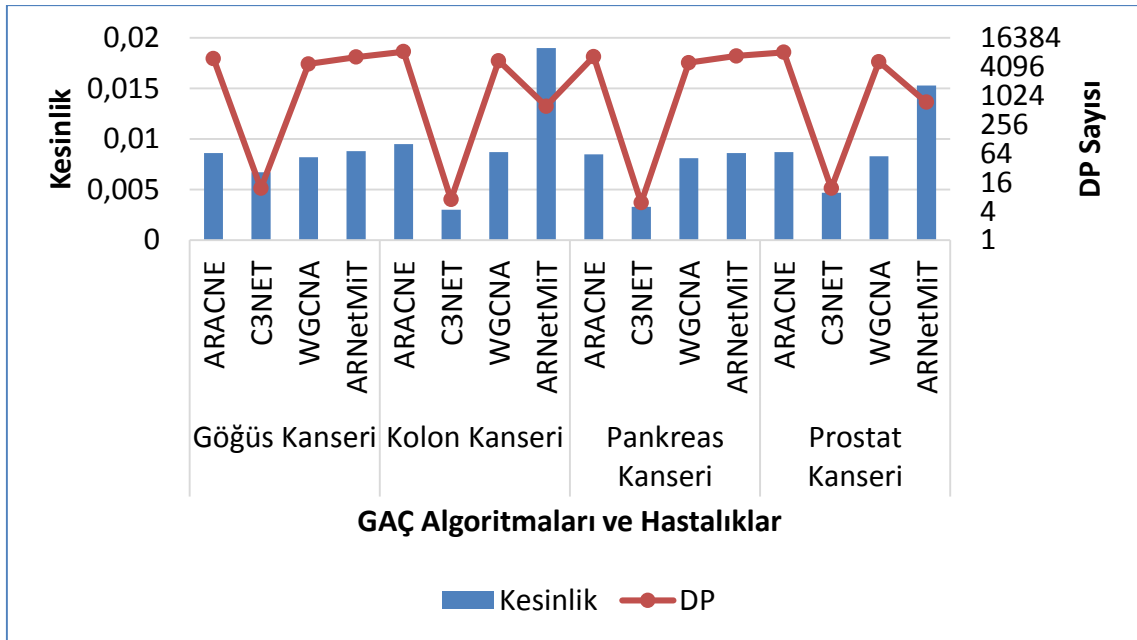


Şekil 6. 2 Dört farklı kanser türüyle ilişkili dgGBA'ların kesinlik ve DP değerleri

Destek değerinden sonra güven değerine bağlı gen- gen ilişki budaması gerçekleştirileceği için Şekil 6.2'de yer alan GBA'lar için destek değeri 0.1 olarak alınmıştır. Minimum güven değerleri 0.5, 0.7 ve 0.9 olarak belirlendikten sonra dört farklı kanser türüyle

ilişkili dgGBA çıkarımı yapılmıştır. Kesinlik değerlerine göre dört kanser türüyle ilişkili GBA'lar incelendiğinde en yüksek kesinlik değeri güven budaması yapılmadan sadece destek budamasıyla elde edilmiş dGBA'lar tarafından elde edildiği gözlemlenmiştir. Güven değeri en düşük değer olan 0.5 seçildiğinde elde edilen kesinlik değerleri genel anlamda diğer güven değerleri kullanılarak elde edilen GBA'lardan yüksek çıkmıştır. Güven değeri 0.5 olduğunda GBA'ların DP sayısı da aynı zamanda yüksek çıkmıştır. Güven değerinin 0.7 ve 0.9 seçildiğinde oluşturulan dgGBA'lardan benzer kesinlik ve DP değerleri elde edilmiştir. Örtüşme analizine göre güven değeriyle gen- gen ilişkisi budaması performansı arttırmamış ve ilişki tahmin gücünü azaltmıştır.

ARNetMiT R paketi seçilen hastalıkla ilişkili miRNA- hedef gen ilişkilerini ikili tablo olarak kullanıcılara sunmaktadır. Bu tabloyu kullanarak farklı gen ağı çıkarım algoritmaları miRNA- hedef gen verilerini kullanarak gen ağı çıkarımı yapabilmektedir. Çalışmamızda güven tabanlı yaklaşımlara göre daha iyi sonuç vermesi ve ilişki tahmin gücü nedeniyle dGBA'lar ile (minimum destek değeri =0.1) ARACNE, C3NET ve WGCNA algoritmalarının performansları birbirleriyle karşılaştırılmış ve sonuçlar Şekil 6.3'te sunulmuştur.

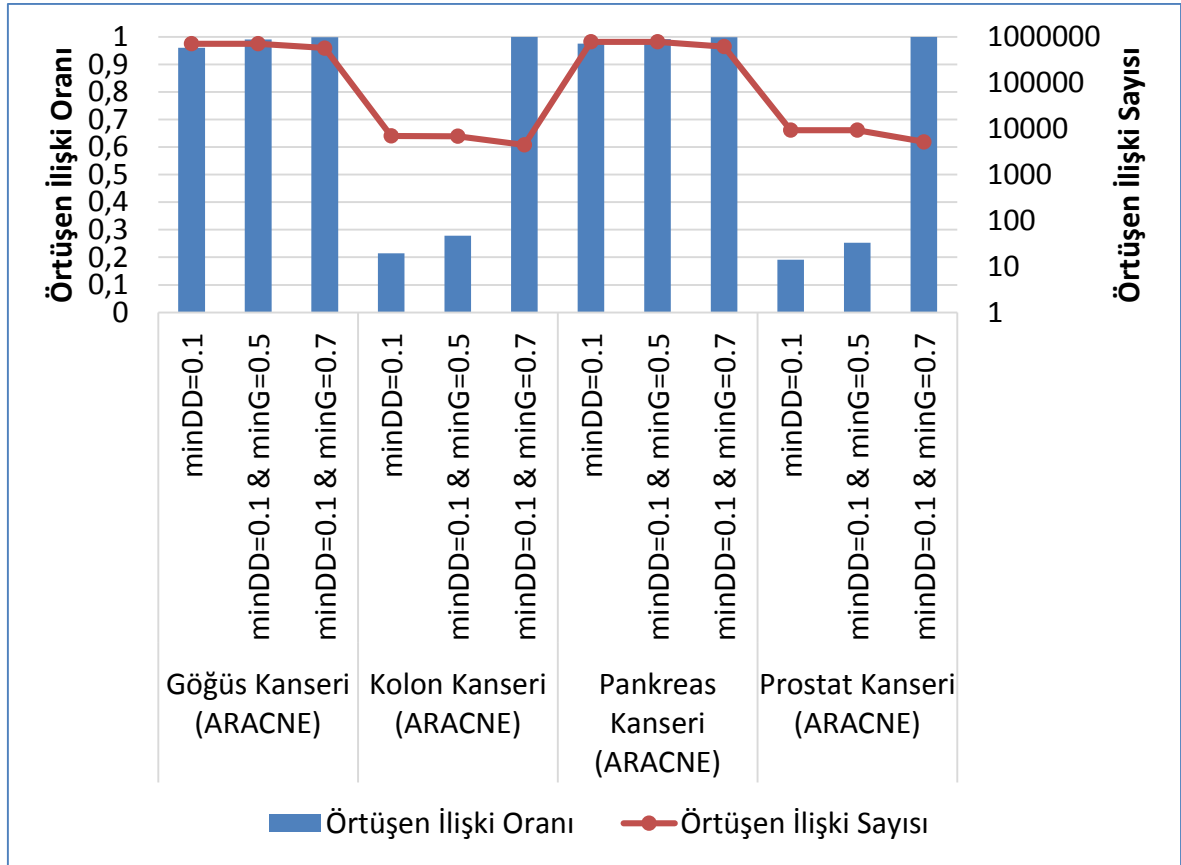


Şekil 6. 3 Farklı GAÇ algoritmalarıyla elde edilmiş GBA'ların kesinlik ve DP değerleri

Şekil 6.3'e göre dört farklı hastalıkla ilişkili miRNA- hedef gen verileri üzerinden en düşük DP sayısı ve kesinlik değerleri C3NET algoritması kullanılarak oluşturulmuş

GBA'lardan elde edilmiştir. ARNetMiT tabanlı dGBA'ların kesinlik değerleri göz önüne alındığında özellikle kolon ve prostat kanserinde daha yüksek sonuçlar alınmıştır. dGBA'lar hem ilişki kestirim gücü hem de kesinlik değerlerine göre dört kanser türü için de diğer algoritmalarla oranla eşit veya daha iyi sonuçlar verdiği gözlemlenmiştir.

Bu bölümde son olarak ARNetMiT tabanlı GBA'ları oluşturan gen- gen ilişkileri ile gen ağı çıkarım algoritmaları tarafından tahmin edilen GBA'larda yer alan gen- gen ilişkileri arasında ortak tahmin edilen ilişkiler belirlenmiştir. Amacımız ARNetMiT tabanlı ilişki tahminleri ile gen ağı çıkarımı algoritmaları kullanılarak elde edilen ilişki tahminleri arasındaki benzerlikleri ölçmektir. Şekil 6.4'te dört farklı kanser türünde gen ağı çıkarım algoritmaları tarafından oluşturulmuş ve en yüksek kesinlik değerleri elde edilmiş GBA'lar ile ARNetMiT tabanlı GBA'larda ortak tahmin edilen ilişkiler gösterilmiştir.



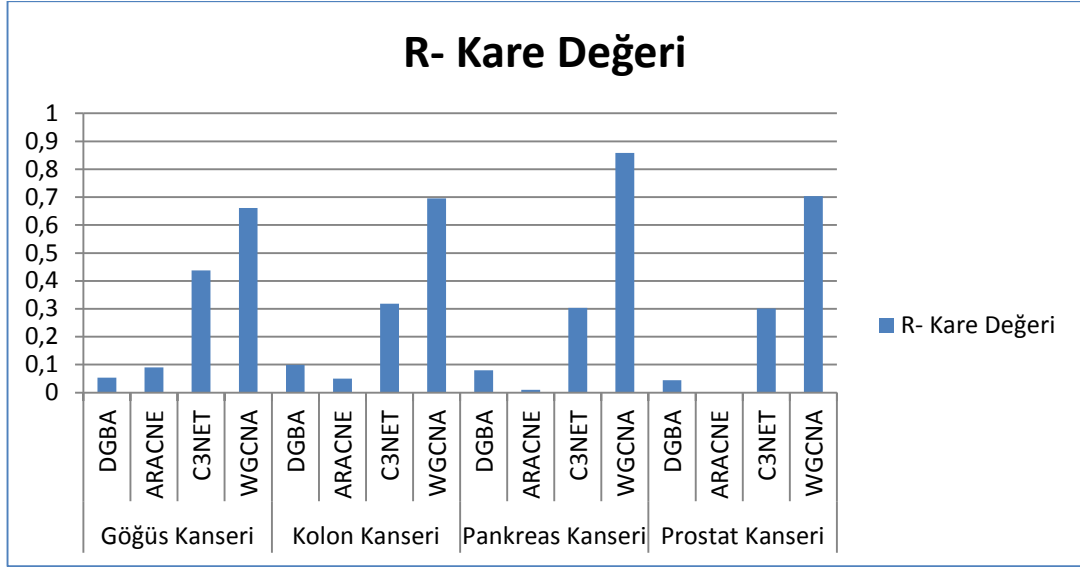
Şekil 6. 4 GAÇ algoritmaları ve ARNetMiT tabanlı GBA'lardaki ortak tahmin edilen ilişki değerleri

Dört farklı kanser türünde de en başarılı sonuçları veren gen ağı çıkarım algoritması Şekil 6.3'de gösterildiği gibi ARACNE olmuştur. Bu nedenle ARACNE tabanlı GBA'lar ile

ARNetMiT tabanlı GBA'lar arasında ortak olarak belirlenmiş gen- gen ilişkileri Şekil 6.4'te gösterilmiştir. Sonuçlara göre ARACNE ve ARNetMiT tabanlı GBA'lardaki ortak örtüşme oranları bazı hastalık türlerinde %95 değerlerinin üzerinde çıkmıştır. Dört kanser türünde de dgGBA'ların dGBA'lara oranla daha fazla oranda ortak ilişki tahmin ettiği sonucuna varılmıştır. Bununla birlikte güven değeri arttıkça ortak ilişki tahmin oranının daha da yükseldiği gözlemlenmiştir. Minimum güven değeri 0.7 olarak belirlendiğinde oluşturulan dgGBA'lardaki gen- gen ilişkilerinin hemen hemen hepsi ARACNE tabanlı GBA'larda da yer almaktadır. Göğüs ve pankreas kanserinde dGBA'ların örtüşme oranı her ne kadar yüksek çıksa da kolon ve prostat kanserinde ortak ilişki oranı %20'lerin altında kalmıştır. Budama yoğunluğu arttıkça ARNetMiT tarafından tahmin edilen gen- gen ilişkileri aynı zamanda gen ağı çıkarım algoritmaları tarafından da tahmin edilmektedir. Kısıtlı budama ile elde edilen gen- gen ilişkilerinin çeşitliliği fazla budamayla elde edilen gen- gen ilişki çeşitliliğinden daha yüksek olduğu gözlemlenmiştir.

6.1.2 ARNetMiT Yaklaşımının Topolojik Değerlendirmesi

Çalışmamızda ARNetMiT ve gen ağı çıkarım algoritmaları tabanlı GBA'ların performansları topolojik özelliklerine göre değerlendirilmiştir. Kesinlik değerlerinin yüksek olması, farklı ve fazla sayıda gen- gen ilişkisi içermesi açısından dGBA'lar (minD=0.1) topolojik değerlendirme aşamasında kullanılmıştır. Şekil 6.5'te GBA'ların ölçeksiz ağ yapısına uygunluk değerlerini gösteren R^2 değerleri bulunmaktadır.



Şekil 6. 5 GAÇ algoritmaları ve ARNetMiT tabanlı GBA'ların R^2 değerleri

Elde edilen R^2 değerlerine göre en yüksek değerler WGCNA tabanlı GBA'lar kullanılarak elde edilmiştir. WGCNA algoritması gen ağı çıkarımı sürecinde topolojik uygunluk değerini sağlayan bir eşik değeri kullanarak gen ağı çıkarımı yapmaktadır. Bu yaklaşım WGCNA tabanlı GBA'ların ölçeksiz ağ özelliklerini daha başarılı bir şekilde yansıtmalarına neden olmaktadır. C3NET tabanlı GBA'lara ait R^2 değerleri WGCNA algoritmasına ait R^2 değerlerini takip etmektedir. En düşük uygunluk değerleri ARACNE tabanlı GBA'lar ve dGBA'lar tarafından elde edilmiştir. Bu sonuçlara göre ARNetMiT tarafından üretilen dGBA'ların ölçeksiz ağ tanımına ait topolojik özellikleri çok yeterli yansıtmadığı anlaşılmaktadır.

