BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MikroRNA VERİ TABANLARINDA BİLGİ GERİ-GETİRİMİ

KORAY AÇICI

YÜKSEK LİSANS TEZİ
2015

MikroRNA VERİ TABANLARINDA BİLGİ GERİ-GETİRİMİ

INFORMATION RETRIEVAL IN MicroRNA DATABASES

KORAY AÇICI

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2015

“MikroRNA Veri Tabanlarında Bilgi Geri-Getirimi” başlıklı bu çalışma, jürimiz tarafından, 18/08/2015 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan : Doç. Dr. Mustafa DOĞAN

Üye (Danışman) : Doç. Dr. Hasan OĞUL

Üye : Yrd. Doç. Dr. Bala Gür DEDEOĞLU

ONAY

..../08/2015

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Sayın Doç. Dr. Hasan OĞUL'a (tez danışmanı), çalışmanın sonuca ulaştırılmasında ve karşılaşılan güçlüklerin aşılmasında her zaman yardımcı ve yol gösterici olduğu için…

Sayın Dr. Mehmet DİKMEN'e her zaman yardımcı ve yol gösterici olduğu için…

Sayın Dr. Fatma AÇICI'ya her zaman yanımda olduğu için…

Sayın Araş. Gör. Tunç AŞUROĞLU'na her zaman yardımcı olduğu için…

Bu çalışma TÜBİTAK tarafından 113E527 nolu proje kapsamında desteklenmiştir.

ÖZ

MikroRNA VERİ TABANLARINDA BİLGİ GERİ-GETİRİMİ

Koray AÇICI

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Büyük ve açık veri tabanlarında bulunan biyolojik deneylerin içerik tabanlı geri getirimi biyoenformatik ve hesaplamalı biyolojide güncel bir problemdir. İçerik-tabanlı getirimde genel olarak, herhangi bir deneysel üst-veri içermeyen örnek sorgu kullanarak bir veri tabanında arama yapılabilmesi hedeflenmektedir. Bu çalışmada özel olarak mikrodizi veri tabanlarında ilgili mikroRNA deneylerinin geri-getirimi problemine odaklanılmıştır. Bunun için işlemsel bir alt yapı önerilmiştir. Mikrodizi profil deneylerinde farklı ifade olan mikroRNA'ları belirlemek için bir normal-tekdüze karışım modeli alt yapıya uyarlanmıştır. Ayrıca mikrodizi deney içeriğini temsil etmek için bilgi-tabanlı bir yöntem önerilmiştir. Ölçülmüş ifade verisi üzerinde istatistiksel zenginleşme analizi için mikroRNA'ların kemoterapi direncini temel alan bir dizi ek açıklamalı mikroRNA kümesi kullanılmıştır. Farklı ifade değerlerini temel alan gerçek-değerli deney imzalarını ikili hale çevirmede kullanmak üzere sıra-tabanlı bir eşikleme yöntemi önerilmiştir. Kategorik imzaları karşılaştırmak için etkili bir benzerlik ölçütü tanıtılarak iki deney arasındaki ilgililiğin ortaya çıkarılmasında kullanılmıştır. Önerilen modelin geri-getirim kabiliyetini ayırt etmek için deneysel ilgililik iki farklı bakış açısı ile değerlendirilmiştir. Birincisi hastalık ilişkisi, ikincisi ise embriyonik köken ortaklığıdır. Bilindiği kadarıyla deney geri getirim problemi, mikroRNA mikrodizileri bağlamında ilk defa bu çalışma ile incelenmiştir.

ANAHTAR SÖZCÜKLER: mikroRNA, mikrodizi, içerik-tabanlı bilgi geri getirimi.

Danışman: Doç. Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

INFORMATION RETRIEVAL IN MicroRNA DATABASES

Koray AÇICI

Baskent University Institute of Science and Engineering

Department of Computer Engineering

Content-based retrieval of biological experiments in large public repositories is a recent challenge in computational biology and bioinformatics. The task is, in general, to search in a database using a query-by-example without any experimental meta-data annotation. Here, a more specific problem that seeks a solution for retrieving relevant microRNA experiments from microarray repositories is considered. A computational framework is proposed with this objective. The framework adapts a normal-uniform mixture model for identifying differentially expressed microRNAs in microarray profiling experiments. Also a knowledge-based approach for representing microarray experiment content is proposed. A group of annotated microRNA sets based on their chemotherapy resistance are used for a statistical enrichment analysis over observed expression data. A rank-based thresholding scheme is offered to binarize real-valued experiment fingerprints based on differential expression. An effective similarity metric is introduced to compare categorical fingerprints, which in turn infers the relevance between two experiments. Two different views of experimental relevance are evaluated, one for disease association and another for embryonic germ layer, to discern the retrieval ability of the proposed model. To the best of one's knowledge, the experiment retrieval task is investigated for the first time in the context of microRNA microarrays.

KEYWORDS: microRNA, microarray, content-based information retrieval.

Advisor: Assoc. Prof. Dr. Hasan OĞUL, Başkent University, Department of Computer Engineering.

İÇİNDEKİLER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

DNA	Deoksiribonükleik asit
RNA	Ribonükleik asit
miRNA	mikroRNA
mRNA	Mesajcı RNA
cDNA	Bütünleyici DNA
GEO	Gene Expression Omnibus
RNAi	RNA interferansı
GEST	Gene Expression Search Tool
GSEA	Gene Set Enrichment Analysis
SNP	Single Nucleotide Polymorphism
miRSNP	microRNA and SNP
CREAM	Chemotherapy ResistancE-Associated MiRSNP
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
TP	True Positive (Doğru Pozitif)
FP	False Positive (Yanlış Pozitif)
DE	Differential Expression (Farklı İfade)
RNA-Seq	RNA Sequencing