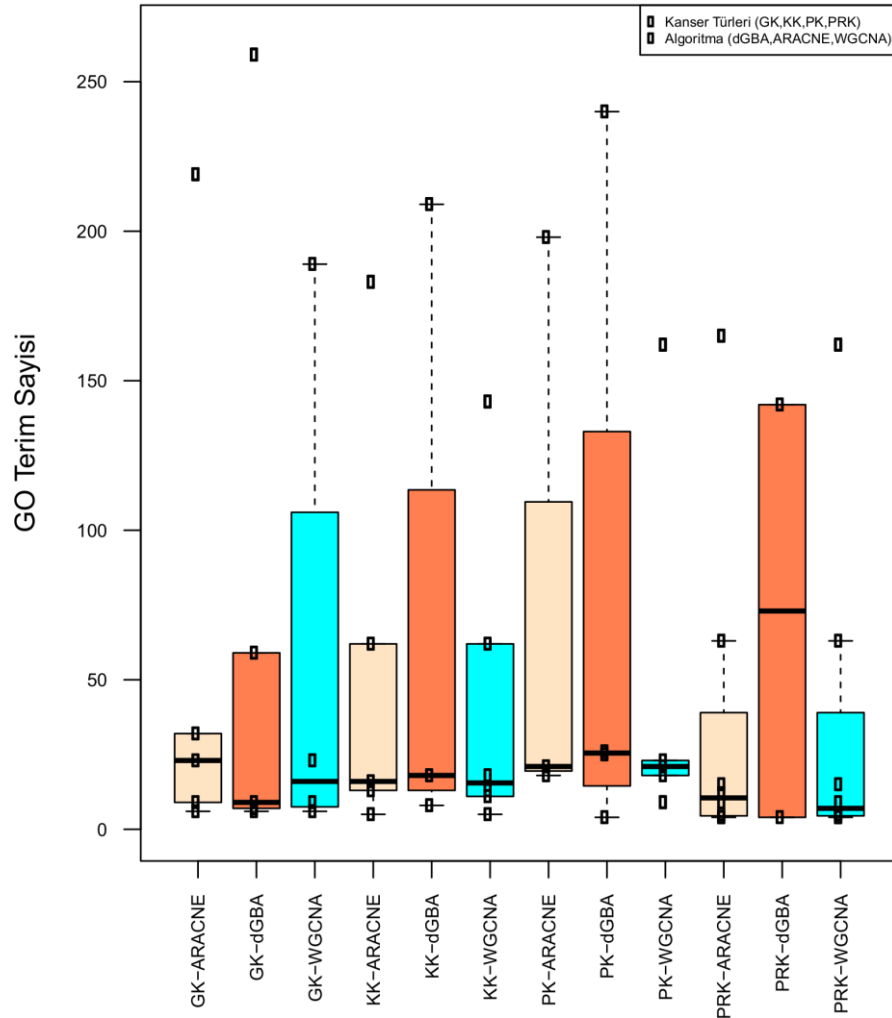
6.1.3 ARNetMiT Yaklaşımının Biyolojik Olarak Değerlendirilmesi

ARNetMiT'in tarafından oluşturulan GBA'ların biyolojik olarak değerlendirmesinin yapılabilmesi için ilk olarak MCODE algoritması ile ARNetMiT tabanlı GBA'lar modüllerine ayrılmıştır. Modüllerde yer alan genler GO zenginleştirme analizinde kullanılarak modüllerle ilişkili GO- terimleri elde edilmiştir. ARACNE, C3NET ve WGCNA algoritmaları tarafından üretilen GBA'lar da modüllerine ayrıldıktan sonra GO zenginleştirme analizi ile ilişkili olduğu GO- terimleri çıkarılmıştır. Dört farklı kanser türüyle ilişkili 4 farklı algoritma tarafından oluşturulan GBA'ların performans değerlendirmesi için OMIM hastalık veri kümesi kullanılmıştır. OMIM hastalık veri kümesi göğüs, prostat, kolon ve pankreas kanseriyle ilişkili genleri içermektedir. Bu

genler her bir hastalık için GO zenginleştirme analizinde kullanılarak dört hastalıkla ilişkili tüm GO- terimleri ayrı ayrı belirlenmiştir. ARNetMiT, ARACNE, C3NET ve WGCNA tabanlı dört hastalığa ait GBA'ların modülleriyle ilişkilendirilmiş GO- terimlerinin dört hastalıkla ilişkili GO- terimleriyle örtüşme analizi incelenmiştir ve biyolojik değerlendirme gerçekleştirilmiştir.

Çalışmamızda topolojik değerlendirmede de kullanılan dGBA'lar biyolojik değerlendirmede de kullanılmıştır. Şekil 6.6'da GBA'lara ait modüllerle ilişkilendirilmiş GO- terimlerinin hastalıkla ilişkili GO- terimleriyle örtüşme sayıları (DP) gösterilmiştir.

GO Terimleri ile Değerlendirme



Şekil 6. 6 GAÇ algoritmaları ve ARNetMiT tabanlı GBA'ların hastalıkla örtüşen GO-terimleri kullanılarak biyolojik olarak değerlendirilmesi

Şekil 6.6'daki göğüs kanseri *GK*, kolon kanseri *KK*, prostat kanseri *PK* ve pankreas kanseri *PRK* kısaltmalarıyla gösterilmiştir. C3NET algoritması kullanılarak elde edilen GBA'lar MCODE algoritması uygulandığında algoritma kriterlerini sağlayan modüller elde edilmediği için sonuçlarda C3NET tabanlı GBA sonuçları yer almamaktadır. GBA'lar MCODE algoritmasıyla birden fazla modüle ayrılabilir. Bir GBA'ya ait olası çok sayıda modülün performansının gösterimi için Şekil 6.6'daki gibi kutu grafiğinden (box plot) faydalanılmıştır. Sonuçlara göre dGBA'lara ait modüllerin ilişkili olduğu GO-terimleri, göğüs kanseri dışındaki diğer üç kanser türüyle ilişkili olan GO-terimleriyle daha çok örtüştüğü gözlemlenmiştir. Özellikle kolon ve prostat kanseriyle verileri kullanılarak oluşturulmuş dGBA'larda yer alan bazı modüllerdeki genler GO zenginleştirme analizi kullanıldığında 200'den fazla hastalıkla örtüşen GO-terimi elde edilmiştir. Örtüşme analizinde istatistiksel olarak anlamlı örtüşen GO-terimlerinin incelenmesi için Fisher Kesinlik Testi kullanılmıştır. Biyolojik değerlendirme sonuçlarına göre ARNetMiT kullanılarak elde edilen GBA'ların hastalıklarla ilgili biyolojik süreç ve olayları başarılı bir şekilde yansıttığı deney sonuçlarında elde edilmiştir.

6.2 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Performans Değerlendirilmesi

Çalışmamızda prostat ve göğüs kanseriyle ilişkili mikrodizin gen ifadesi, RNA-Seq ve miRNA- hedef gen veri kümeleri kullanılarak iki farklı kanser türüne ait GBA'lar tez kapsamında geliştirdiğimiz iki katmanlı mimariyle entegre edilmiştir. Bu bölümde entegrasyon çalışmasında kullanılan veri kümelerine ait bilgiler, veri kümelerinin normalizasyon adımları, entegrasyon sonucunda elde edilen GBA'lar örtüşme analizi, topolojik ve biyolojik değerlendirme kriterlerine göre performans değerlendirmesine yer verilmiştir.

6.2.1 Entegrasyon Çalışmasında Kullanılan Prostat ve Göğüs Kanseri Ait Veri Kümeleri

Çalışmamızda prostat kanseriyle ilişkili mikrodizin gen ifadesi verileri 69 prostat kanseri tümör örneğinden oluşmuştur. Her bir örnek 22,277 gen parçasını (probe) içermektedir. Gen parçaları Affymetrix HGU133PLUS2 mikrodizin tipi kullanılarak

oluşturulmuştur ve NCBI veri tabanı sayfasından GSE6956 numarasıyla erişilmektedir. Göğüs kanseriyle ilişkili veriler 56 göğüs kanseri tümör örneğinden oluşmakta ve her örnek prostat kanserinde olduğu gibi Affymetrix HG-U133PLUS2 mikrodizin tipi kullanılarak elde edildiği için 22,277 gen içermektedir. Göğüs kanserine ait mikrodizin gen ifadesi verilerine NCBI ulaşım numarası GSE10797'dir.

Prostat kanserine ait RNA- Seq verileri TCGA (The Cancer Genome Atlas)[81] projesinden elde edilmiştir. Bu veriler 495 prostat kanserine ait örnekten alınmıştır ve her bir örnek 20,502 gen parçası içermektedir. Göğüs kanseriyle ilişkili RNA- Seq verileri ise 39 göğüs tümör örneğinden alınmıştır ve her bir örnek Illumina HiSeq 2500 dizileme teknolojisiyle üretildiği için 57,736 gen içermektedir. Bu genlerden okuma değerleri sıfırdan büyük olan genler sadece gen ağı çıkarımında kullanıldığında geriye 15,830 gen kalmıştır. Göğüs kanserine ait RNA- Seq verilerinin NCBI ulaşım numarası GSE68086'dır.

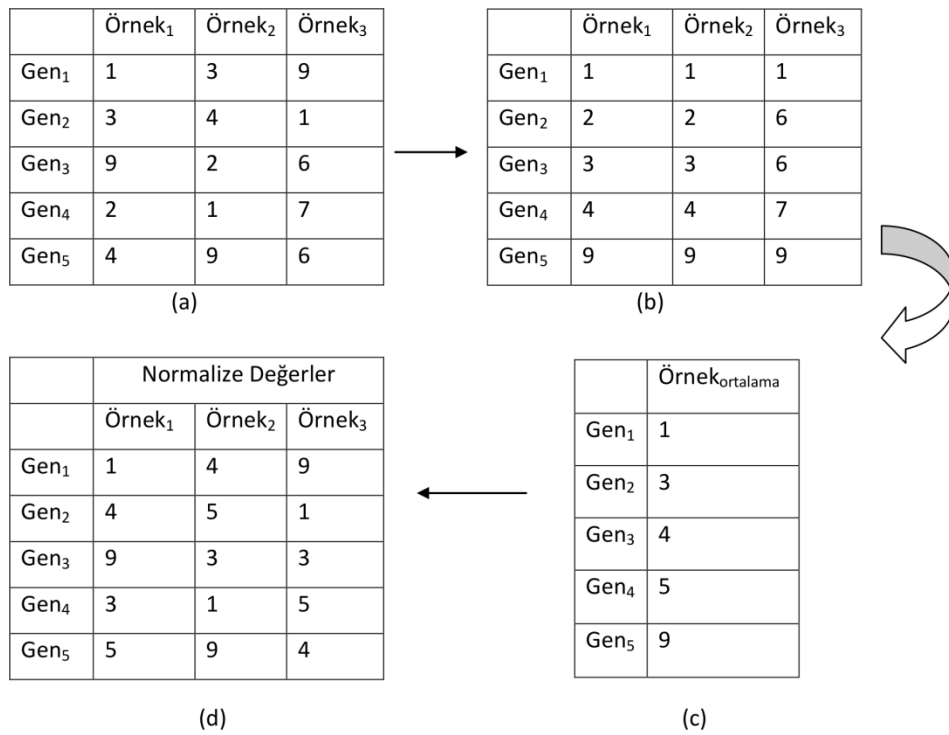
Çalışmamızda kullanılan prostat ve göğüs kanserine ait miRNA- hedef gen verileri ARNetMIT R paketi kullanılarak elde edilmiştir. Çizelge 6.1'de gösterildiği üzere prostat kanserine ait miRNA- hedef gen ilişki sayısı 3232 iken, göğüs kanserine ait miRNA- hedef gen ilişki sayısı 2179'dur.

6.2.2 Entegrasyon Çalışmasında Kullanılan Prostat ve Göğüs Kanserine Ait Veri Kümelerinin Normalizasyonu

Ham veri kümeleri temel olarak deney süreçlerinde ortaya çıkan gürültüler ve farklı koşullarda elde edilen verilerin birlikte kullanımı nedeniyle veri kümelerine ön işleme adımları uygulandıktan sonra GBA çıkarımı yapılmaktadır. Çalışmamızda mikrodizin gen ifadesi veri kümeleri için dağılım normalizasyonu (quantile normalization) yöntemi, RNA- Seq verileri için ise M- Değerlerinin Düzenlenmiş Ortalaması (Trimmed Mean of M- values, TMM) yöntemi kullanılmıştır. miRNA- hedef gen verileri deneysel olarak ispatlanmış ikili ilişkileri içerdiği için herhangi bir normalizasyon bu verilere uygulanmamıştır. Veriler normalize edildikten sonra gen ağı çıkarım algoritmaları uygulanarak GBA oluşumu gerçekleştirilmiş olur.

Dağılım Normalizasyonu: Mikrodizin gen ifadesi verileri üzerinde normalizasyon uygulanmasının en büyük nedeni biyolojik temeli olmayan ve deney sürecinde ortaya

çıkan değişkenlerin oluşturduğu hataların minimize edilme isteğidir. Normalizasyon sonucunda aykırı verilerin (outlier) yoğunluk değerleri de aynı zamanda belirlenmiş olur. Aykırı veri keşfi biyolojik olarak önem taşımaktadır. Normalizasyon ile farklı tip hücrelerden (normal- kanserli hücre gibi) farklı zamanlarda alınmış mikrodizinlerdeki değerler karşılaştırılabilir hale gelmektedir. Farklı mikrodizinlerin bir arada kullanımı için normalizasyon uygulanması önem taşımaktadır. Çalışmamızda mikrodizin gen ifadesi verilerinin normalizasyonu için dağılım normalizasyonu (quantile normalization) yöntemi kullanılmıştır. Şekil 6.7'de dağılım normalizasyonun adımları bir örnek üzerinde gösterilmiştir. Şekil 6.7a'da genler ve genlerin üç örnekte aldığı yoğunluk değerleri verilmiştir. Şekil 6.7b'de ise bu değer sütun bazlı (örnek bazlı) sıralanmıştır. Bir sonraki adımda sütun bazlı değerlerin ortalaması alınmıştır. Ortalama değerleri sırayla Şekil 6.7.a'daki yoğunluk değerlerinin yerine sıralama gözeterek yerleştirilmiştir.



Şekil 6. 7 Dağılım normalizasyonu

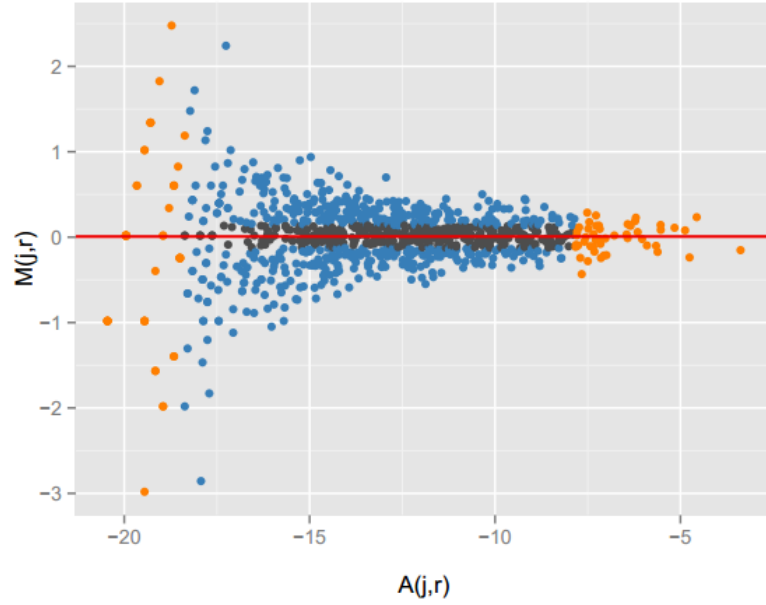
M- Değerlerinin Düzenlenmiş Ortalaması: RNA- Seq verileri referans genome üzerinde örtüşen gen parçalarının (transkript) okuma sayısını içermektedir. RNA- Seq verileri üzerinde iki temel durum nedeniyle normalizasyon işleminin yapılması gerekir. Bunlardan ilki okuma sayılarının örnekten örneğe değişmesi (between lane normalization) diğer neden ise uzun gen dizilimine sahip genlerin okuma sayısının

daha fazla olmasıdır (within-lane normalization) [82]. Çalışmamızda RNA-Seq verilerinin normalizasyonu için literatürde en iyi sonuçları veren normalizasyon yöntemlerinden olan M- Değerlerinin Düzenlenmiş Ortalaması (Trimmed Mean of M-values, TMM) kullanılmıştır [83], [84]. MDDO örnekteki toplam okuma sayısının büyük kısmı az sayıda gene aittir. Bu durum az sayıda okuması olan genlerin arasından fark yaratan genlerin bulunması (differentially expressed genes) ve bu genler arasındaki gen ağı oluşturmada normalizasyon aşamasında probleme neden olmaktadır. Bu problemi gidermek için MDDO, düzenlenmiş ortalama değerleri önceden belirlenen oranlarda (%x) üst ve alt okuma sayısı içeren genleri eleyerek normalizasyon işlemini gerçekleştirir. MDDO bunun için Log- Örnek Değişikliği (Log Fold Change, M- değerleri) ve Mutlak İfade Seviyesi (Absolute Expression Level, A- değeri) gibi iki metriği kullanmaktadır. Bir genin M- değeri, bir örneğin referans örneğe göre okuma sayısındaki değişimini gösterir. A- değeri ise genin örnek ve referans değerdeki ortalama değerini vermektedir.

$$M_g(j, r) = \log_2 \left(\frac{K_{gj}}{D_j} \right) - \log_2 \left(\frac{K_{gr}}{D_r} \right)$$

$$A_g(j, r) = 0.5 * (\log_2 \left(\frac{K_{gj}}{D_j} \right) + \log_2 \left(\frac{K_{gr}}{D_r} \right)) \quad (6.1)$$

K_{gj} ve K_{gr} , g geninin sırasıyla j örneğinde ve r referans örneğindeki okuma sayısını göstermektedir. Veri kümesindeki tüm örneklerin içerdiği okuma sayıları kutu grafiğinde gösterilir ve üçüncü çeyrek değerine en yakın okuma değerine sahip örnek referans örnek (örnek r) olarak belirlenir. D_j ve D_r , j örneğinde ve r referans örneğindeki okuma sayısını göstermektedir. $M_g(j, r)$, g geninin örnek r 'ye göre örnek j 'deki M- değerini gösterirken, $A_g(j, r)$, g geninin örnek r 'ye göre örnek j 'deki A- değerini göstermektedir. Örnek j 'de yer alan genlere ait M ve A- değerleri elde edilir. M- değerlerinin %30'luk alt ve üst değer diliminde yer alan değerlere sahip genler ve A- değerlerinin %5'lik alt ve üst değer diliminde yer alan değerlere sahip genlerden ortak olan genler elenir. Şekil 6.8'de MDDO normalizasyonu ile %30-%5 kriterlerine göre gerçekleştirilen gen eleme işlemi gösterilmiştir [83], [85].



Şekil 6. 8 MDDO normalizasyon örneği [85]

$$w_g(j, r) = \left(\frac{D_j - K_{gj}}{D_j K_{gj}} + \frac{D_r - D_{gr}}{D_r D_{gr}} \right)$$

$$f_j = \frac{\sum_{g \in G} w_g(j, r) \cdot M_g(j, r)}{w_g(j, r)} \quad (6.2)$$

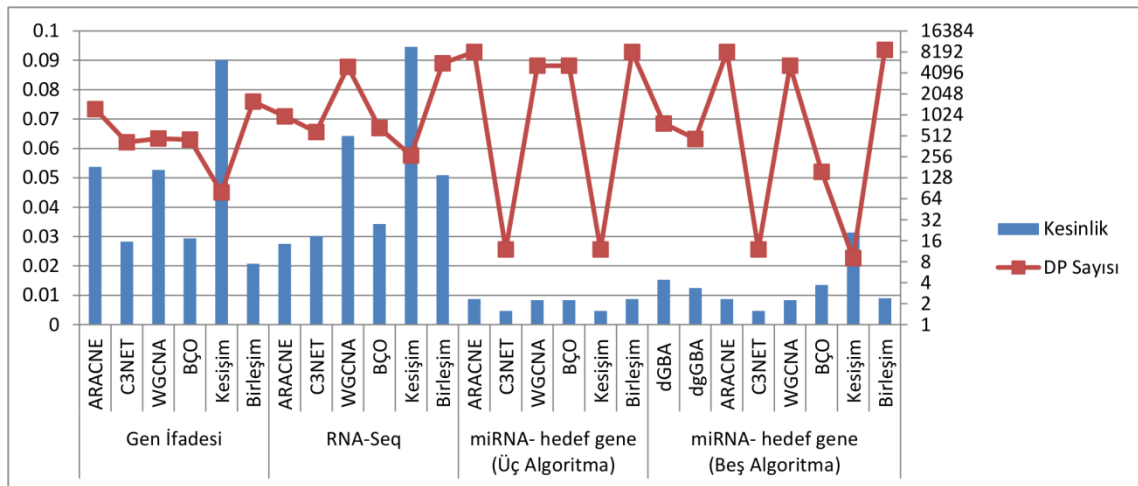
f_j değeri j örneğinin düzeltme değeri olarak denklem 6.2'de gösterildiği gibi hesaplanır. Bu örnekteki tüm genlere ait okuma değerleri $\sqrt{f_j}$ değeri ile çarpılır ve referans örnek için tekrar denklem 6.2'deki eşitlikle hesaplanan $\sqrt{f_r}$ değerine bölünür. Son adımdan sonra genlerin j örneğindeki normalize edilmiş değerleri elde edilir. Normalizasyon işlemi gerçekleştirilmiş G gen kümesi M ve A - değer elemelerini geçen genleri gösteren kümedir ve sadece bu kümede yer alan genlere ait okuma değerleri düzeltilir. Çalışmamızda MDDO normalizasyonu için *edgeR* R [86] paketinden yararlanılmıştır.

6.2.3 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Örtüşme Analizi Değerlendirilmesi

Bu bölümde iki katmanlı GBA entegrasyon mimarisinde yer alan başlangıç, birinci ve ikinci entegrasyon aşamalarında elde edilen GBA'lar örtüşme analiziyle performansları değerlendirilmiştir. Çalışmamızda kullandığımız prostat ve göğüs kanseriyle ilişkili her

bir aşamada elde edilen GBA'lar birlikte değerlendirilmiş ve iki kanser türüne ait sonuçlar karşılaştırılmıştır. İlk olarak iki kanser türüne ait başlangıç aşaması ve birinci entegrasyon aşamasında elde edilen GBA'ların performansları değerlendirilmiştir. Başlangıç aşamasında elde edilen dokuz GBA'nın örtüşme analizi değerlendirmesiyle gen ağı çıkarım algoritmalarının performansları her üç veri kümesine göre karşılaştırılmıştır. Başlangıç aşamasından elde edilen sonuçları kullanarak aynı zamanda üç farklı biyolojik veri kümesinin performansları da gen ağı çıkarım algoritmalarının bu veri kümeleri üzerindeki başarı değerlerine bakılarak kıyaslanması gerçekleştirilmiştir. Başlangıç aşaması ve birinci entegrasyon çalışması sonucunun aynı grafikler üzerinde gösterilmesiyle birinci entegrasyon aşamasının performansa etkisi gözlemlenmeye çalışılmıştır.

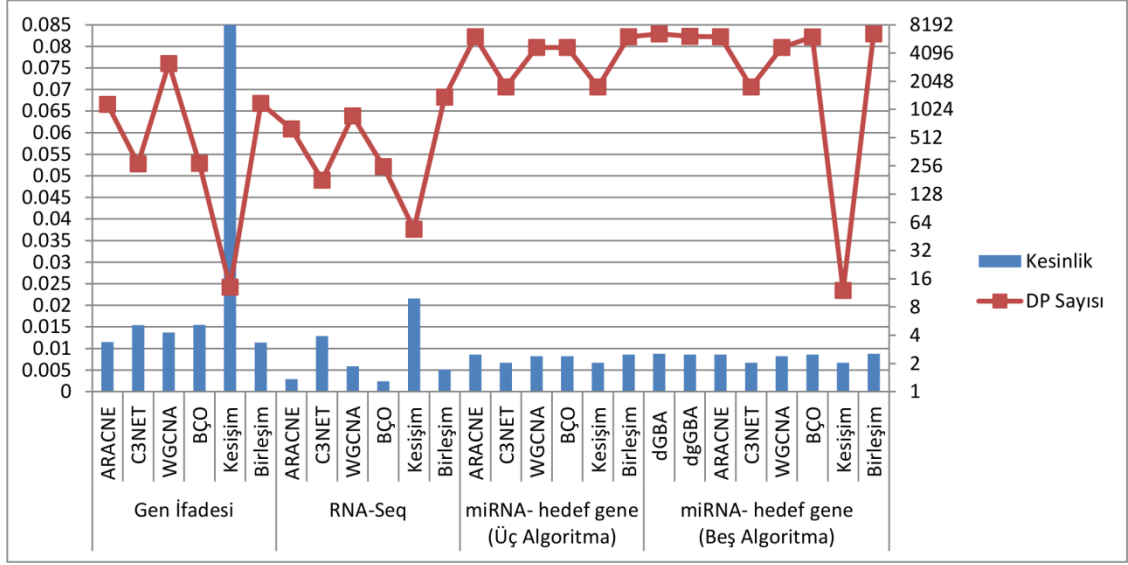
Mikrodizin gen ifadesi ve RNA- Seq verilerinde üç gen ağı çıkarım algoritması uygulayarak üç farklı GBA elde edilirken, miRNA- hedef gen verilerine ARNetMiT tabanlı GBA'lar da eklenerek beş GBA entegrasyon aşamasında kullanılmıştır. ARNetMiT bazlı GBA'lar kullanarak ve kullanmayarak entegrasyon işlemi ayrı ayrı gerçekleştirilmiş ve ARNetMiT tabanlı GBA'ların entegrasyona etkisi gözlemlenmiştir. Şekil 6.9'da prostat kanseriyle ilişkili başlangıç ve birinci entegrasyon aşamalarından sonra elde edilen GBA'ların örtüşme analizi sonuçları sunulmuştur.



Şekil 6. 9 Prostat kanserine ait başlangıç ve birinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları

Her üç veri kümesinden elde edilen sonuçlara göre tüm gen ağı çıkarım algoritmaları en düşük kesinlik değerlerini miRNA- hedef gen verilerini kullanarak elde etmiştir. Bu

nedenle örtüşme analizine göre değerlendirme yapıldığında en başarısız veri kümesi miRNA- hedef gen veri kümesi olarak gözlemlenmektedir. Mikrodizin gen ifadesi ve RNA-Seq verileri kullanılarak elde edilen GBA'ların örtüşme sonuçları ise benzerlik göstermektedir. Gen ağı çıkarım algoritmalarının sonuçları karşılaştırıldığında C3NET algoritmasının gen ifadesi ve miRNA- hedef gen verileri üzerinde en düşük kesinlik değerlerini elde ettiği gözlemlenmektedir. Sadece RNA- Seq verisi üzerinde C3NET algoritmasının ARACNE'den daha iyi performans gösterebilmiştir. Aynı zamanda C3NET algoritmasının yapısı gereği ilişki tahmin sayısının da diğer algoritmalarından az olduğu DP değerlerine bakılarak anlaşılmaktadır. Birinci entegrasyon sonuçları aynı veri kümesi üzerinde farklı gen ağı çıkarım algoritmaları tarafından oluşturulan GBA'ların entegrasyonu olarak tanımlandığı için her bir veri seti için ayrı ayrı Şekil 6.9'da gösterilmiştir. Buna göre kesişim metoduyla elde edilen GBA'ların içerdiği gen ilişkisi sayısı az dahi olsa kesinlik değerleri başlangıç aşamasında yer alan GBA'lardan daha yüksek çıktığı gözlemlenmiştir. Diğer iki algoritma ise başlangıç aşamasındaki performans değerlerini genel olarak yükseltmemiştir. Birleşim metoduyla entegrasyon sonucunda beklenildiği üzere GBA'larda yer alan gen ilişkisi sayısının arttığı gözlemlenmiştir. miRNA- hedef gen verileri üzerinde birinci entegrasyon aşamasında ARNetMiT tabanlı GBA'ların entegrasyon işlemine katılmasıyla kesinlik değerlerinde az da olsa bir artış olduğu görülmektedir. Şekil 6.9'da RNA- Seq verileri üzerinde birinci entegrasyon sonucunda BTGBA'nın kesinlik sonucunun BÇOGBA'dan yüksek çıktığı görülmüştür. Bunun sebebi en az iki tane GBA tarafından ortak olarak tahmin edilen gen-gen ilişkilerinin genellikle yanlış pozitif tahmin olarak değerlendirilmesidir. KTGBA'ların aynı veri kümesi üzerinde en yüksek kesinlik değeri alıp en düşük değerlerin BÇOGBA'dan elde edilmesi aynı zamanda üç GBA tarafından ortak tahmin edilen gen ilişkilerinin doğru pozitif olmasından kaynaklanır.



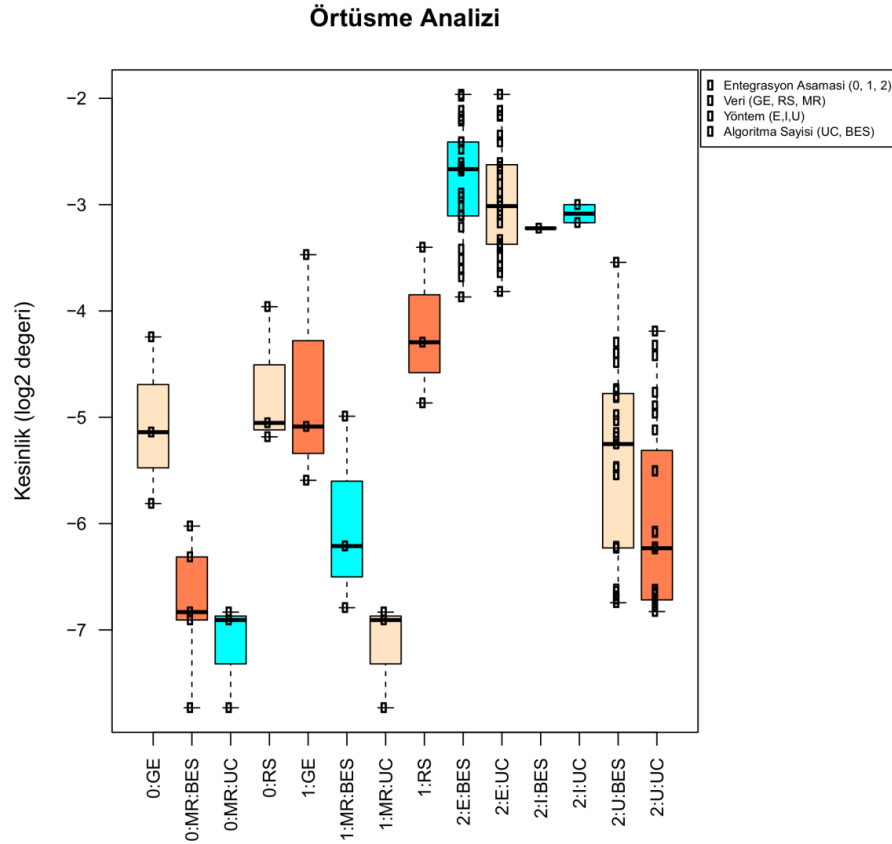
Şekil 6. 10 Göğüs kanserine ait başlangıç ve birinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları

Şekil 6.10'da göğüs kanserine ait başlangıç ve birinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizin sonuçları gösterilmiştir. Prostat kanserinden farklı olarak miRNA- hedef gen verileri kullanılarak elde edilen GBA'ların performans değerleri RNA- Seq verileri kullanılarak elde edilen GBA'ların performans değerlerine yakın çıktığı gözlemlenmektedir. Şekil 6.9'dan farklı olarak C3NET algoritmasının kesinlik sonuçları diğer algoritmalara çok yakın çıkmasına rağmen en kısıtlı ilişki tahmin gücü yine C3NET tabanlı GBA'lardan elde edilmiştir. Birinci entegrasyon sonuçlarında kesişim metodunun prostat kanseri sonuçlarında olduğu gibi performansı arttırıp, tahmin gücünü azalttığı gözlemlenmiştir. Bunun dışında birinci entegrasyon seviyesinde BÇO ve birleşim tabanlı yaklaşımların göğüs kanseri verileri üzerinde de performansı arttırmadığı görülmüştür.

Birinci entegrasyon sonucunda elde edilen performans değerlerinin başlangıç seviyesindeki performans değerlerini net bir biçimde arttıramamasının en büyük sebebi grup içi entegrasyondur. Grup içi entegrasyon sonucunda GBA'lardaki farklı gen ilişki sayısı az olduğu için entegrasyon metotları yeterince performansları arttıramamıştır.

İkinci entegrasyon aşamasında olası 81 GBA oluşturulması nedeniyle başlangıç ve birinci entegrasyon aşamalarında elde edilen GBA'lar ile ikinci entegrasyon aşamasında elde edilen GBA'ların karşılaştırılması için kutu grafiğinden

yararlanılmıştır. Şekil 6.11'de prostat kanseriyle ilgili tüm aşamalarda elde edilmiş GBA'ların kesinlik değerlerini sunan kutu grafiği gösterilmiştir.