1. GİRİŞ

1.1 Motivasyon ve Tezin katkıları

Mikrodizi deneyleri sonucunda elde edilen gen ifade verileri, genomik süreçlerin ortaya çıkarılıp aydınlatılmasında yaygın olarak kullanılmaktadır. Mikrodizi deneylerinin sonuçları deneyleri yapan kişiler tarafından GEO (Gene Expression Omnibus) gibi büyük ve merkezi veri tabanlarına yüklenmekte ve diğer araştırmacıların kullanımına sunulmaktadır. Son yıllarda bu tip veri tabanlarının boyutlarının çok yüksek seviyelere ulaştığı gözlenmektedir. Araştırmacılar için bu dev derlemler, bilimsel araştırma ve klinik uygulamalar bakımından bir hazine olarak düşünülürken, bu derlemlerde yapısal olmayan sorguların kullanılamaması deney arama işlemini sınırlandırmaktadır. Veri tabanlarına daha önce yüklenmiş bir deneyi arayan kullanıcılar, organizma adı, yazar adı, deney açıklaması, hastalık türü, mikrodizi platformu gibi metinsel üst-veri (meta-data) bilgilerini kullanarak sorgulama yapmak zorundadırlar. Bu yüzden yapılan deney, araştırmacının yükleme sırasında deneyle ilgili hangi ek bilgileri girdiğiyle sınırlı kalmakta, içeriği geri plana düşmekte ve sorgulama sonucunda bulunma olasılığı azalmaktadır. Fakat son yıllarda deneyin içeriğine göre benzer deneylerin bulunması problemi, ileri biyolojik bilgi keşfinin ve biyomedikal karar destek sistemlerinin gelişimine sağlayacağı potansiyel olanaklar nedeniyle önem kazanmakta ve popülerleşmektedir. Bu problem bilgi geri getirim kuramında "içerik-tabanlı arama" olarak adlandırılmaktadır. İçerik-tabanlı arama, gen ifade veri tabanları için güncel bir konu haline gelmiş ve son beş yılda bu amaçla yapılmış çalışmalara rastlanmıştır. İlgili çalışmalardaki ulaşılmak istenilen sonuç; kullanıcının kendi deney sonucunu sisteme yükleyebileceği ve bu işlemin sonucunda benzer içeriğe sahip deneylerin kullanıcıya sunulabileceği bir ortamın ortaya çıkmasıdır. Bu sayede benzer sonuçlar alınan deneylerden hipotezler üretilebilecek, yeni algoritmalar doğrulanabilecek, yeni deneyler tasarlanabilecek ve bunlardan faydalanma imkanı doğacaktır.

Bu tez çalışmasında, büyük veri tabanlarına araştırmacılar tarafından yüklenen deney sonuçları içerisinde gizlenen bilgiden gerçek anlamıyla yararlanılmasını sağlamak için içerik-tabanlı bir bilgi geri getirim sisteminin oluşturulması hedeflenmiştir. Mevcut bilgi geri getirim sistemleri mRNA gen ifadelerini içeren

mikrodizi deneyleri üzerine yoğunlaşmıştır. MikroRNA gen ifadesine dayalı mikrodizi deneylerini temel alan içerik-tabanlı bilgi geri-getirim çalışmaları literatürde bulunmamaktadır. Bu tez, mikroRNA mikrodizi deney sonuçlarını barındıran veri tabanlarında içerik-tabanlı bilgi geri getirimini sağlamaya yönelik ilk çalışma olma özelliğine sahip olacaktır.

Tez kapsamında deneyler GEO veri tabanından toplanmış, ön işlemlerden geçirilerek hazır hale getirilmiştir. MikroRNA mikrodizi deney sonuçlarını karşılaştırmada her bir deneyi temsil edecek imzaların oluşturulması sağlanmıştır. İmza çıkarma işleminden sonra deney imzaları, bilinen ve tez kapsamında yeni geliştirilen benzerlik veya uzaklık ölçütleri kullanılarak karşılaştırılmış ve her ölçüt için bilgi geri getirim performansı hesaplanmıştır. Mevcut ölçütler kendi aralarında karşılaştırılmış, performansı en yüksek olan ölçüt için iyileştirmeler yapılmaya çalışılmıştır.

Bu çalışmanın katkısı aşağıdaki maddelerde sunulmaktadır:

- İlgili deney geri-getirimi problemi mikroRNA mikrodizileri bağlamında ilk defa bu tezde ele alınmıştır.
- Mikrodizi profil deneylerinde farklı ifade olan mikroRNA'ları belirlemek için olasılıksal bir normal-tekdüze karışım modeli uyarlanmıştır.
- Farklı ifade olma durumlarına dayalı deney imzalarını ikili hale getirmek için sıra-tabanlı dinamik bir eşikleme yöntemi önerilmiştir.
- İki deney arasındaki ilgililiği tespit etmek için kategorik imzaları karşılaştıran etkili bir benzerlik ölçütü tanıtılmıştır.
- Kapsamlı bir karşılaştırma kümesi biyoenformatik araştırmacılarının gelecekte yararlanması için sunulmuştur.

1.2 Alan Bilgisi

MikroRNA'lar (miRNA) 18-25 nt uzunluğunda, protein kodlamayan küçük RNA'lardır. Bu yapılar post-transkripsiyonel degredasyon veya translasyonel represyon süresince gen ifadesini baskılar [1]. İlk miRNA, lin-4, 1993 yılında genetik çalışmalarda sıkça kullanılan Caenorhabditis elegans türünde tespit edilmiştir [2]. MiRNA'ların en temel fonksiyonu RNA interferansıdır (RNAi). miRNA'lar, mRNA'ların bir komplementer (tamamlayıcı) dizilim (sekans) ile

çözülmesiyle, gen ifadesinin susturulmasını sağlayan mekanizmayı çalıştırırlar [3]. RNAi mekanizmasının keşfi moleküler biyoloji araştırmalarını hızlandırarak istenilen herhangi bir genin etkisiz hale getirilmesini sağlayan genetik teknolojilerin gelişimine öncülük etmiştir [4]. miRNA'ların gen düzenlemesi sürecindeki etkisi son yıllarda yapılan çeşitli çalışmalarla da ortaya konmuştur [5]. miRNA'lar bu süreçlerin dışında, kanser gibi bazı hastalıkların ilerlemesine de sebebiyet verir [6]. E2F1 proteini hücre bölünmesini düzenleyen bir proteindir. Bu protein iki tip miRNA tarafından inhibe edilmektedir. miRNA, mRNA'ya bağlanarak gen aktivitesine (hücre bölünmesi) etki eden proteinin çevrimini engelleyerek kontrolsüz hücre bölünmesine dolayısıyla kansere sebep olmaktadır [7].