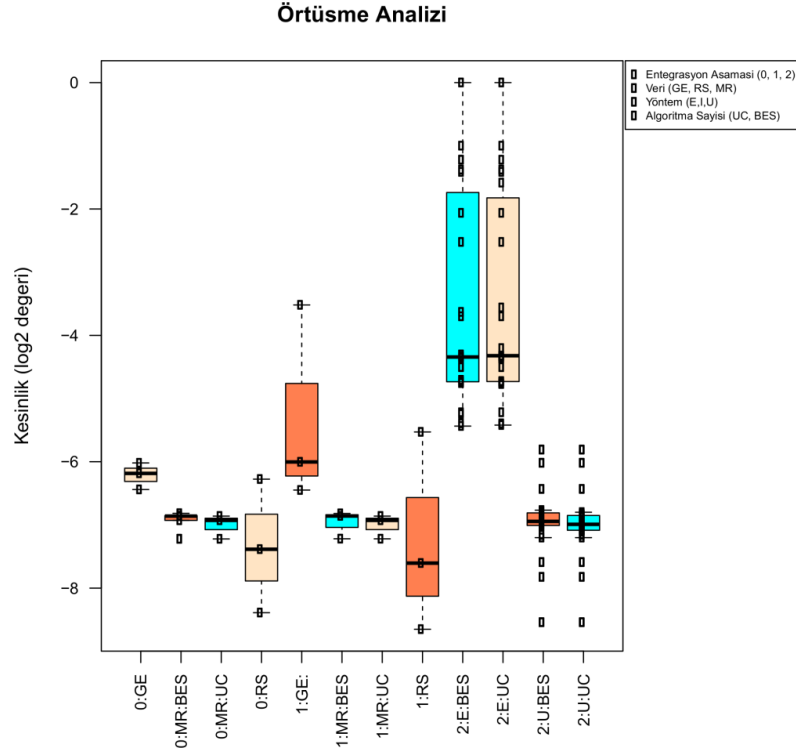


Şekil 6. 11 Prostat kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları

Kesinlik değerleri küçük değerler olması nedeniyle kutu grafiğindeki kesinlik değerlerinin $\log_2$ tabanında karşılıkları Şekil 6.11'de gösterilmiştir. Kutu grafiğinde yer alan sıfıncı düzey (başlangıç düzeyi) ve birinci entegrasyon düzeyindeki sonuçlar Şekil 6.9 ile aynı olduğu gözlemlenmektedir. Şekil 6.11'de yer alan etiketlerde *E* harfi basit çoğunluk oylaması işlemini, *I* harfi kesişim, *U* harfi ise birleşim metotlarını göstermektedir. Etiketlerde yer alan *UC* (*ARNetMiT* bazlı GBA'lar dahil edilmemiş) ve *BES* (*ARNetMiT* bazlı GBA'lar dahil edilmiş) miRNA-hedef gen verileri üzerinde elde edilen GBA'ların sayısını göstermektedir. Şeki 6.11'e göre en başarılı kesinlik değerleri ikinci aşama entegrasyondan sonra elde edilen GBA'lardan elde edilmiştir.

Birinci entegrasyon aşamasındaki BÇO tabanlı GBA'ların örtüşme analizi performansları başlangıç aşamasına yakın iken, ikinci entegrasyon aşamasında BÇO yaklaşımı kullanılarak elde edilen GBA'ların en yüksek performans değerlerini sağlamaktadır.

İkinci entegrasyon aşamasındaki kesişim tabanlı GBA'larında kesinlik değerlerini BÇO tabanlı GBA'ların kesinlik değeri takip etmiştir. İkinci aşamada kesişim tabanlı GBA sayısı sıkı eliminasyon yaklaşımından dolayı az sayıda oluşturulmuştur. miRNA üzerinde entegrasyon aşamasında kullanılan ARNetMiT tabanlı GBA'ların kısıtlı da olsa performansı arttırdığı Şekil 6.11'de de belirtilmiştir.



Şekil 6. 12 Göğüs kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların örtüşme analizi sonuçları

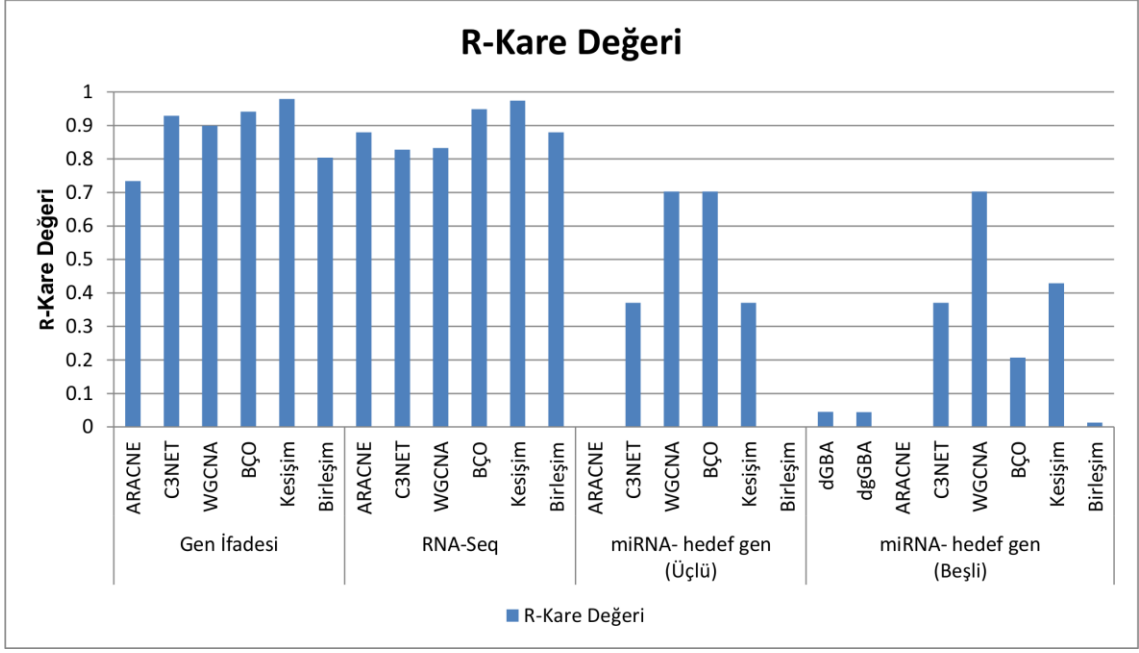
Şekil 6.12'de göğüs kanserine ait tüm aşamalarda elde edilen GBA'ların örtüşme analizi sonucunda elde ettiği kesinlik değerleri gösterilmiştir. Şekil 6.10'da başlangıç ve birinci entegrasyon adımıyla elde edilen sonuçlar Şekil 6.12'de tekrarlanmıştır. Prostat kanseri ile ilgili GBA'lardaki sonuçlara benzer örtüşme analizi sonuçları göğüs kanseri verisi üzerinde de elde edilmiştir. İkinci entegrasyon adımıyla BÇO metoduyla elde edilen GBA'ların kesinlik değerleri tüm aşamalarda elde edilen GBA'ların kesinlik değerlerinden daha yüksek olarak gözlemlenmektedir. Göğüs kanserine ait veriler kullanılarak oluşturulan GBA'ların ikinci entegrasyon adımıyla kesişim entegrasyon metodu uygulandığında herhangi bir GBA elde edilmemiştir. Buradan da üç farklı veri seti üzerinde oluşturulan GBA'ların içerdiği gen ilişkilerinin farklı olduğu sonucu

çıkmaktadır. Göğüs kanserine ait miRNA- hedef gen verileri üzerinde GBA sayısının artmasının kesinlik değerini arttırmadığı sonucuna varılmıştır.

Örtüşme analizinin sonucunda grup içi entegrasyonun performans arttırımında çok etkili olmadığı ama gruplar arası entegrasyonun GBA'ların performans değerlerini arttırdığı görülmüştür. Gruplar arası entegrasyon sonucunda elde edilen performans artışında farklı biyolojik veri kümeleri kullanılarak oluşturulan GBA'ların içerdiği gen ilişkilerinin değişkenliğinin yüksek olması büyük önem taşımaktadır.

6.2.4 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Topolojik Özelliklerinin Değerlendirilmesi

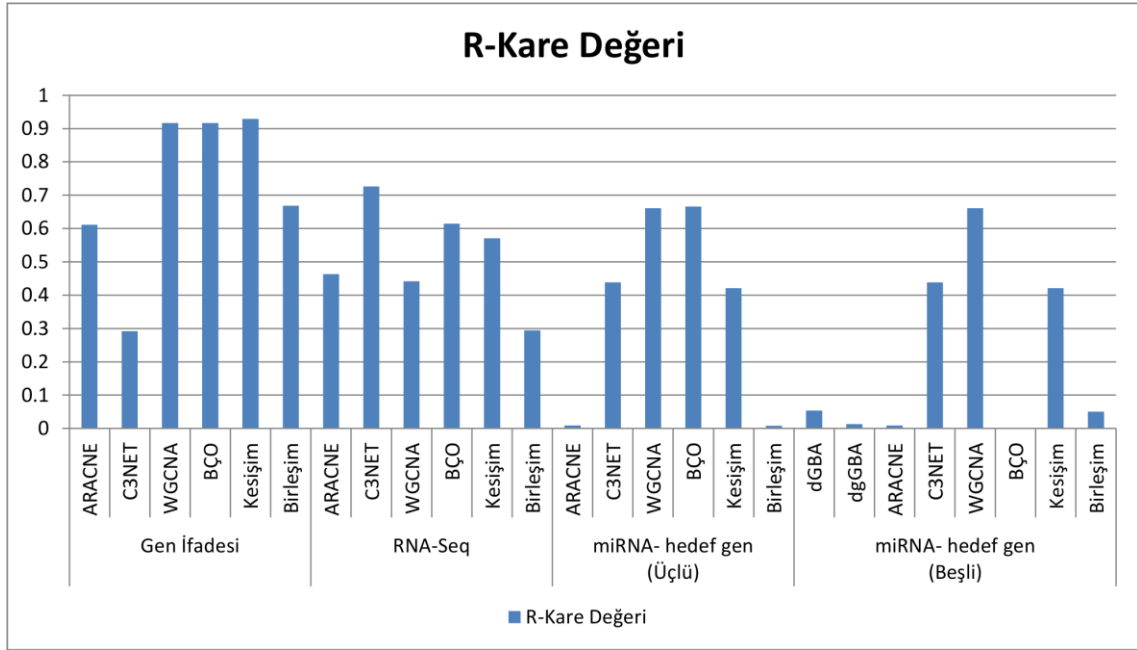
İki katmanlı entegrasyon mimarisiyle elde edilen GBA'ların ölçeksiz ağ özelliklerini yansıtmaya gücü, uygunluk değeri olan belirleme katsayısı (R^2) ile ölçülmüştür. R^2 değerinin bire yakın olması GBA'ların ölçeksiz ağ tanımını iyi yansıttığını belirtir. Belirleme katsayısı sıfır veya sıfıra yakın ise bu ağlar rassal oluşturulan ağ özellikleri gösterir. Bu bölümde prostat ve göğüs kanseriyle ilişkili iki katmanlı entegrasyon mimarisi ile elde edilen GBA'ların topolojik olarak değerlendirmesi yapılmıştır. Örtüşme analizinde olduğu gibi ilk olarak başlangıç aşaması ve birinci entegrasyon aşaması sonuçları birlikte verilmiş ve daha sonra tüm aşamalarda elde edilen GBA'ların R^2 değerleri birbirleriyle karşılaştırılmıştır. Şekil 6.13'te prostat kanseri verileri kullanılarak başlangıç ve birinci entegrasyon aşamalarından elde edilmiş GBA'ların R^2 değerleri gösterilmiştir.



Şekil 6. 13 Prostat kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların R^2 değerleri

Elde edilen sonuçlara göre başlangıç aşamasında WGCNA ve C3NET algoritmasıyla elde edilen GBA'lara ait R^2 değeri bire yakın çıkmıştır. Bu iki algoritmanın özellikle mikrodizin gen ifadesi ve RNA- Seq verileri üzerinde ürettiği GBA'ların ölçeksiz ağ modeline uygun olduğu Şekil 6.13'te gösterilmiştir. En düşük R^2 değerleri miRNA- hedef gen verileri kullanıldığında elde edilmiştir. ARNetMiT tabanlı dGBA ve dgGBA'ların R^2 değerleri ARACNE tabanlı GBA'lardan yüksek C3NET ve WGCNA tabanlı GBA'lardan düşük çıkmıştır. Birinci entegrasyonun topolojik özelliklere etkisine baktığımızda ise mikrodizin gen ifadesi ve RNA- Seq verilerinde BÇO ve kesişim bazlı entegrasyonun R^2 değerini arttırdığı gözlemlenmiştir. miRNA- hedef gen verilerinde kesişim ve BÇO tabanlı entegrasyonun R^2 değerini arttırmadığı görülmektedir. Bunun nedeni Şekil 6.4'de gösterilen tüm gen ağı çıkarım algoritmalarının miRNA- hedef gen verileri üzerinde benzer ilişki tahmin etmesine dayanmaktadır. Bu durum aynı zamanda miRNA- hedef gen üzerinde elde edilen GBA sayısının artmasına rağmen R^2 değerinin artmamasının da nedenidir. Birinci entegrasyon aşamasında BTGBA'lar üç gen ağı çıkarım algoritması tarafından sadece tahmin edilen tüm gen- gen ilişkilerini kullanarak ağ çıkarımı gerçekleştirdiği için bu ilişkiler arasından merkez genlere ait ilişkilerinin oranı azalmakta bu durum da R^2 değerinin düşmesine neden olmaktadır.

Şekil 6.14'te ise göğüs kanseri verileri kullanılarak başlangıç ve birinci entegrasyon aşamalarından elde edilmiş GBA'ların R^2 değerleri gösterilmiştir.

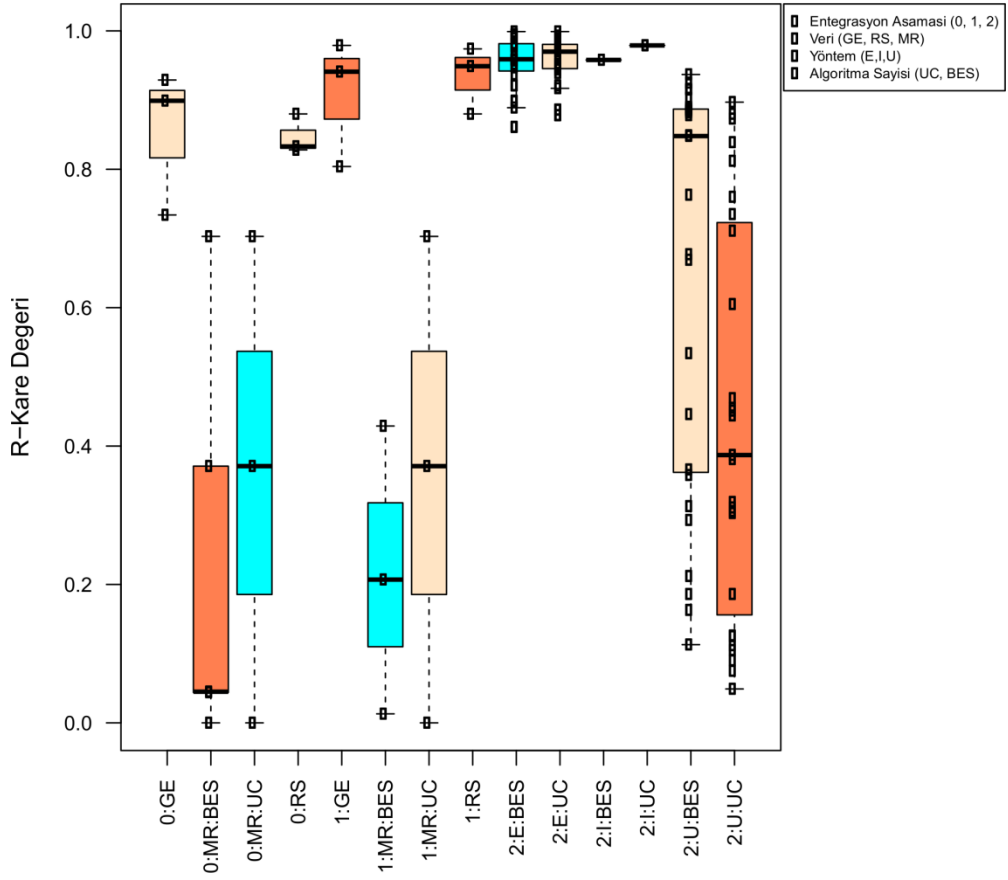


Şekil 6. 14 Göğüs kanserine ait başlangıç, birinci ve ikinci entegrasyon aşamasında elde edilen GBA'ların R^2 değerleri

Şekil 6.14'te belirtilen göğüs kanseri verileri kullanılarak başlangıç ve birinci entegrasyon aşamasında oluşturulan GBA'lara ait R^2 sonuçları prostat kanseri verileri kullanılarak elde edilen sonuçlarla paralellik göstermektedir. Genel olarak WGCNA ve C3NET algoritmaları kullanılarak elde edilen GBA'ların R^2 değerleri daha yüksek çıkmıştır. En düşük R^2 sonuçları yine miRNA- hedef gen verileri kullanılarak elde edilen GBA'lar tarafından oluşturulmuştur. Birinci entegrasyon aşamasında prostat kanseri verilerinde olduğu gibi BÇO ve kesişim tabanlı GBA'ların R^2 değerleri başlangıç aşamasındaki GBA'ların R^2 değerinden yüksek çıkmıştır. Birinci entegrasyon aşamasında birleştirme metodu kullanıldığında ise zayıf gen ilişkisi eleme kriterinden dolayı oluşturulan BTGBA'ların R^2 değeri göğüs kanseri verilerinde de düşük çıkmıştır.