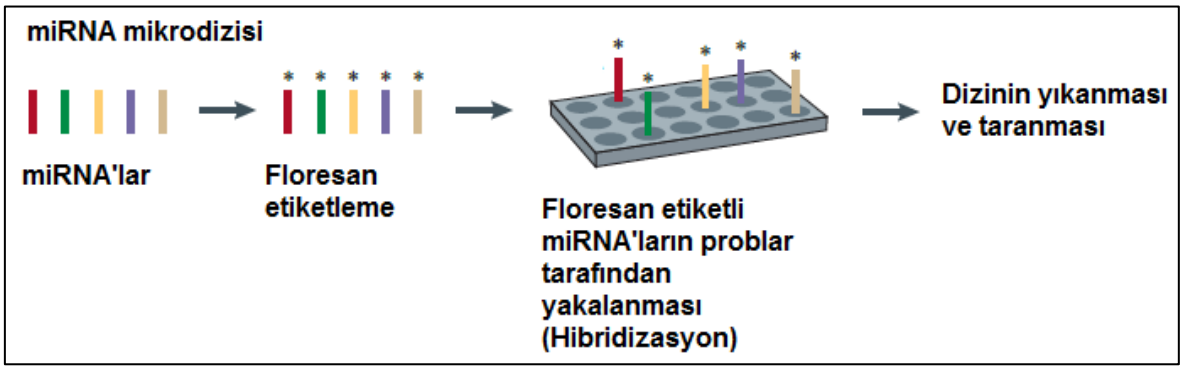
Mikrodiziler genomik araştırmalarda yaygın olarak kullanılan araçlardır. Son 20 yıllık zaman zarfında mikrodizi deneyleri sonucunda oluşan veri kümelerinin sayısındaki artış biyoenformatik ve makine öğrenme konularında yeni araştırmaları teşvik etmektedir. Mikrodiziler kullanılarak doku ve hücre numunelerindeki gen ifadesi farklılıklarına ilişkin veri toplanması hedeflenir. Mikrodizi deneyleri sonucunda elde edilen veriler, hastalık tanısında kullanılmak üzere veya spesifik tipteki tümörleri ayırt etmek için kullanılmaktadır [8].

Gen ifadesi mikrodizi teknolojisi, bir organizmada binlerce gen ifadesini eş zamanlı olarak ölçme yeteneğine sahip çok kuvvetli bir teknolojidir [9]. Prob-hedef hibridizasyonu bu teknolojinin merkezi konseptidir. Nükleik asit dizilimlerinin belirlenmesi ışınım-tabanlı (floresan) algılama yoluyla yapılmaktadır [10]. Biyolojik sistemlerdeki karmaşık moleküler etkileşimleri incelemek üzere mikrodizi teknolojisi giderek artan oranda kullanılmaktadır. miRNA'lara olan artan ilgiyle beraber mikrodizi teknolojisini de içeren en iyi yerleşmiş moleküler ve biyolojik teknolojiler miRNA araştırma alanına başarıyla transfer edilmiştir [11]. Günümüzde birçok ticari miRNA mikrodizi platformu mevcuttur. Bunlara örnek olarak Affymetrix, Agilent, Illumina, Ambion ve Exiqon verilebilir.

miRNA profilleme işleminde (miRNA ifadelerinin ölçülmesi) miRNA mikrodizisi kullanıldığında 4 temel adım vardır (Şekil 1.1):

- miRNA'ların floresan etiketlenmesi
- DNA-tabanlı yakalama probları dizisi ile hibridizasyon
- Dizinin yıkanması ve taranması

- Veri çıkarımı ve işleme



Şekil 1.1 miRNA deneylerinde adımlar (Pritchard et al.[12]'dan değiştirilerek)

Bu adımlar sonucunda bir grup miRNA için belirlenmiş koşullar altında ifade değerleri elde edilmiş olur.

1.3 Önceki Çalışmalar

Mikrodizilerin kullanıldığı bilimsel deneyler sonucunda elde edilen gen ifade verileri GEO ve ArrayExpress gibi büyük veri tabanlarına yüklenmekte ve buralarda saklanmaktadır [13; 14]. Bu veri tabanları bütün araştırmacıların kullanımına açıktır. Genel kullanım amaçlı veri tabanları haricinde spesifik organizmalara veya miRNA gibi moleküllere ait özel amaçlı veri tabanları da oluşturulmuştur [15; 16; 17]. İster genel amaçlı ister özel amaçlı olsun tüm gen ifadesi verilerini barındıran veri tabanlarında deney arama işlemi kullanıcıya sağlanan arayüz aracılığıyla çeşitli metinsel üst-verilerin (üzerinde deneyin yapıldığı organizmanın adı, mikrodizi platformu, miRNA adı, deneyin yayımlandığı bilimsel makale, yazar adı vb.) sisteme girilmesiyle yapılmaktadır [14; 18; 19; 20]. Bu yüzden arama sonucunda getirilen deney veya deneyler, kullanıcının aradığı deney için verdiği bilgiler ile sınırlanmaktadır. Ayrıca bu deneyler ve sonuçları biyolojik olarak anlam taşıyan önemli gizli bilgileri de içlerinde barındırabilmektedir. Tüm bu sebeplerden ötürü bahsi geçen büyük veri tabanlarında, deneylerin içerik-tabanlı geri getirimi için işlemsel yöntemlere ihtiyaç duyulmuştur.

Gen ifade veri tabanlarında içerik-tabanlı aramaya yönelik bilinen ilk çalışma Hunter et al. [21] tarafından önerilen Gene Expression Search Tool (GEST)

yöntemidir. GEST yönteminde iki deney Bayes tabanlı bir benzerlik ölçütü kullanılarak karşılaştırılmaktadır. Bahsedilen çalışmada herhangi bir koşulda birden fazla gen ifade değerini içeren bir dizi profili bir deney olarak kabul edilmektedir. Makalede yöntemin benzer deneyi bulma açısından başarılı olmadığı belirtilse de fikri ilk ortaya atan çalışma olması sebebiyle önem arz etmektedir.

Önceden sonuçları elde edilmiş deneyleri tekrar kullanarak yeni bir deneyde elde edilen verilerin analizinin yapılabileceği fikri Tanay et al. [22] tarafından tekrar vurgulanmıştır. Yazarlar çalışmalarında, daha önce yaptıkları deneylerden elde ettikleri ikili-küme (bicluster) aktiviteleriyle yeni yaptıkları deneyde bulunan ikili-kümeleri ilişkilendirebilmişlerdir. Daha önceden yapılmış deneylerin sonuçlarını kullanarak herhangi bir hastalıkla veya ilaç uygulamasına karşı oluşan hücresel tepkilerle ilgili biyolojik hipotezler üretmeyi başarabilen çalışmalar da bulunmaktadır [23; 24; 25].

Gen ifade verilerini içeren bir veri tabanı üzerinde içerik-tabanlı arama amaçlı çalışmalar 2006 yılından itibaren ortaya çıkmaya başlamıştır. Bu çalışmalar, verilen iki durum arasında (kontrol durumu ve farklı herhangi bir koşul) genlerin birlikte ifade olması (co-expression) veya farklı ifade olması (differential expression) üzerine kurulmuştur.

Horton et al. [26] RaPiDS olarak adlandırdıkları basit bir profil karşılaştırma algoritması önermişlerdir. Çalışmalarında iki gen ifadesi profili arasındaki benzerliği hesaplamak için Spearman sıralama katsayısı kullanmışlardır. RaPiDS algoritmasını kullanarak Fujibuchi et al. [27] CellMontage isimli uygulamayı yaratmışlardır. CellMontage uygulaması kullanıcı tarafından girilen bir sorgu profilini büyük bir veri tabanında arayan ilk uygulama olma özelliğine sahiptir. Uygulamanın kullanılarak arama işleminin yapılması sonucunda yalnız aynı hücre veya doku türünden olan profillerin belirli bir doğruluk oranında getirilebildiği ortaya konmuştur. Bahsi geçen yöntemle ilgili ikinci bir problem, gen sayısının çok fazla sayıda olması halinde (genelde 30000'in üzerinde) arama işleminin zaman açısından çok verimsiz olmasıdır. Chen et al. [28] farklı ifade olan genleri temel alan görsel bir uygulama geliştirmişlerdir. Uygulama bir gen listesini girdi olarak kabul edip aynı genlerin farklı ifade olduğu benzer deneyleri basit bir karşılaştırma ölçütü kullanarak arayabilme özelliğine sahiptir. GeneChaser olarak isimlendirilen

bu uygulamanın dezavantajı aynı anda az sayıda genin aranmasına olanak tanımasıdır. Bu sebeple bir deneyi temsil edebilecek tüm genlerin modellenmesini sağlayamamaktadır. Verilen gen listesi ile ilgili benzer deneylerin bulunma olasılığı azalmaktadır. Hibbs et al. [29] tarafından benzer bir algoritma birlikte ifade olan genler için geliştirilmiştir. Engreitz et al. [30] deneylerin daha hızlı aranmasını sağlayacak verimli bir yöntem üzerine çalışmışlardır. İlgili çalışmada deneydeki bütün farklı ifade değerleri alınmış, bir özellik azaltma algoritması kullanılarak (Bağımsız Bileşen Analizi, ICA) deneyi temsil eden daha küçük boyutta bir imzaya dönüştürülmüştür. Deney karşılaştırmaları bu imzalar üzerinden yapılmıştır. Deneyin imzaya dönüştürülerek temsil edilmesi yöntemi, ProfileChaser olarak adlandırılan bir web sunucusu haline getirilerek kullanıma sunulmuştur [31]. Bell and Sacan [32] çalışmalarında ikili (binary) imza yöntemini kullanarak içerik tabanlı arama için daha verimli bir yaklaşım sunmuşlardır. Bu yaklaşımda özellik sayısının azaltılması yerine farklı ifade değerleri kategorize edilerek bu kategorilerin ikili sayılara çevrilmesi sağlanmıştır. Böylelikle deneyler için daha hızlı karşılaştırma yapabilme olanağı sağlanmıştır. İlgili çalışma ciddi bir bilgi kaybı yaşanmadan farklı ifade kategorilerinin ifade farkının gerçek değeri yerine kullanılabileceğini göstermiştir.

Caldas et al. [33] bir deneyi genlerin kullanılmasıyla oluşturulan bir imza ile değil, gen kümeleri kullanılarak tanımlanan bir imza ile temsil etmeyi önermişlerdir. Bu düşünceyi hayata geçirmek için gen kümesi zenginleşme analizi (Gene Set Enrichment Analysis, GSEA) olarak adlandırılan bir yöntemden yararlanmışlardır. Subramanian et al. [34] tarafından geliştirilen GSEA yöntemi, iki farklı durum (fenotip) için ifade verileri olan bir grup gen içerisinde önceden tanımlanmış gen kümelerinden hangilerinin fenotip farklılaşması açısından etkin olduğunu ortaya çıkarmaktadır. Başka bir ifade ile iki durum arasında farklılaşan genlerin oluşturduğu fonksiyonel alt kümeleri tespit etmektedir. Bir deney için bu kümelerin tespit edilebilmesi durumunda deney artık genler ile değil gen kümeleri ile indekslenmiş bir imza ile temsil edilebilmektedir. Bu sayede deney karşılaştırmaları bu imzalar kullanılarak yapılabilmektedir. Suthram et al. [25] benzer bir çalışmayı GSEA kullanılarak elde edilen değil ağ-tabanlı bir yöntem ile elde ettikleri gen kümeleri üzerinden yapmışlardır. Geliştirdikleri yöntemi kullanarak herhangi bir hastalıkla ilgili deneyleri geri getirmeye çalışmışlardır.

Çalışmalarındaki gen kümeleri, protein-protein etkileşim ağları kullanılarak çıkarılan fonksiyonel gen modüllerine göre oluşturulmuştur.

Le et al. [35] çalışmalarında Spearman ilintisine dayalı karşılaştırma yöntemi kullanarak bir organizmada yapılan deneylerin başka bir organizmada yapılan deneyler üzerinde sorgulanabilmesini sağlamışlardır. Georgii et al. [36] önceden tanımlanmış bir gen listesi kullanılarak oluşturulan düzenleyici bir modelle arama yapma üzerine çalışmışlardır.