Prostat kanseriyle ilgili başlangıç, birinci ve ikinci entegrasyon aşamalarında elde edilen GBA'lara ait R^2 değerleri kutu grafiği olarak 6.15'te gösterilmiştir.

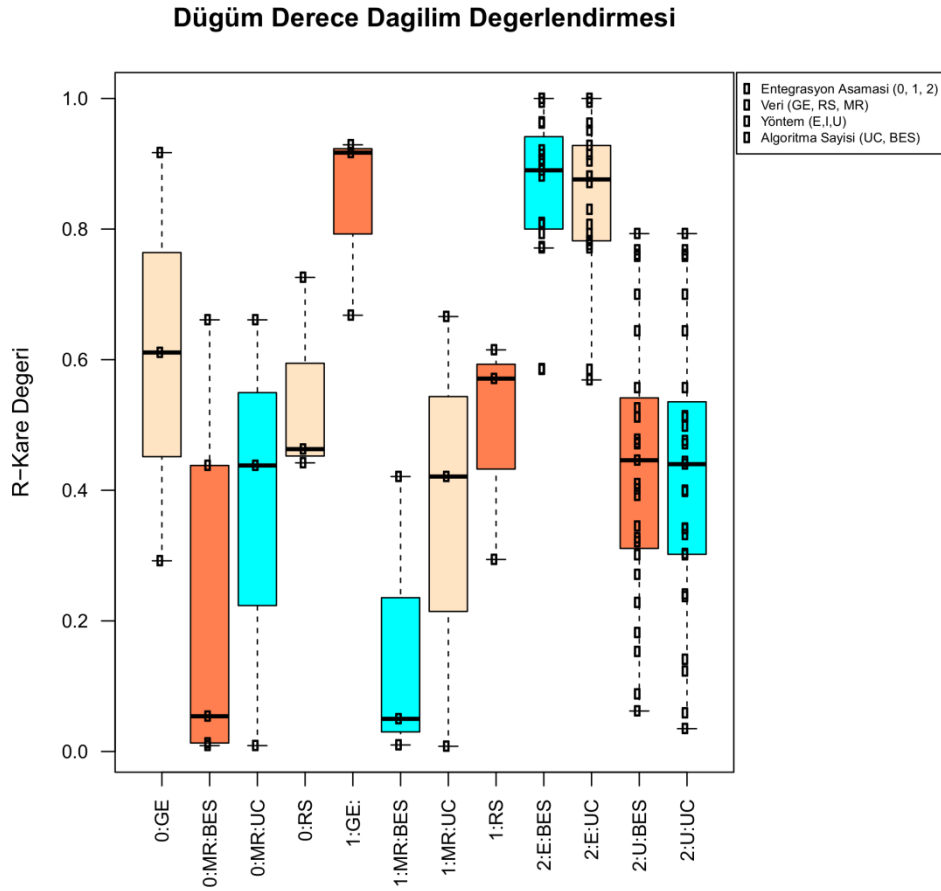
Düğüm Derece Dağılım Değerlendirmesi



Şekil 6. 15 Prostat kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların R^2 değerleri

Başlangıç ve birinci entegrasyon aşamasına ait sonuçları gösteren Şekil 6.13'te elde edilen sonuçlar Şekil 6.15'te kutu grafiğinde de gösterilmiştir. Şekil 6.15'te miRNA-hedef gen verileri üzerinde birinci entegrasyon aşamasında üç GBA'nın oluşturduğu kutu grafiğinin medyan değerinin beş GBA'nın oluşturduğu kutu grafiğinin medyan değerinden fazla olduğu gözlemlenmektedir. Bunun sebebi dGBA ve dgGBA'lardan elde edilen düşük R^2 değerleridir. İkinci entegrasyon aşamasında BÇO ve kesişim metotlarıyla oluşturulan GBA'ların R^2 değerlerinin bire çok yakın olduğu, birleşim metoduyla elde edilenlerin R^2 değerlerinin medyan değerlerinin ARNetMiT bazlı GBA'lar kullandığında 0.8'e yaklaştığı görülmektedir. İkinci entegrasyon aşamasında, ARNetMiT tabanlı GBA'ların entegrasyonda kullanımıyla özellikle birleşim GBA'ların R^2 değerlerinin belirgin bir şekilde arttığı gözlemlenmiştir.

Göğüs kanseriyle ilgili başlangıç, birinci ve ikinci entegrasyon aşamalarında elde edilen GBA'lara ait R^2 değerleri kutu grafiği olarak 6.16'da gösterilmiştir.



Şekil 6. 16 Göğüs kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların R^2 değerleri

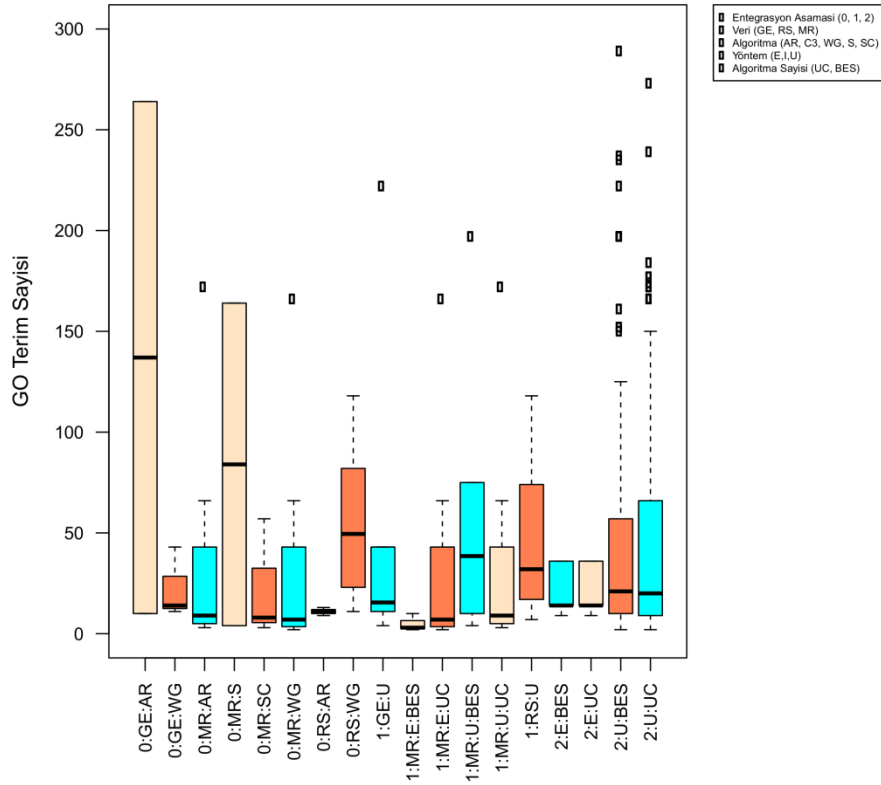
Göğüs kanserine ait tüm entegrasyon aşamalarından çıkarılan GBA'ların R^2 değerleri prostat kanseriyle ilişkili GBA'lardan çıkartılan R^2 değerlerinden daha düşük olduğu gözlemlenmektedir. Birinci entegrasyon sonucunda özellikle gen ifadesi verilerinden elde edilen GBA'ların başlangıç aşamasına göre daha yüksek R^2 değerine sahip olduğu gözlemlenmiştir. ARNetMiT tabanlı GBA'ların R^2 değerlerinin düşük olması prostat kanserinde olduğu gibi göğüs kanserinde de başlangıç ve birinci entegrasyon aşamasının uygunluk değerlerini gösteren kutu grafiğindeki medyan değerinin düşmesine sebep olmuştur. İkinci entegrasyon aşamasında kesişim tabanlı herhangi bir GBA oluşturulamadığı için uygunluk değerleri elde edilmemiştir. İkinci entegrasyon aşamasında BÇO yaklaşımıyla entegrasyonu ile oluşturulan GBA'ların R^2 değerlerinin bire yaklaştığı gözlemlenmiştir.

İki katmanlı GBA entegrasyon mimarisi ile birinci ve ikinci entegrasyon aşamalarından elde edilen GBA'ların başlangıç aşamasına göre ölçeksiz ağ özelliklerini daha iyi yansıttığı çalışmamızda gösterilmiştir. Örtüşme analizinde grup içi entegrasyon aşamasının (birinci entegrasyon aşaması) performansa etkisi kısıtlı iken topolojik değerlendirmede grup içi entegrasyonun da uygunluk değerlerini arttırdığı gözlemlenmiştir. Gruplar arası entegrasyon değişik gen- gen ilişkilerinin entegrasyonu içermesine rağmen R^2 değerlerinin kısıtlı da olsa artmasını sağlamıştır. Özellikle ikinci entegrasyon aşamasındaki BÇO yaklaşımıyla üretilen GBA'ların ölçeksiz ağ tanımına daha uygun olduğu görülmüştür. Aynı aşamada birleşim tabanlı GBA'ların zayıf ilişki elemesi kriterlerinden dolayı R^2 değerleri BÇO ve kesişim tabanlı GBA'ların R^2 değerlerinden düşük çıkmıştır.

6.2.5 İki Katmanlı Entegrasyon Mimarisi ile Elde Edilen GBA'ların Biyolojik Olarak Değerlendirilmesi

Çalışmamızda GBA'ların son değerlendirme kriteri olarak GBA'larda yer alan modüllerin hastalıklarla ilişkili olduğu GO- terimleri incelenmiştir. Prostat ve göğüs kanseri verilerini kullanarak elde edilen GBA'lardaki modüllerin ilişkili olduğu GO- terimleri ile OMIM hastalık listesinden elde edilen prostat ve göğüs kanseriyle ilişkili genler kullanılarak belirlenen GO- terimleri arasında istatistiksel olarak anlamlı bir şekilde örtüşen GO- terimleri belirlenerek biyolojik performans değerlendirilmesi yapılmıştır. Bir GBA birden fazla modül içerebileceği için örtüşen GO- terim sayısını göstermek için kutu grafiği kullanılmıştır. Prostat kanseri verileri kullanılarak başlangıç, birinci ve ikinci entegrasyon aşamalarında elde edilen GBA'lara ait modüllerin hastalıkla ilişkilendirilmiş örtüşen GO- terim sayısı Şekil 6.17'de gösterilmiştir. Şekil 6.17'nin etiketlerinde diğer grafiklerde olmayan *AR*, *C3*, *WG*, *S* ve *SC* etiketleri gösterilmiştir. *AR* ARACNE tabanlı GBA'ları, *C3* C3NET tabanlı GBA'ları, *WG* WGCNA tabanlı GBA'ları, *S* destek tabanlı GBA'ları, *SC* ise destek-güven tabanlı GBA'ları göstermektedir.

GO Terimleri ile Değerlendirme



Şekil 6. 17 Prostat kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların hastalıkla örtüşen GO- terimleri kullanarak biyolojik olarak değerlendirilmesi

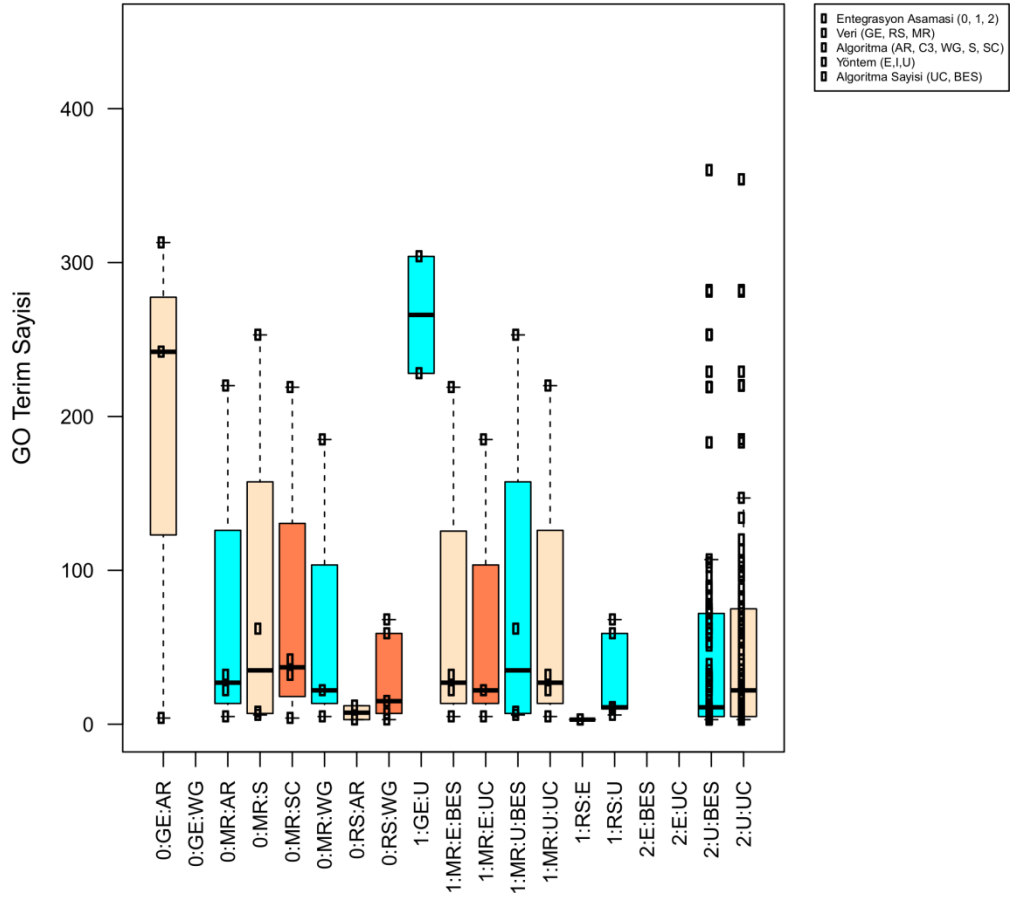
Başlangıç düzeyinde mikrodizin gen ifadesi verileri kullanılarak sadece ARACNE ve WGCNA algoritmaları tarafından belirlenen GBA'lara ait modüllerin hastalıkla ilişkili olduğu görülmektedir. Gen ifadesi verilerinde ARACNE tabanlı modül sayısı az olmasına rağmen bu modüllerin medyan değeri incelendiğinde 140'dan fazla prostat kanseriyle ilişkili GO- terimiyle ilişkilendirilmiştir. WGCNA tabanlı GBA'lardaki bazı modüllerde bu sayının 25'in altında çıktığı gözlemlenmiştir. miRNA- hedef gen verilerinde C3NET algoritması dışındaki tüm algoritmaların oluşturduğu GBA'larda yer alan modüller prostat kanseriyle ilişkili süreçlerle ilişkilendirilmiştir. Kutu grafiğindeki medyan değerleri göz önüne alındığında miRNA- hedef gen verileri üzerinde en yüksek GO- terimiyle ilişkili modüller ARNetMiT tabanlı dGBA tarafından başlangıç aşamasında elde edilmiştir. RNA- Seq verileri üzerinde gen ifadesi verilerinde olduğu gibi sadece ARACNE ve WGCNA tabanlı GBA'lardaki modüller prostat kanseriyle ilişkilendirilmiştir. RNA- Seq verileri üzerinde WGCNA tabanlı GBA'nın çok sayıda modülü hastalıkla ilişkilendirilmekle birlikte bazı modüllerin prostat kanseriyle ilişkili GO- terimlerin

medyan değeri 50'den fazla çıkmıştır. İlk entegrasyon aşamasında örtüşme analizi ve topolojik değerlendirme bölümlerindeki sonuçların aksine en başarılı sonuçlar birleştirme yöntemiyle elde edilen GBA'lar tarafından çıkarılmıştır. Birleştirme tabanlı GBA'lardan fazla sayıda prostat kanseriyle ilişkili GO- terimlerinin belirlenmesinin en temel nedeni kesişim ve BÇO tabanlı GBA'ların daha az sayıda gen ilişkisi içermesinden kaynaklanmaktadır. Birinci entegrasyon aşamasında miRNA- hedef gen verilerinde entegrasyonda kullanılan GBA sayısı BÇO tabanlı GBA'lara ait modüllerde hastalıkla örtüşen GO- terim değerlerini (medyan bazlı) değiştirmezken, GBA sayısının fazla olması nedeniyle birleşim tabanlı GBA'larda örtüşen GO- terim sayısı artmıştır. Birinci entegrasyon sonucunda oluşturulan GBA'ların biyolojik değerlendirmede başlangıç aşamasına göre performansı değiştirmedeği sonucuna varılmıştır.