Literatürdeki mevcut yöntemler bir deneyi, deneyi temsil edecek sabit uzunlukta bir imzaya dönüştürmekte ve bu imzalar arasında, imzanın yapısına uygun, bir uzaklık veya benzerlik ölçütü kullanarak karşılaştırma yapmaktadır. Oluşturulan imzalar genelde iki türlü gösterilmektedir:

- Global bir gen ifadesi listesinde yer alan gen ifadelerinin nasıl veya hangi seviyede farklı ifade olduklarını gösteren nümerik değerler ile
- Önceden tanımlanan gen ifade kümelerinin bir listesi üzerinde, kümelerin ilgili deneyde etkin olup olmadıklarını gösteren bir etiket ile

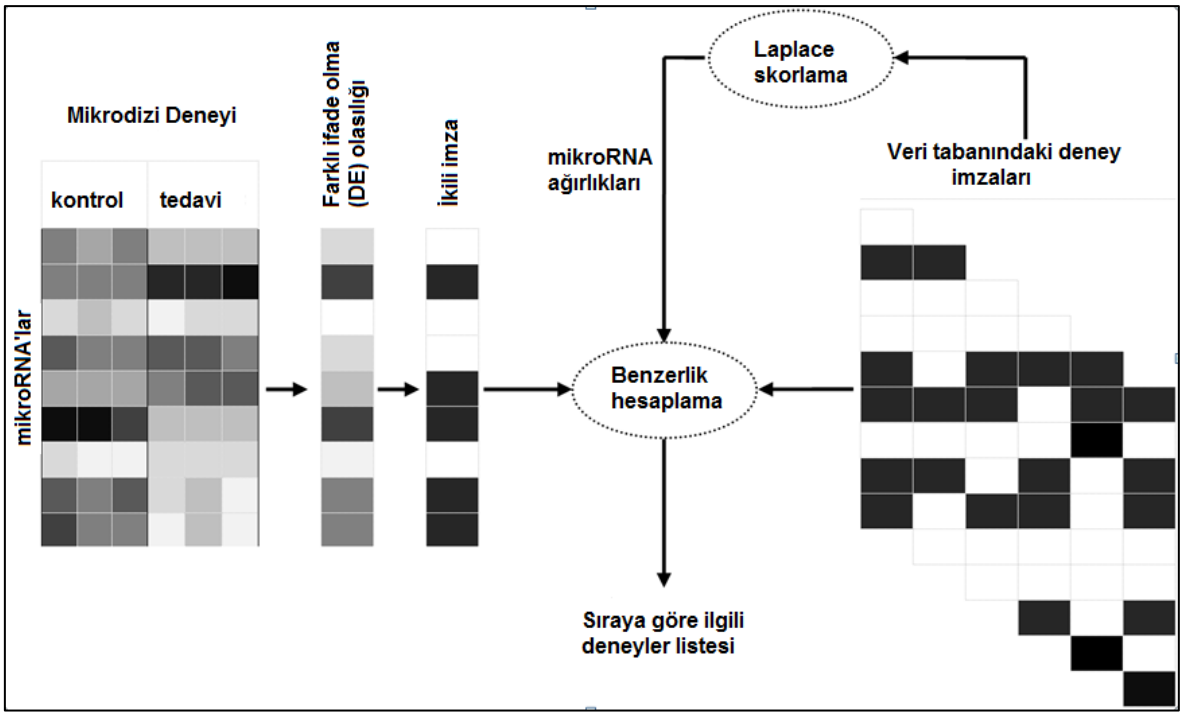
İmzaların çıkarımına ve karşılaştırma ölçütlerine göre kullanılan yöntemler çalışmalarda farklılık göstermektedir. Çalışmalardaki ortak yan, binlerce gen ifadesinden oluşan bir deneyin, iki farklı durum (koşul veya fenotip) için ölçülmüş gen ifade değerleri olarak kabul edilmesidir. Gen ifade veri tabanları üzerine, bildiğimiz kadarıyla, yapılan çalışmalarda miRNA deney geri getirimi şimdiye kadar dikkate alınmamıştır. Chen et al. [37] çalışmalarında deney arama yerine sadece bir gen profilinin aranması üzerine eğilmişlerdir.

2. YÖNTEMLER

2.1 Bilgi Geri Getirim Modeli

Deney geri-getirim problemi, örnek olarak verilen bir sorgu deneyine içerik olarak benzer deneylerin veri tabanında aranarak benzerlik sırasına göre raporlanması şeklinde tanımlanabilir. Bu çalışmada deney diye tabir edilen, bir grup miRNA için iki farklı koşulda ölçümleri yapılmış mikrodizi çalışmasının sonucudur. Daha biçimsel haliyle ilgili miRNA mikrodizi deney geri getirimi aşağıda belirtildiği gibi anlatılabilir.

Verilen bir miRNA ifade matrisi E'de e_{ij}, j. durumdaki i. mikroRNA'nın ifadesini temsil eder. Mikrodizi veri havuzundaki $\{M_1, M_2, \ldots, M_t\}$ matris derlemleri arasında, k'nın en yüksek benzerliği $s(E, M_k)$ gösterdiğinin kabul edildiği M_k matrisinin bulunması ise amaç olarak tanımlanır. İki tam ifade matrisi arasında bir benzerliğin tanımlanması gerçekleştirilebilir olmadığından karşılaştırma geri getirim uygulama çerçevesi içinde deney içeriğini temsil eden, imza adı verilen, daha basit, bir boyutlu bir vektör üzerinden yapılır. Bir veri tabanı araması sırasında sorgu olarak verilen deneyin karşılaştırılması, veri tabanında halihazırda mevcut olan diğer imzalar üzerinden, deneyin imzası aracılığıyla yapılır. Bu yüzden içerik-tabanlı geri getirim stratejisinin başarılı bir gerçekleştirimi; birinci olarak, verilen deneyden temsilci bir imzanın nasıl türetileceği ve ikinci olarak ise iki imzanın etkin bir biçimde nasıl karşılaştırılacağı sorularına verilecek cevaplara bağlıdır. Karşılaştırılan deney matrislerinin satırlarının (miRNA listeleri) eşli olması veya bu şekilde düzenlenebileceği mantıklı bir varsayımdır. Daha sonrasında bilgi geri getirim modeli iki matrisin bütün miRNA ifade profillerinin imzalarının eşleşmesi üzerine kurulabilir. Diğer bir ifade ile tüm iki matris yerine E ve M_k'nın imzaları üzerinden benzerlik $s(E, M_k)$ tanımlanır. Bu çalışmada kontrol ve tedavi olmak üzere iki koşula sahip olan miRNA mikrodizi deneyleri üzerine odaklanılmıştır. Bilgi geri getirim modelinin genel bir görünümü Şekil 2.1'de görülmektedir.