İkinci entegrasyon aşamasında kesişim tabanlı GBA bulunmadığı için GBA'larla örtüşen GO- terim bulunmamaktadır. İkinci entegrasyonda da birleşim tabanlı GBA'ların daha fazla sayıda prostat kanseriyle ilişkili modül içerdiği gözlemlenmiştir. Bazı modüllerde bu sayı 250 GO- teriminin üzerindedir. İkinci entegrasyon aşamasında ARNetMiT tabanlı GBA'ların entegrasyon aşamasına eklenmesi performans sonuçlarını etkilememiştir. İkinci entegrasyon aşamasında BÇO tabanlı GBA'lardaki modüllerin hastalıkla ilgili örtüşen GO- terim sayısı, diğer aşamalarda elde edilen GBA'lardaki modüllere ait hastalıkla ilişkili örtüşen GO- terim sayısına yakın elde edilmiştir. Prostat kanseri ile ilişkili GBA'lardaki modüllerin hastalıkla ilişkisi incelendiğinde birinci ve ikinci entegrasyon sonucunda başlangıç aşamasına göre hastalıkla ilişkili GO- terimlerini içeren GBA modüllerinin sayısında belirgin bir artış veya azalış olmamıştır.

Göğüs kanseriyle ilgili başlangıç, birinci ve ikinci entegrasyon aşamalarında elde edilen GBA'ların modüllerinin hastalıkla ilişkili olarak belirlenen GO- terimleri sayısı Şekil 6.18'de gösterilmiştir.

GO Terimleri ile Degerlendirme



Şekil 6. 18 Göğüs kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların hastalıkla örtüşen GO- terimleri kullanarak biyolojik olarak değerlendirilmesi

Göğüs kanseri verilerinden elde edilen GBA'lardaki modüllerin göğüs kanseriyle ilişkisi GO- terim sayısı prostat kanseri verileri kullanılarak elde edilen prostat kanseriyle ilişkili GO- terim sayısından fazla olduğu Şekil 6.17 ve Şekil 6.18 karşılaştırıldığında gözlemlenmektedir. Başlangıç aşamasındaki GBA'lara ait modüllerin performansları değerlendirildiğinde en az göğüs kanseriyle ilişkili GO- terimleri RNA- Seq verileri kullanımında elde edilmiştir. Mikrodizin gen ifadesi verileri üzerinde ARACNE tabanlı GBA çok sayıda göğüs kanseriyle bağlantılı GO- terimiyle ilişkilendirilmiştir. ARNetMiT tabanlı algoritmalar tarafından oluşturulan GBA'lara ait modüller göğüs kanseriyle ilişkili GO- terimi belirlemede diğer algoritmalarından daha başarılı olduğu görülmüştür. miRNA- hedef gen verileri üzerindeki GBA sayısının fazla olması aynı veri kümesindeki birinci entegrasyon aşamalarında örtüşen GO- terimlerin medyan değerleri incelendiğinde az da olsa performans artışı sağladığı görülmüştür.

Birinci ve ikinci entegrasyon aşamalarında birleşim tabanlı GBA'ların modüllerinden çok sayıda göğüs kanseriyle ilişkili GO- terimleri elde edilmiştir. İkinci entegrasyon aşamasında BÇO tabanlı ve kesişim tabanlı GBA'lara ait modüllerin göğüs kanseriyle ilişkili GO- terim ilişkilendirilmediği Şekil 6.18'de gösterilmiştir. İkinci entegrasyon aşamasında ARNetMIT tabanlı GBA'ların entegrasyon aşamasında kullanılması biyolojik değerlendirmede belirgin bir performans artışı sağlamamıştır.

GBA'lar ve GBA'lara ait modüllerin biyolojik değerlendirmesinde entegrasyon işleminin biyolojik değerlendirme performansı net bir şekilde arttırmadığı sonucu elde edilmiştir. Bununla birlikte birinci ve ikinci entegrasyon aşamalarında birleşim yaklaşımıyla oluşturulan GBA'ların kesişim ve BÇO tabanlı GBA'lara göre biyolojik değerlendirmede daha başarılı olduğu gözlemlenmiştir.

SONUÇ VE ÖNERİLER

Kanser, kalp hastalıkları, şeker hastalığı gibi kompleks hastalıklar bir gene ait protein sentezine ait değişimlerin ötesinde çok sayıda genin birlikte yer aldığı biyolojik süreç ve olayların sonucunda meydana gelmektedir. İnsanlarda protein senteziyle ilişkilendirilmiş yaklaşık 20 bin kadar gen keşfedilmekle birlikte bu genlerin farklı fenotiplerde ilişkilerinin biyolojik deneylerle belirlenmesi zaman alıcı ve pahalı bir süreçtir. Çeşitli biyolojik veri kümeleri üzerinde gen ağı çıkarım algoritmalarının uygulanması sonucunda birlikte ifade edilen genler bulunmakta ve bu genlerle GBA'lar oluşturulmaktadır. Böylece hastalıklarla ilişkili biyolojik süreçlerde ortak olarak yer alma eğiliminde olan genler arasındaki ilişkiler hızlı ve az maliyetli olarak belirlenmektedir. GBA'larda ortak ifade edilen genlerin keşfi ile birlikte bu genlerin ilişkili olduğu prostat ve göğüs kanseriyle ilişkili biyolojik süreçlerin belirlenmesi hastalıkların altında yatan karmaşık mekanizmaların anlaşılması için önem taşımaktadır.

Literatürde farklı biyolojik veri kümeleri birlikte kullanılarak hastalıklarla ilişkili moleküler seviyede ilişkiler belirlenmeye çalışılmıştır. Bu çalışmalar temel olarak biyolojik veri kümelerindeki çeşitliliği kullanarak daha güvenilir biyolojik sonuçların belirlenmesini hedeflemektedir. Tez çalışmamız kapsamında mikrodizin gen ifadesi, RNA- Seq ve miRNA- hedef gen verileri üzerinde farklı gen ağı çıkarım algoritmaları kullanılarak elde edilen prostat ve göğüs kanseriyle ilişkili GBA'lar İki Katmanlı GBA Entegrasyon Mimarisi kullanılarak entegrasyon işlemi gerçekleştirilmiştir. Amacımız prostat ve göğüs kanseriyle ilgili daha kapsamlı ve doğruluklu GBA'ları belirlemektir.

Literatürde farklı gen ağı çıkarım algoritmaları ve biyolojik veri kümelerini ortak bir şekilde entegre eden çalışma sayısı çok kısıtlıdır. Oluşturduğumuz iki katmanlı entegrasyon mimarisiyle farklı gen ağı çıkarım algoritmalarına ait gen- gen ilişki tahmini çeşitliliği ve biyolojik veri kümelerinden elde edilen GBA oluşumundaki çeşitlilik bir arada kullanılmıştır. Tezimizin diğer çalışmalara göre en büyük farkı tek bir biyolojik veri kümesi veya tek bir gen ağı çıkarım algoritmasına dayalı sonuçlar elde etmek yerine çeşitli gen ağı çıkarım algoritmaları ve biyolojik veri kümeleri kullanarak hastalıklarla ilişkili, kapsamlı ve doğruluklu GBA çıkarımını gerçekleştirmektir.

İki Katmanlı GBA Entegrasyon Mimarisi başlangıç, birinci ve ikinci entegrasyon aşamalarını içeren üç temel düzeyden oluşmaktadır. Başlangıç düzeyinde farklı gen ağı çıkarım algoritmaları üç farklı biyolojik veri kümesine ayrı ayrı uygulanmış ve GBA'lar elde edilmiştir. Birinci entegrasyon düzeyinde farklı gen ağı çıkarım algoritmalarının aynı veri kümeleri üzerinde elde ettikleri GBA'lar kesişim, basit çoğunluk oyu ve birleşim metotlarıyla birbirlerine entegre edilmiştir. Birinci aşamadaki bu entegrasyona grup içi entegrasyon denilmiştir. İkinci entegrasyon aşamasında ise farklı biyolojik veri kümelerinde birinci entegrasyon sonucunda oluşturulmuş GBA'ların yine kesişim, basit çoğunluk oyu ve birleşim metotlarıyla birleştirilmesine dayanmaktadır. Bu aşamadaki entegrasyona ise gruplar arası entegrasyon denilmiştir. İki katmanlı GBA entegrasyondaki başlangıç aşamasında gen ağı çıkarım algoritmalarının performansları karşılaştırılmıştır. Birinci entegrasyon aşamasında elde edilen sonuçlar başlangıç aşamasındaki sonuçlarla, ikinci entegrasyon sonucunda elde edilen sonuçlar ise başlangıç ve birinci entegrasyon aşamasında elde edilen sonuçlarla karşılaştırılmıştır. Çalışmamızda literatür örtüşme analizi, topolojik ve biyolojik değerlendirme kriterleriyle GBA'ların performansları karşılaştırılmıştır.

Sonuçlara göre örtüşme analizindeki en düşük kesinlik değerleri miRNA- hedef gen verileri kullanılarak elde edilmiştir. Gen ağı çıkarım algoritmalarının arasında tüm veri kümelerinde en iyi performans gösteren bir algoritma olmamasına rağmen en az gen ilişki tahmini C3NET algoritması ile elde edilmiştir. Birinci entegrasyon sonucunda kesişim tabanlı GBA'ların performanslarının başlangıç aşamasındaki GBA'lardan daha başarılı olduğu sonucuna varılmıştır. Birinci entegrasyon aşamasında basit çoğunluk oyu ve birleşim tabanlı GBA'ların örtüşme analizi performansı başlangıç aşamasındaki

GBA'ların performansına yakın elde edilmiş ve herhangi bir performans artışı gözlemlenmemiştir. Örtüşme analizindeki en başarılı GBA sonuçları ikinci entegrasyon sonucunda elde edilmiştir. Özellikle basit çoğunluk oyu tabanlı ve kesişim tabanlı GBA'ların kesinlik değerleri diğer aşamalarda elde edilen GBA'ların kesinlik değerlerinden yüksek çıkmıştır. Bu sonuçlara göre grup içi entegrasyonun örtüşme analizinde performansı çok etkilemediği, gruplar arası entegrasyonun ise örtüşme analizinde performans artışına neden olduğu gözlemlenmiştir.

Çalışmamızdaki bir diğer GBA performans değerlendirme kriteri ise GBA'ların ölçeksiz ağ özelliklerine uygunluğunun ölçülmesidir. WGCNA algoritması GBA çıkarımında GBA'ların ölçeksiz ağ olmasına göre eşik değeri belirlediği için başlangıç aşamasında en başarılı GBA'lar WGCNA tarafından elde edilmiştir. GBA'lara ait en düşük ölçeksiz ağ uygunluk değerleri yine miRNA- hedef gen verileri kullanılarak elde edilmiştir. Birinci ve ikinci entegrasyon aşamalarında özellikle basit çoğunluk oyu yaklaşımıyla elde edilen GBA'ların uygunluk değerleri başlangıç aşamasındaki GBA'lara oranla daha yüksek çıkmıştır. Özellikle ikinci entegrasyon aşamasında birleşim tabanlı GBA'ların ölçeksiz ağ uygunluk değerleri diğer yaklaşımlara oranla düşük çıkmıştır. Bunun en büyük nedeni birleşim tabanlı GBA oluşumu sırasındaki zayıf gen- gen ilişki budama aşamasından kaynaklanmaktadır.

GBA'lar son olarak içerdikleri modüllerdeki genlerin prostat ve göğüs kanseriyle ilişkili GO- terim sayıları kullanılarak biyolojik olarak değerlendirilmiştir. Başlangıç aşamasında WGCNA ve ARACNE algoritmaları tabanlı GBA'lara ait modüller hastalıkla ilişkili GO- terimleriyle ilişkilendirilmiştir. C3NET ise bu aşamada performansı en düşük algoritma olarak gözlemlenmiştir. Birinci ve ikinci entegrasyon aşamalarının biyolojik olarak değerlendirmede performans artışına etkisi kısıtlı olmuştur. Birinci ve ikinci entegrasyon aşamalarında birleştirme tabanlı GBA'ların hastalıklarla ilişkili GO- terim sayısı çıkarımında daha iyi olduğu gözlemlenmiştir.

GBA'ların entegrasyonu ile birlikte daha önce miRNA- hedef gen verileri üzerinde daha önce hiç uygulanmamış birliktelik kuralı tabanlı algoritma olan ARNetMiT yaklaşımı tez kapsamında geliştirilmiş ve olası 152 hastalıkla ilişkili GBA'lar oluşturulabilir hale gelmiştir. miRNA- hedef gen verileri üzerinde yapılan değerlendirmede ARNetMiT

tabanlı GBA'ların örtüşme analizi ve biyolojik değerlendirmede WGCNA, C3NET ve ARACNE tabanlı GBA'lardan daha başarılı olduğu gözlemlenmiştir. Topolojik değerlendirmede ise WGCNA ve C3NET algoritması tarafından oluşturulan GBA'ların ölçeksiz ağ yapısına daha uygun olduğu anlaşılmıştır. Birinci ve ikinci entegrasyon aşamalarında miRNA- hedef gen verileri üzerinde ARNetMiT algoritması kullanılarak GBA sayısı arttırıldığında örtüşme analizi ve topolojik performans değerlendirmesinde artış olduğu gözlemlenmiştir.

Tez kapsamında değişkenliğin arttırımı için entegrasyon aşamasında farklı biyolojik veri kümeleri ve gen ağı çıkarım algoritmalarına yer verilmiştir. Kullandığımız gen ağı çıkarım yöntemlerinin hepsi bilgi teorisi tabanlı algoritmalarından oluşmaktadır. Sadece bilgi teorisi tabanlı gen ağı çıkarım algoritmaları değil, Bayes Ağları, Regresyon yaklaşımıyla gen ağı çıkarımı gibi farklı yaklaşımlara ait algoritmalara yer verilmesi birinci entegrasyon aşamasında performansı arttırabilecek olan yaklaşımlardandır. Bununla birlikte entegrasyon aşamasında çekirdek fonksiyon tabanlı, olasılık tabanlı entegrasyon yöntemlerinin kullanılması da performansı olumlu olarak etkileyeceği düşünülmektedir. Biyolojik veri kümelerinin ve gen ağı çıkarım algoritmaların arttırılması da daha kapsamlı ve doğruluklu GBA'ların elde edilmesine olanak sağlayabilmektedir.

Çalışmamız sonunda farklı biyolojik veri kümelerinden ve algoritmalarından elde edilen GBA'lar entegre edilmiş ve farklı değerlendirme kriterlerine göre daha kapsamlı ve doğruluklu GBA elde edimini sağlanmıştır. Bundan sonra yukarıda bahsettiğimiz algoritma sayısının ve biyolojik veri sayısının artması, farklı tip gen ağı çıkarım algoritmasının uygulanması, entegrasyon aşamasında olasılıksal ve makine öğrenmesine dayalı farklı yaklaşımlarla GBA'ların entegre edilmesi planlanmaktadır.

KAYNAKLAR

- [1] Gligorijević, V. ve Pržulj, N. (2015). "Methods for biological data integration: perspectives and challenges.", *Journal of the Royal Society Interface*, 12(112), 20150571.
- [2] Agro, A. F., Rotilio, G., Strom, R., Moricca, G. ve Fanelli, A. R. (1969). "The biochemical mechanism of selective heat sensitivity of cancer cells: II. Studies on nucleic acids and protein synthesis", *European Journal of Cancer* (1965), 5(2), 137-146
- [3] Ruggero, D. (2013). "Translational control in cancer etiology", *Cold Spring Harbor perspectives in biology*, 5(2), a012336
- [4] Heather, J. M. ve Chain, B. (2016). "The sequence of sequencers: the history of sequencing DNA", *Genomics*, 107(1), 1-8.
- [5] Mardis, E. R. (2011). "A decade/s perspective on DNA sequencing technology", *Nature*, 470(7333), 198-203.
- [6] Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics", *Trends in genetics*, 24(3), 133-141.
- [7] Ballouz, S., Verleyen, W. ve Gillis, J. (2015). "Guidance for RNA-seq co-expression network construction and analysis: safety in numbers", *Bioinformatics*, 31(13), 2123-2130.
- [8] de Matos Simoes, R., Dalleau, S., Williamson, K. E. ve Emmert-Streib, F. (2015). "Urothelial cancer gene regulatory networks inferred from large-scale RNAseq, Bead and Oligo gene expression data", *BMC systems biology*, 9(1), 21.
- [9] Iancu, O. D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R. ve McWeeney, S. (2012). "Utilizing RNA-Seq data for de novo coexpression network inference", *Bioinformatics*, 28(12), 1592-1597.
- [10] Mamdani, M., Williamson, V., McMichael, G. O. ve Bacanu, S. A. (2015). "Integrating mRNA and miRNA weighted gene co-expression networks with eQTLs in the nucleus accumbens of subjects with alcohol dependence", *PloS one*, 10(9), e0137671.