Şekil 2.1 Geri getirim modelinin genel görünümü

2.2 İmza Tasarımı

2.2.1 Farklı ifade tabanlı imza çıkarımı

Bir miRNA mikrodizi deneyi, deneyde ölçülen bütün miRNA'ların farklı ifade olma değerlerinin bir vektörü olarak tasarlanan bir imza ile temsil edilmektedir. miRNA i'nin farklı ifade olması, z_i değişkeni ile ifade edilir, iki deneysel koşul (kontrol, tedavi) arasında farklı ifadelenen miRNA'nın olasılık değeri ile ölçülür. Dean and Raftery [38] tarafından ortaya konulan, veriyi sabit ve farklı ifade olmuş miRNA'ların normal-tekdüze karışımına uydurarak, z_i'yi tahmin etmeye çalışan yöntem kullanılarak miRNA mikrodizi deney sonuçlarından deney imzası çıkarılmıştır. Model (2.1) eşitliğindeki ifade ile verilmiştir.

$$r_i \sim pN(r_i \mid \mu, \sigma^2) + (1-p)U(r_i) \qquad (2.1)$$

Burada r_i, miRNA i için gözlenmiş normalize edilmiş logaritma oranını; p, farklı ifade olunma öncel olasılığını; $N(r_i|\mu,\sigma^2)$, ortalama μ ve varyans σ^2 ile normal dağılımı; $U(r_i)$ ise sınırlı bir aralıkta tekdüze dağılımı temsil etmektedir. Bir beklenti büyütme (expectation maximization) algoritması; p, μ ve σ^2 için parametre

tahminleri verilen z_i sonsal olasılığını direkt olarak hesaplayabilir. Bu da imza vektöründe karşılık gelen miRNA profili için imzayı niceleyen gizli değişkenin değerine işaret eder. k. beklenti adımında z_i'nin değeri mevcut parametre tahminleri kullanılarak (2.2) eşitliğinde belirtildiği gibi kestirilebilir.

$$\widehat{z_i}^{(k)} = \frac{\left(1-\hat{p}^{(k-1)}\right)U(r_i)}{\hat{p}^{(k-1)}N\left(r_i|\hat{\mu}^{(k-1)},(\hat{\sigma}^{(k-1)})^2\right)+\left(1-\hat{p}^{(k-1)}\right)U(r_i)} \tag{2.2}$$

Bir büyütme adımı beklenti adımını takip eder ve farklı ifade olma olasılıklarının verilen mevcut kestirimlerini kullanarak modelin p, μ ve σ^2 parametrelerini (2.3), (2.4) ve (2.5) eşitliklerinde belirtildiği gibi tahmin eder.

$$\hat{p}^{(k)} = \frac{\sum_i(1-\widehat{z_i}^{(k)})}{n} \tag{2.3}$$

$$\hat{\mu}^{(k)} = \frac{\sum_i((1-\widehat{z_i}^{(k)})\times r_i)}{\sum_i(1-\widehat{z_i}^{(k)})} \tag{2.4}$$

$$(\hat{\sigma}^{(k)})^2 = \frac{\sum_i((1-\widehat{z_i}^{(k)})\times(r_i-\hat{\mu}^{(k)})^2)}{\sum_i(1-\widehat{z_i}^{(k)})} \tag{2.5}$$

Farklı ifade olma olasılığı değeri aktif bir eşikleme yöntemi ile ikili sisteme çevrilir. İkili gösterimin, ham ifade verisi ve normal-tekdüze karışım modeli kullanılarak verinin işlenmiş örnekleri içindeki gürültünün etkisini azaltacağı düşünülmektedir. İkili sisteme çevirme her deney için ayrı bir eşik değeri seçilerek yapılır. Eşik değeri, sıralı listenin üst basamaklarında yer alan miRNA'ların, bütün miRNA'ların belirli bir yüzdesine karşılık gelen (%K olarak kabul edilirse), farklı ifade edilmiş olarak deney imzasında kapsanmasını garanti altına alır. Sabit bir eşik değeri seçmek yerine bu dinamik eşik yönteminin kullanılması bir deneyin yeterince yoğun bir vektör tarafından temsil edilmesine olanak sağlamaktadır. K değeri herhangi bir ilgililik (benzerlik) yönteminde geri getirim performansını maksimize etmeyi amaçlamak üzere global olarak kapsamlı bir arama tarafından belirlenir. Böylece deneye özgü t eşik değeri sıralı listenin en üst %K bölümündeki son

miRNA'nın farklı ifade olma olasılık değerine eşit olur. Bir deney için optimal bir t değeri bulunduğunda her farklı ifade olma değeri, z_i, ikili bir farklı ifade durumuna, x_i, eşlenir. Eşleme işlemi (2.6) eşitliğinde gösterilmektedir.

$$x_i = \begin{cases} 1, & eğer\ z_i > t \\ 0, & diğer \end{cases} \quad (2.6)$$

Bir *E* deneyi, n tane miRNA girdisi içeren bir mikrodizi için $X=\{x_1,x_2, \ldots, x_n\}$ imza vektörü ile temsil edilir. Kontrol ve tedavi koşulları arasında farklı ifadedeki yönün miRNA'nın davranışının çıkarımında değerli bir bilgi olduğu göz önünde bulundurularak, imzada azalan ifadeye sahip girdiler -1 ile çarpılıp negatif yönlü olarak işaretlenmektedir.

2.2.2 miRNA kümeleri tabanlı imza çıkarımı

Mikrodizi deney sonuçlarından imza oluşturmak için bir diğer yöntem ayrı ayrı gen ifadelerinin yerine gen ifade kümelerinin kullanılmasıdır. Bu yöntemde oluşturulan imzadaki her bir girdi ölçülmüş ifade verisine bağlı olarak zenginleşme skorunu belirtir [33; 36]. Bu yaklaşım sonuçların biyolojik olarak yorumlanabilirliği açısından daha fazla umut vaat eden bir yaklaşımdır. Fakat her içerik için güvenilir kümelerin eksikliği bu yaklaşımın yaygın olarak kullanılmasına engel teşkil etmektedir. Örneğin, mevcut küme zenginleşme analiz uygulamalarından hiç biri miRNA'lar için önceden tanımlanmış fonksiyonel kümeler sunamamaktadır. Bu yüzden şimdiye kadar miRNA mikrodizi deneyleri için küme zenginleşme analizi kavramı kullanılmamıştır. Bu amaçla kemoterapi direnci üzerindeki benzerliklerine göre oluşturulan miRNA kümelerinin kabiliyetleri değerlendirilmiştir. Kemoterapi direnci, tedavi koşulu altındaki tümörün tipiyle çok ilişkilidir. Buradan hareketle deneylerin miRNA kümeleri kullanılarak bilgi-tabanlı imzasının oluşturulması ve aynı hastalıkla ilgili farklı mikrodizi deneyleri arasındaki ilgililiğin belirlenmesi hedeflenmiştir.