- [11] Lai, X., Bhattacharya, A., Schmitz, U., Kunz, M., Vera, J. ve Wolkenhauer, O. (2013). "A systems' biology approach to study microRNA-mediated gene regulatory networks", *BioMed research international*, 2013.
- [12] Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B. ve Marchal, K. (2006). "Inferring transcriptional modules from ChIP-chip, motif and microarray data", *Genome biology*, 7(5), R37.
- [13] Vu, T. T. ve Vohradsky, J. (2009). "Inference of active transcriptional networks by integration of gene expression kinetics modeling and multisource data", *Genomics*, 93(5), 426-433.
- [14] Guan, D., Shao, J., Deng, Y., Wang, P., Zhao, Z., Liang, Y. ve Yan, B. (2014). "CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data", *Bioinformatics*, 30(8), 1190-1192.
- [15] Li, H., Wang, N., Gong, P., Perkins, E. J. ve Zhang, C. (2011). "Learning the structure of gene regulatory networks from time series gene expression data", *BMC genomics*, 12(5), S13.
- [16] Liu, F., Zhang, S. W., Guo, W. F., Wei, Z. G. ve Chen, L. (2016). "Inference of gene regulatory network based on local bayesian networks", *PLoS computational biology*, 12(8), e1005024.
- [17] Singh, N. ve Vidyasagar, M. (2016). "blars: An algorithm to infer gene regulatory networks", *IEEE/ACM transactions on computational biology and bioinformatics*, 13(2), 301-314.
- [18] van Dam, S., Vösa, U., van der Graaf, A., Franke, L. ve de Magalhães, J. P. (2017). "Gene co-expression analysis for functional classification and gene-disease predictions", *Briefings in bioinformatics*, bbw139.
- [19] Ha, M. J., Baladandayuthapani, V. ve Do, K. A. (2015). "DINGO: differential network analysis in genomics", *Bioinformatics*, 31(21), 3413-3420.
- [20] Gov, E. ve Arga, K. Y. (2016). "Interactive cooperation and hierarchical operation of microRNA and transcription factor crosstalk in human transcriptional regulatory network", *IET systems biology*, 10(6), 219-228
- [21] Bussey, K. J., Chin, K., Lababidi, S., Reimers, Kuo, W. L. ve Collins, C. (2006). "Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel", *Molecular cancer therapeutics*, 5(4), 853-867.
- [22] Sun, Z., Asmann, Y. W., Kalari, K. R., Bot, B., Eckel-Passow, J. E., Baker, T. R. ve Schroth, G. P. (2011). "Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing", *PloS one*, 6(2), e17490.
- [23] Olsen, C., Fleming, K., Prendergast, N., Rubio, R., Emmert-Streib, F., Bontempi, G. ve Quackenbush, J. (2014). "Inference and validation of predictive gene networks from biomedical literature and gene expression data", *Genomics*, 103(5), 329-336. ISO 690

- [24] Hériché, J. K., Lees, J. G., Morilla, I., Walter, T., Petrova, B., Roberti, M. J. ve Haering, C. H. (2014). "Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation", *Molecular biology of the cell*, 25(16), 2522-2536.
- [25] Gönen, M. ve Alpaydın, E. (2011). "Multiple kernel learning algorithms", *Journal of machine learning research*, 12(Jul), 2211-2268.
- [26] Semenza, G. L. (2003). "Targeting HIF-1 for cancer therapy", *Nature reviews cancer*, 3(10), 721-732.
- [27] Levine, B. (2007). "Cell biology: autophagy and cancer", *Nature*, 446(7137), 745-747.
- [28] Chu, M., Zhang, R., Zhao, Y., Wu, C., Guo, H., Zhou, B. ve Ma, H. (2013). "A genome-wide gene–gene interaction analysis identifies an epistatic gene pair for lung cancer susceptibility in Han Chinese", *Carcinogenesis*, 35(3), 572-577.
- [29] Lo, S. H., Chernoff, H., Cong, L., Ding, Y. ve Zheng, T. (2008). "Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer", *Proceedings of the National Academy of Sciences*, 105(34), 12387-12392.
- [30] Liao, S., Davoli, T., Leng, Y., Li, M. Z., Xu, Q. ve Elledge, S. J. (2017). "A genetic interaction analysis identifies cancer drivers that modify EGFR dependency", *Genes & Development*, 31(2), 184-196.
- [31] Price, R. A., Ness, R. O. B. E. R. T. A. ve Laskarzewski, P. (1990). "Common major gene inheritance of extreme overweight", *Human biology*, 747-765.
- [32] International Human Genome Sequencing Consortium. (2004). "Finishing the euchromatic sequence of the human genome", *Nature*, 431(7011), 931-945.
- [33] Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N. ve Klitgord, N. (2005). "Towards a proteome-scale map of the human protein–protein interaction network", *Nature*, 437(7062), 1173-1178.
- [34] Emmert-Streib, F. ve Dehmer, M. (Eds.). (2010). "Medical biostatistics for complex diseases", John Wiley & Sons.
- [35] Schadt, E. E. (2009). "Molecular networks as sensors and drivers of common human diseases", *Nature*, 461(7261), 218-223.
- [36] Obayashi, T., Hayashi, S., Shibaoka, M., Saeki, M., Ohta, H. ve Kinoshita, K. (2007). "COXPRESdb: a database of coexpressed gene networks in mammals", *Nucleic acids research*, 36(suppl_1), D77-D82.
- [37] Kim, H., Lee, J. K. ve Park, T. (2007). "Boolean networks using the chi-square test for inferring large-scale gene regulatory networks", *BMC bioinformatics*, 8(1), 37.
- [38] Berger, S. I. ve Iyengar, R. (2009). "Network analyses in systems pharmacology", *Bioinformatics*, 25(19), 2466-2472.

- [39] Albert, R. (2005). "Scale-free networks in cell biology", *Journal of cell science*, 118(21), 4947-4957.
- [40] 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(5586), 1551-1555.
- [41] Jordan, I. K., Mariño-Ramírez, L., Wolf, Y. I. ve Koonin, E. V. (2004). "Conservation and coevolution in the scale-free human gene coexpression network", *Molecular biology and evolution*, 21(11), 2058-2070.
- [42] Thompson, G. W. ve Kello, C. T. (2014). "Walking across Wikipedia: a scale-free network model of semantic memory retrieval", *Frontiers in psychology*, 5.
- [43] Chung, F. R. ve Lu, L. (2006). "Complex graphs and networks", American Mathematical Society,. ISO 690
- [44] Giorgi, F. M., Del Fabbro, C. ve Licausi, F. (2013). "Comparative study of RNA-seq-and microarray-derived coexpression networks in *Arabidopsis thaliana*", *Bioinformatics*, 29(6), 717-724.
- [45] Shinozaki, K., Yamaguchi-Shinozaki, K. ve Seki, M. (2003). "Regulatory network of gene expression in the drought and cold stress responses", *Current opinion in plant biology*, 6(5), 410-417.
- [46] Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. ve Friedman, N. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data", *Nature genetics*, 34(2), 166-176.
- [47] Elowitz, M. B., Levine, A. J., Siggia, E. D. ve Swain, P. S. (2002). "Stochastic gene expression in a single cell", *Science*, 297(5584), 1183-1186.
- [48] Dudoit, S., Yang, Y. H., Callow, M. J. ve Speed, T. P. (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", *Statistica sinica*, 111-139.
- [49] Reiner, A., Yekutieli, D. ve Benjamini, Y. (2003). "Identifying differentially expressed genes using false discovery rate controlling procedures", *Bioinformatics*, 19(3), 368-375.
- [50] Pan, W. (2002). "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray", *Bioinformatics*, 18(4), 546-554
- [51] Jenson, S. D., Robetorye, R. S., Bohling, S. D., Schumacher, J. A., Morgan, J. W., Lim, M. S. (2003). "Validation of cDNA microarray gene expression data obtained from linearly amplified RNA", *Molecular Pathology*, 56(6), 307.
- [52] Buhler, J. (2002). "Anatomy of a comparative gene expression study", Washington university in St. Louis.
- [53] Ozsolak, F. ve Milos, P. M. (2011). "RNA sequencing: advances, challenges and opportunities", *Nature reviews genetics*, 12(2), 87-98.

- [54] Garber, M., Grabherr, M. G., Guttman, M. ve Trapnell, C. (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq", *Nature methods*, 8(6), 469-477.
- [55] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R. ve Pachter, L. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", *Nature protocols*, 7(3), 562-578.
- [56] Xu, C. F., Yu, C. H. ve Li, Y. M. (2009). "Regulation of hepatic microRNA expression in response to ischemic preconditioning following ischemia/reperfusion injury in mice", *OMICS A Journal of Integrative Biology*, 13(6), 513-520.
- [57] Cingiz, M. Ö., & Diri, B. (2017). "The inference of gene co-expression networks of breast and colon cancer using miRNA-target gene interactions data", In *Signal Processing and Communications Applications Conference (SIU)*, 2017 25th (pp. 1-4). IEEE.
- [58] Bueno, M. J., de Castro, I. P. ve Malumbres, M. (2008). "Control of cell proliferation pathways by microRNAs", *Cell cycle*, 7(20), 3143-3148.
- [59] Tomé, M., López-Romero, P., Albo, C., Sepúlveda, J. C., Fernández-Gutiérrez, B., Dopazo, A. ve González, M. A. (2011). "miR-335 orchestrates cell proliferation, migration and differentiation in human mesenchymal stem cells", *Cell Death & Differentiation*, 18(6), 985-995.
- [60] Altay, G., & Emmert-Streib, F. (2010). "Inferring the conservative causal core of gene regulatory networks", *BMC Systems Biology*, 4(1), 132.
- [61] 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*, 7(1), S7
- [62] Roy, S., Lane, T. ve Werner-Washburne, M. (2009, June). "Learning structurally consistent undirected probabilistic graphical models", In *Proceedings of the 26th annual international conference on machine learning* (pp. 905-912). ACM.
- [63] Meyer, P. E., Lafitte, F. ve Bontempi, G. (2008). "minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information", *BMC bioinformatics*, 9(1), 461.
- [64] Langfelder, P. ve Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1), 559.
- [65] Cingiz, M. Ö., Biricik, G. ve Diri, B. (2017). "ARNetMiT R Package: association rules based gene co-expression networks of miRNA targets", *Cellular and molecular biology (Noisy-le-Grand, France)*, 63(3), 18.
- [66] Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X. ve Liu, Y. (2008). "miR2Disease: a manually curated database for microRNA deregulation in human disease", *Nucleic acids research*, 37(suppl_1), D98-D104.

- [67] Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S. D., Lin, Y. L., Lee, W. H ve Tsai, T. R. (2015). "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database", *Nucleic acids research*, 44(D1), D239-D247.
- [68] Hsu, S. D., Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H. ve Jian, T. Y. (2014). "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions", *Nucleic acids research*, 42(D1), D78-D85.
- [69] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S. ve Coello, C. A. C. (2014). "Survey of multiobjective evolutionary algorithms for data mining", Part II. *IEEE Transactions on Evolutionary Computation*, 18(1), 20-35.
- [70] Agrawal, R. ve Ramakrishnan, S. (1994). "Fast algorithms for mining association rules", *Proc VLDB'94*, 487-99.
- [71] Zaki, M. J. (2000). "Scalable algorithms for association mining", *Transactions on Knowledge and Data Engineering*, 12(3): 372-90.
- [72] Brown, C. (2015). hash: Full feature implementation of hash/associated arrays/dictionaries, <https://cran.rstudio.com/web/packages/hash/index.html>
- [73] Altay, G., Altay, N. ve Neal, D. (2013). "Global assessment of network inference algorithms based on available literature of gene/protein interactions", *Turkish Journal of Biology*, 37(5), 547-555.
- [74] Castro, M. A., Wang, X., Fletcher, M. N., Meyer, K. B. ve Markowetz, F. (2012). "RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations", *Genome biology*, 13(4), R29.
- [75] Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y. ve Michoel, T. (2009). "Module networks revisited: computational assessment and prioritization of model predictions", *Bioinformatics*, 25(4), 490-496.
- [76] Li, B., Zhang, Y., Yu, Y., Wang, P., Wang, Y., Wang, Z. ve Wang, Y. (2015). "Quantitative assessment of gene expression network module-validation methods", *Scientific reports*, 5.
- [77] Bader, G. D. ve Hogue, C. W. (2003). "An automated method for finding molecular complexes in large protein interaction networks", *BMC bioinformatics*, 4(1), 2.
- [78] Wu, X. Y. ve Xia, X. Y. (2015). Pronet: biological network construction, visualization and analyses.
- [79] Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H. ve Vilo, J. (2016). "g: Profiler—a web server for functional interpretation of gene lists (2016 update)", *Nucleic acids research*, 44(W1), W83-W89.
- [80] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. ve McKusick, V. A. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of

human genes and genetic disorders", *Nucleic acids research*, 33(suppl_1), D514-D517.

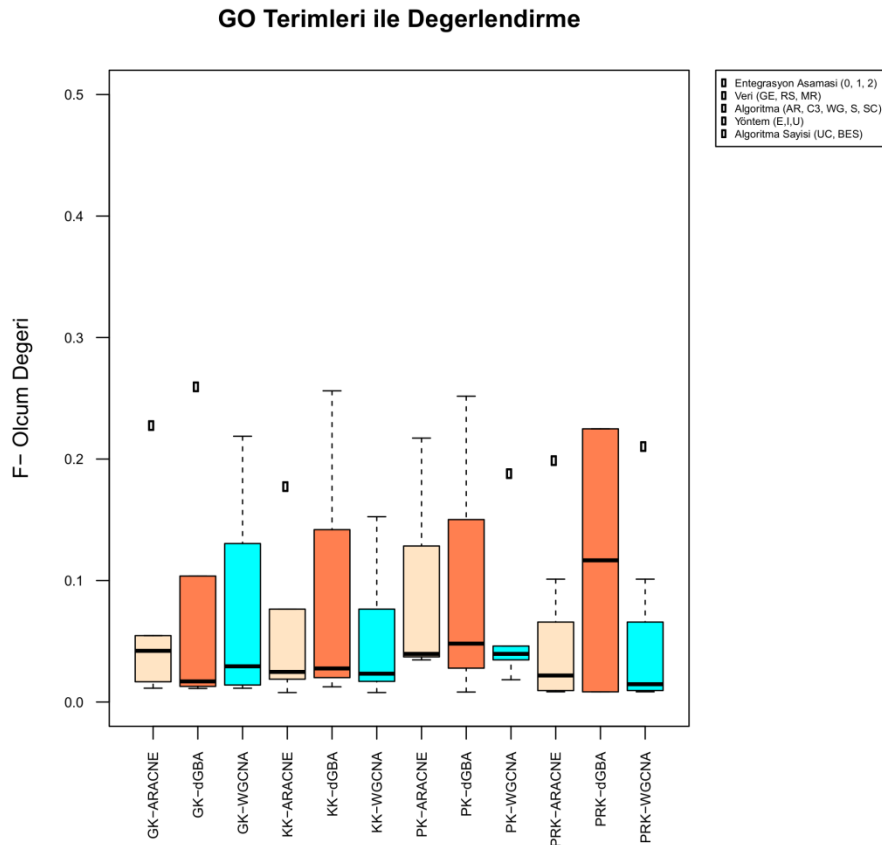
- [81] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K ve Cancer Genome Atlas Research Network. (2013). "The cancer genome atlas pan-cancer analysis project", *Nature genetics*, 45(10), 1113-1120.
- [82] Risso, D., Schwartz, K., Sherlock, G. ve Dudoit, S. (2011). "GC-content normalization for RNA-Seq data", *BMC bioinformatics*, 12(1), 480.
- [83] Robinson, M. D. ve Oshlack, A. (2010). "A scaling normalization method for differential expression analysis of RNA-seq data", *Genome biology*, 11(3), R25.
- [84] Maza, E., Frasse, P., Senin, P., Bouzayen, M. ve Zouine, M. (2013). "Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes", *Communicative & integrative biology*, 6(6), e25849.
- [85] Gonzalez, I. (2014). "Statistical analysis of RNA- Seq data", Plateforme Bioinformatique- INRA Toulouse, Plateforme Biostatistique- IMT Universit'e Toulouse III
- [86] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", *Bioinformatics*, 26(1), 139-140.