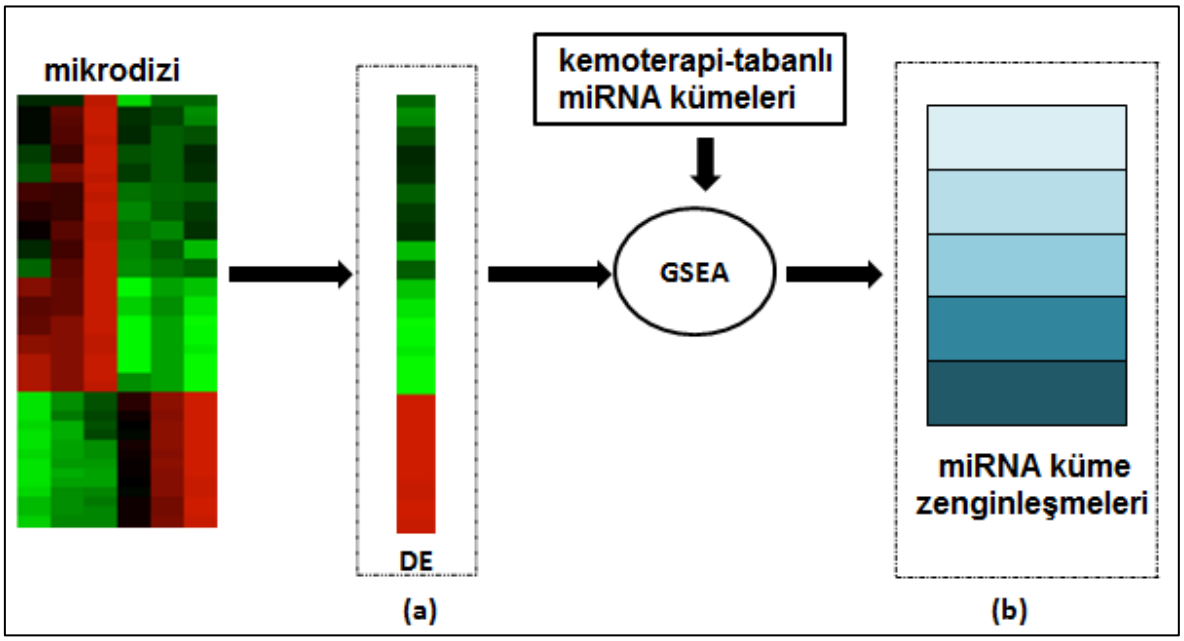
miRNA'lardan direkt olarak bir imza oluşturmak yerine kemoterapi direncini temel alan önceden tanımlanmış miRNA kümeleri kullanılarak ikinci imza çıkarma yöntemi oluşturulmuştur. Bu durumda bir deney imzasının uzunluğu veri havuzundaki miRNA kümelerinin sayısına eşit hale gelmektedir (Şekil 2.2). Deney

imzasının oluşturulmasına katkıda bulunan her miRNA kümesinin değeri Gene Set Enrichment Analysis (GSEA) algoritması kullanılarak hesaplanmaktadır.

GSEA, muhtemel tanımlı gen kümelerinin iki biyolojik durum arasında istatistiksel olarak kayda değer, uyumlu farklar gösterip göstermediğini belirleyen bir hesaplama metodudur [34]. Farklı ifade değerleriyle birlikte verilen bir miRNA listesi için GSEA algoritması aşağıdaki adımları takip etmektedir:

1) Farklı ifade değerine göre miRNA'ları sırala
2) Her bir miRNA kümesi için:
    2.1) Sıralanmış miRNA'lar için kümülatif toplamı hesapla
    2.2) Sıfırdan maksimum sapmayı zenginleşme skoru (ES) olarak bildir
3) ES değerlerine göre sıralanmış miRNA kümelerini bildir

2.1 adımı toplama ilk değer atanmasıyla başlar. Daha sonra sıralı miRNA listesi kullanılarak, mevcut miRNA kümede bulunuyor ise toplam değeri artırılır, bulunmuyor ise toplam değeri azaltılır. Artırımın büyüklüğü miRNA'nın incelenmekte olan fenotip ile olan korelasyonuna bağlıdır.



Şekil 2.2 miRNA mikrodizi deneyleri için imzalar: (a) farklı ifade tabanlı, (b) yeni kemoterapi direnci-tabanlı imza

Mevcut GSEA uygulamaları fonksiyonel miRNA kümeleri için hazır listeler sağlayamadığından kemoterapi direncine dayalı miRNA kümeleri oluşturulmuştur. Gerekli ilişkiler CREAM veri tabanından temin edilmiştir [39]. CREAM veri tabanı yüksek-çıktılı multi-omik veriden tespit edilen kemoterapi direnç-ilişkili miRSNP'lerin keşfi ve depolanması için geniş kapsamlı bir veri havuzu sağlamaktadır. Bu koleksiyon kullanılarak aynı kemoterapi direnci ile ilişkili miRNA'ların kümesi çıkarılabilmektedir. Buradan her biri farklı bir direnç tipine karşılık gelen 276 tane miRNA kümesi oluşturulmuştur.

GSEA her bir kemoterapi-tabanlı miRNA kümesi için bir zenginleşme skoru üretmektedir. Bu skorun olduğu gibi kullanılması yerine deneyde önemli ölçüde zenginleşen miRNA kümeleri seçilir ve imza değerleri 1 yapılır. Bu da miRNA kümesinin deneyde fonksiyonel olduğunu belirtmektedir. Eğer zenginleşme skoru bir eşik değerinden küçük ise miRNA kümesine karşılık gelen imza değeri 0 olarak belirlenir. Eşik değeri her bir deney için aktif olarak seçilir. Aktif bir seçim gerçekleştirmek için kümelerin %K kısmı, ki bunlar miRNA kümelerinin sıralı listesi içinde 0.05'ten daha küçük bir p-değerine erişen kümelerdir, seçilir. Seçilen girdiler önemli ölçüde zenginleşmiş olarak belirlenir. K'nın optimal değeri, geri getirim performansını maksimize etmek için, bütün deneyler üzerinde yapılacak geniş kapsamlı bir arama ile bulunur.