GEN BİRLİKTELİK AĞLARININ F-ÖLÇÜM DEĞERİ İLE BİYOLOJİK DEĞERLENDİRMESİNİN GERÇEKLENMESİ

Hastalıklara ilişkili gen birliktelik ağlarına ait modüllerdeki genler GO Zenginleştirme Analizi'nde kullanılmasıyla elde edilen GO- terimleri ve OMIM hastalık- gen veri kümesindeki hastalıkla ilgili genlerin GO Zenginleştirme Analizi'nde kullanılmasıyla elde edilen GO- terimleri örtüşme analizinde DP değerleri kullanılarak biyolojik değerlendirme çalışmamızda gerçekleştirilmiştir. DP değerleri tek başlarına kullanıldığında GBA'lardaki modüllere ait hastalıklarla ilişkili istatistiksel olarak anlamlı GO- terim sayılarını vermekle birlikte her bir modülle ilişkili toplam GO- terim sayısı ve OMIM hastalık veri kümesi kullanılarak hastalıklarla ilişkilendirilmiş toplam GO- terim sayısı bilgilerini vermekte eksik kalmaktadır. Bu bölümde gen birliktelik ağlarının biyolojik olarak değerlendirmesinde F-ölçüm değeri (F_1 -ölçüm değeri veya F-ölçeği) kullanarak daha kapsamlı analiz yapılması hedeflenmiştir. OMIM veri kümesi kullanılarak hastalıkla ilişkilendirilmiş GO- terimleri test veri kümesini, GBA'lara ait modüllerle ilişkilendirilmiş GO- terimleri ise eğitim verisi olarak değerlendirilmiştir. Bir modüle ait GO- terim sayısı hastalıkla ilişkili GO- terimleri listesinde de yer alıyorsa DP olarak, yer almıyorsa FP olarak örtüşme analizinde değerlendirilmiştir. Hastalıkla ilişkili olmasına rağmen gen birliktelik ağlarına ait modüllerle ilişkilendirilmemiş GO- terimleri ise YN olarak belirlenerek örtüşme analizinde F-ölçüm değeri hesaplanmıştır.

A-1 ARNetMiT Tabanlı Gen Birliktelik Ağlarının F-Ölçüm Değeri İle Değerlendirilmesi

Prostat, göğüs, pankreas ve kolon kanseriyle ilişkili miRNA- hedef gen verileri kullanarak dört kanser türüne ait gen birliktelik ağları ARNetMiT R paketi kullanılarak çalışmamızda elde edilmiştir. Dört kanserle ilişkili OMIM hastalık- gen veri kümesi yer alan hastalıklarla ilişkili genler GO Zenginleştirme Analizi'nde kullanıldığında göğüs kanseriyle ilişkili 920 GO- terimi, göğüs kanseriyle ilişkili 1032 GO- terimi, kolon kanseriyle ilişkili 1253 GO- terimi ve pankreas kanseriyle ilişkili 941 GO- terimi elde edilmiştir. Dört kanser türüne ait veriler kullanılarak ARACNE, WGCNA, C3NET ve dGBA tabanlı gen birliktelik ağlarındaki modüllerin hastalıklarla ilişkili örtüşen GO- terim sayısı Şekil 6.6'da gösterilmiştir. GAÇ algoritmaları ve ARNetMiT tabanlı dGBA'ların F-ölçüm değeriyle biyolojik olarak değerlendirilmesine ait sonuçlar Şekil Ek-A.1'de gösterilmiştir.

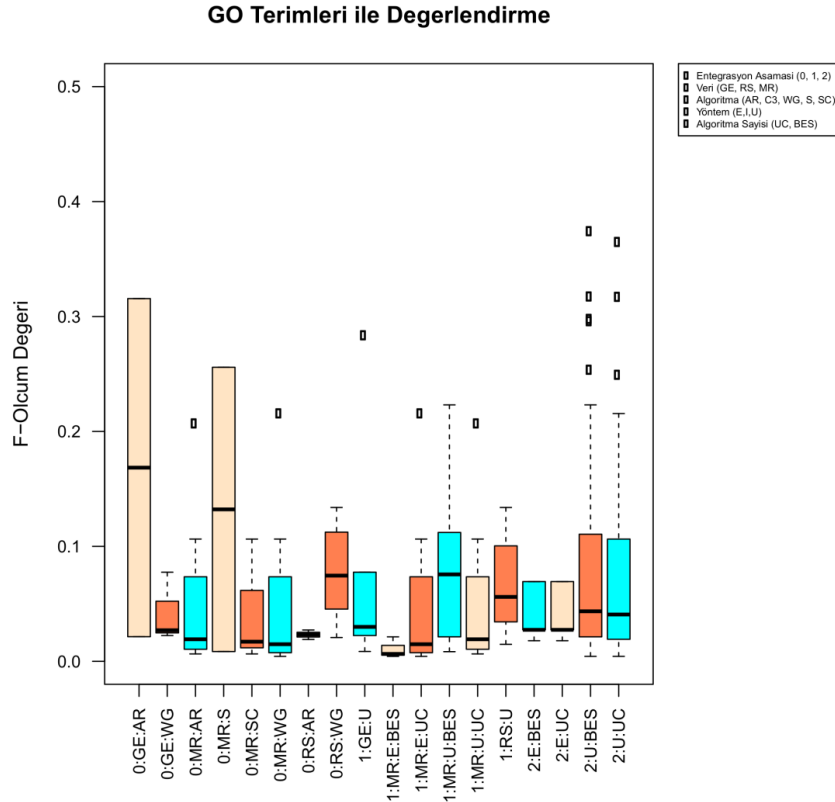


Şekil Ek-A. 1 GAÇ algoritmaları ve ARNetMiT tabanlı GBA'ların F-ölçüm değeri ile biyolojik olarak değerlendirilmesi

Şekil Ek-A.1'de yer alan sonuçlar Şekil 6.6'da elde edilen hastalıkla ilişkili örtüşen GO-terim sayılarından elde edilen sonuçlarla paralellik göstermektedir. Kutu grafiğindeki medyan değerleri incelendiğinde göğüs kanseri dışındaki diğer üç kanser türünde dGBA'ların en yüksek F-ölçüm değerlerine sahip olduğu gözlemlenmiştir.

A-2 İki Katmanlı Entegrasyon Mimarisiyle Elde Edilen Gen Birliktelik Ağlarının F-Ölçüm Değeri İle Değerlendirilmesi

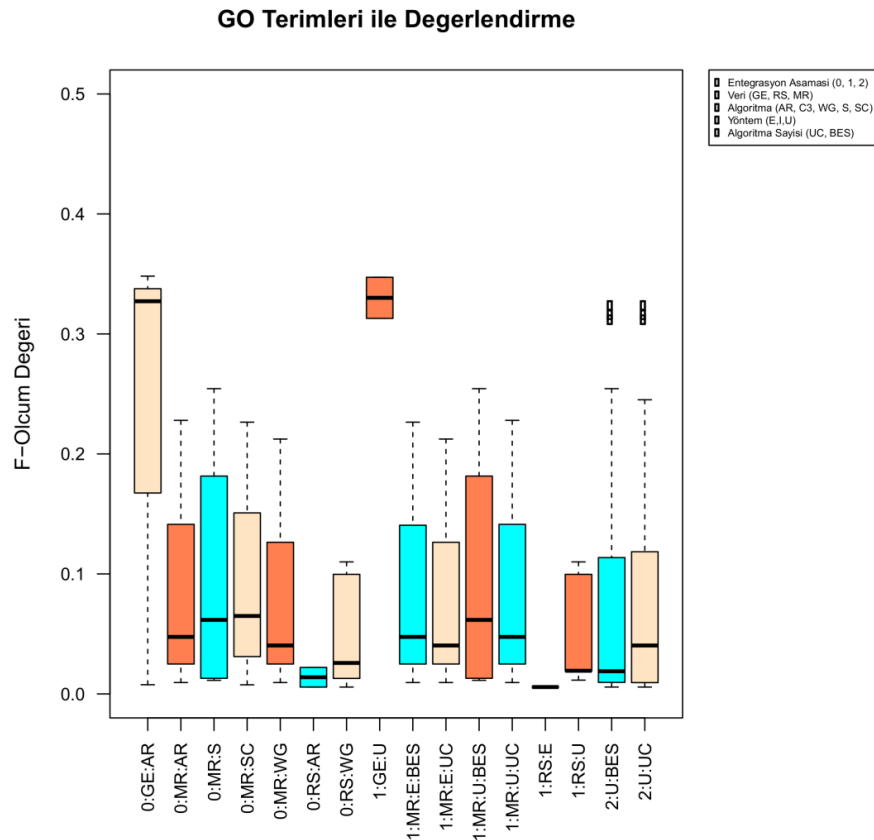
Prostat ve göğüs kanseriyle ilişkili iki katmanlı entegrasyon mimarisiyle başlağıç, birinci ve ikinci entegrasyon düzeyinde elde edilen gen birliktelik ağlarının F-ölçüm değerleri kullanılarak biyolojik olarak değerlendirilmiştir. OMIM hastalık- gen veri kümesindeki genler kullanılarak prostat kanseriyle ilişkili 920 GO-terimi, göğüs kanseriyle ilişkili 1032 GO-terimi elde edilmiştir. Prostat kanseriyle ilişkili GBA'larda yer alan modüllerinin F-ölçüm değeriyle biyolojik olarak değerlendirilmesine ait sonuçlar Şekil Ek-A.2'de gösterilmiştir.



Şekil Ek-A. 2 Prostat kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların F-ölçüm değeriyle biyolojik olarak değerlendirilmesi

Sonuçlara göre birinci ve ikinci entegrasyon düzeyinde elde edilen GBA'lara ait modüllerin F-ölçüm değerleri ile başlangıç aşamasında elde edilen GBA'lardaki modüllerin F-ölçüm değerleriyle karşılaştırıldığında performans artışı gerçekleşmediği gözlemlenmiştir. F-ölçüm değerleri kutu grafiğinde medyan değerlerine göre karşılaştırıldığında en yüksek değerlerin başlangıç aşamasında gen ifadesi ve miRNA-hedef gen verileri üzerinde elde edildiği sonucuna varılmıştır. Şekil EK- A.2'de elde edilen sonuçlar sadece DP değerleri kullanılarak Şekil 6.17'de gösterilen biyolojik değerlendirme sonuçlarıyla örtüşmektedir.

Benzer şekilde göğüs kanseriyle ilişkili GBA'larda yer alan modüllerinin örtüşme analizi sonucunda değerlendirilmesiyle elde edilen F-ölçüm değerleri Şekil Ek-A.3'te gösterilmiştir.



Şekil Ek-A. 3 Göğüs kanserine ait tüm entegrasyon aşamalarında elde edilen GBA'ların F-ölçüm değeriyle biyolojik olarak değerlendirilmesi

Göğüs kanserine ait GBA'ların örtüşme sonuçları F-ölçüm değerleriyle analiz edildiğinde prostat kanseri verilerinde elde edilen sonuçlara benzer sonuçlar çıktığı

gözlemlenmiştir. Göğüs kanseri verilerinde de grup içi ve gruplar arası entegrasyon biyolojik değerlendirmede performans artışı kısıtlı kalmıştır. Şekil EK- A.3'te elde edilen sonuçlar sadece DP değerleri kullanılarak Şekil 6.18'de gösterilen biyolojik değerlendirme sonuçlarıyla örtüşmektedir.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Mustafa Özgür CİNGİZ
Doğum Tarihi ve Yeri : 16 Şubat 1984, Kütahya- Simav
Yabancı Dili : İngilizce
E-posta : ozgurcingiz@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	Bilgisayar Mühendisliği	Yıldız Teknik Üniversitesi	2012
Lisans	Endüstri Mühendisliği	İstanbul Kültür Üniversitesi	2009
Lisans	Bilgisayar Mühendisliği	İstanbul Kültür Üniversitesi	2008
Lise	Sayısal	Kadıköy Anadolu Lisesi	2003

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2010- 2018	Yıldız Teknik Üniversitesi	Araştırma Görevlisi
2009-2010	İstanbul Kültür Üniversitesi	Araştırma Görevlisi

YAYINLARI

Makale

1. Cingiz, M. Özgür., Diri, B., Two-tier combinatorial structure to integrate various gene co-expression networks of prostate cancer, Genomics (In review) (Doktora Tezinden)
2. Cingiz, M. Özgür, Biricik, G., Diri, B. ARNetMiT R Package: association rules based gene co-expression networks of miRNA targets. Cellular and Molecular Biology, 2017, 63.3: 18-25. (Doktora Tezinden)
3. Cingiz, M. Özgür., Banu Diri., and Göksel Biricik. "Am I typing fresh tweets: Detecting up-to-dateness and worth of categorical information in microblogs." Expert Systems with Applications 42.12 (2015): 5256-5263, DOI: 10.1016/j.eswa.2015.02.025

Bildiri

1. Cingiz, M. Özgür, Biricik, G. and Diri, B. " A combinatorial approach to construct core and generic gene co-expression networks of colon cancer" INnovations in Intelligent SysTems and Applications (INISTA), International Symposium on. IEEE, 2017 (Doktora Tezinden)
2. Cingiz M. Özgür, Diri, B., "The inference of gene co-expression networks of breast and colon cancer using miRNA- target gene interactions data" 25. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı, 2017 (Doktora Tezinden)
3. Yamuç, A., Cingiz, M. Özgür, Biricik, G. and Kalıpsız O. " Solving Test Suite Reduction Problem Using Greedy and Genetic Algorithms." Electronics, Computers and Artificial Intelligence (ECAI), 2017 9th International Conference on. IEEE, 2017
4. Cingiz, M. Özgür, Şefik Temei, and Oya Kahpsız. "Test case prioritization with

- improved genetic algorithm." 2014 22nd Signal Processing and Communications Applications Conference (SIU). IEEE, 2014.
5. Temel, Sefik, Mustafa Özgür Cingiz, and Oya Kalipsiz. "Hibrid Algoritma ve Isıl İşlem Algoritmasıyla Test Kümesi Önceliklendirilmesi." UYMS. 2014.
 6. Unudulmaz, Ahmet, Oya Kalipsiz, and M. Özgür Cingiz. "Risk Faktörleri ve Risk Değerlendirme Modellerinin Farklı Veri Setleri Üzerinde Gerçeklenmesi.", UYMS, 2013
 7. Cihan Pınar., Kalipsiz Oya, Cingiz M. Özgür., & Doksöz Mahir. "Yazılım Geliştirme Dersleri Öğrenci Projelerinin Birliktelik Kuralı Göre Değerlendirilmesi", UYMS, 2013
 8. Albayrak, Abdülkadir, M. Özgür Cingiz, and M. Fatih Amasyalı. "Evaluation of robustness of ensemble learners to noisy data." Signal Processing and Communications Applications Conference (SIU), 2013 21st. IEEE, 2013.
 9. Cingiz, M. Özgür, Ahmet Unudulmaz, and Oya Kalipsiz. "Prediction of project problem effects on software risk factors." Intelligent Software Methodologies, Tools and Techniques (SoMeT), 2013 IEEE 12th International Conference on. IEEE, 2013
 10. Cingiz, M. Özgür, and Banu Diri. "Classification of microblogging users." 2012 20th Signal Processing and Communications Applications Conference (SIU). IEEE, 2012.
 11. Cingiz, M. Özgür, Ahmet Unudulmaz, and O. Kalipsiz. "Farklı Kestirim Yaklaşımları Kullanılarak Yazılım Projelerindeki Problem Etkilerinin Risk Yönetiminin Gerçeklemesi." Ulusal Yazılım Mühendisliği Sempozyumu (UYMS), Ankara (2012).
 12. Unudulmaz, Ahmet, M. Özgür Cingiz, and Oya Kalipsiz. "Yazılım Projelerindeki Problem Etkilerinin Yazılım Mimarisi ile İlişkilendirilmesi.", 4th Turkish National Software Architecture Symposium (UYMK), 2012

13. Cingiz, M. Özgür, and Banu Diri. "Content Mining of Microblogs." Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on. IEEE, 2012.

Kitap

1. Cingiz, M. Ö., & Diri, B. (2014). Content Mining of Microblogs. State of the Art Applications of Social Network Analysis (pp. 359-369): Springer. (extended version of ASONAM 2012)