2.3 İmza Karşılaştırma

İki deneyin ilgililiğinden sonuç çıkarma işlemi deneylerin imzaları arasında bir benzerlik skoru hesaplanarak yapılır. İkili olmayan deney imzalarının karşılaştırılmasında kullanılan ölçütler Euclid uzaklığı, Pearson korelasyon katsayısı, Spearman korelasyon katsayısı, Bhattacharyya uzaklığı ve Cosine benzerliğidir. İkili imzaların karşılaştırılmasında Jaccard benzerlik katsayısı ve Tanimoto benzerliği ölçütleri kullanılmıştır. Ayrıca deney imzasındaki her bir miRNA için farklı bir ağırlık kullanan, Tanimoto ölçütünün genişletilmiş bir versiyonu olan yeni bir ölçüt, 'Ağırlıklandırılmış Tanimoto' ölçütü de ikili imzaların karşılaştırılmasında kullanılmıştır.

2.3.1 Euclid uzaklığı

İkili olmayan iki imza arasındaki basit geometrik uzaklığı ölçmek için kullanılmaktadır. İki imza arasında Euclid uzaklığı kullanılarak ölçülen değer ne kadar düşükse imzalar arasındaki benzerlik de o kadar yüksektir. X ve Y imzaları arasındaki Euclid uzaklığı (2.7) eşitliğinde belirtildiği gibi hesaplanır.

$$s(X,Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{2.7}$$

2.3.2 Bhattacharyya uzaklığı

Bhattacharrya uzaklığı iki sürekli olasılık dağılımı arasındaki uzaklığı ölçmek için kullanılmaktadır. İkili olmayan bir imza da sürekli olasılık değerlerinden oluştuğu için iki imza arasındaki uzaklığı ölçmek için kullanılabilir. İki imza arasındaki uzaklık ne kadar düşükse imzalar arasındaki benzerlik de o kadar yüksek olmaktadır. X ve Y imzaları arasındaki uzaklık (2.8) eşitliğinde belirtildiği gibi hesaplanır.

$$s(X,Y) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\sigma_x^2}{\sigma_y^2} + \frac{\sigma_y^2}{\sigma_x^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(\mu_x - \mu_y)^2}{\sigma_x^2 + \sigma_y^2}\right) \tag{2.8}$$

X ve Y iki farklı dağılım yani imza olmak üzere σ^2, her iki imza için varyansları; μ ise her iki imzadaki ortalama değerleri ifade etmektedir.

2.3.3 Pearson korelasyon katsayısı

Pearson korelasyon katsayısı iki değişken arasındaki doğrusal korelasyonu (bağımlılığı) bulmaya yarayan bir ölçüttür. İkili olmayan iki imzadaki aynı (eşli) miRNA'ların farklı ifade olma olasılıkları arasındaki korelasyonu ölçmek için kullanılabilir. Pearson korelasyon katsayısı kullanılarak elde edilen benzerlik değerleri [-1, 1] aralığında bulunmaktadır. İki imza arasındaki benzerlik değeri 1 ise iki imza arasında tamamen pozitif korelasyon olduğu sonucu çıkarılır. Başka bir ifade ile bir doğrusal eşitlik X ve Y imzaları arasındaki ilişkiyi mükemmel bir şekilde

tanımlamaktadır. Benzerlik değerinin 0 olması X ve Y imzaları arasında herhangi bir doğrusal korelasyon olmadığı anlamına gelmektedir. Benzerlik değerinin -1 olması durumunda X ve Y imzaları arasında tamamen negatif korelasyon olduğu sonucu çıkarılmaktadır. Diğer bir ifade ile imzalar arasında mükemmel negatif bir ilişki söz konusudur. Pearson korelasyon katsayısı (2.9) eşitliğinde belirtildiği gibi hesaplanır.

$$s(X,Y) = \frac{N\sum x_i y_i - \sum x_i y_i}{\sqrt{N\sum x_i^2 - \left(\sum x_i\right)^2} - \sqrt{N\sum y_i^2 - \left(\sum y_i\right)^2}} \tag{2.9}$$

X ve Y iki imza olmak üzere x_i ve y_i ifadeleri imzalardaki eşli miRNA'lar için farklı ifade olma olasılık değerlerini; Nise toplam eşli miRNA sayısını temsil etmektedir.

2.3.4 Cosine benzerliği

Cosine benzerliği iki vektör arasındaki açının kosinüsünü ölçerek vektörler arasındaki benzerliği hesaplamada kullanılan bir benzerlik ölçütüdür. Her imza bir vektör olarak ifade edildiğinden iki imzanın benzerliğini ölçmek için kullanılabilir. Cosine benzerlik ölçütü kullanıldığında alınan sonuçlar [0,1] aralığında bulunmaktadır. Eğer benzerlik sonucu 1 ise iki vektör aynı yönde ve aralarındaki açı 0°'dir. Bu durumda iki vektör aynıdır. Benzerlik sonucunun 0 çıkması durumunda ise iki vektör arasında dekorelasyon (ortogonallik) olduğu ve aralarındaki açının 90° olduğu ortaya çıkmaktadır. Sonuç olarak daha yüksek bir değer daha yüksek benzerliği ifade eder. Cosine benzerliği (2.10) eşitliğinde belirtildiği gibi hesaplanır.

$$s(X,Y) = \frac{\sum_{i=1}^{N} x_i \times y_i}{\sqrt{\sum_{i=1}^{N}\left(x_i\right)^2} \times \sqrt{\sum_{i=1}^{N}\left(y_i\right)^2}} \tag{2.10}$$

X ve Y iki imza olmak üzere x_i ve y_i ifadeleri imzalardaki eşli miRNA'ların farklı ifade olma olasılık değerlerini temsil etmektedir.

2.3.5 Spearman sıra korelasyon katsayısı

Spearman'ın sıra korelasyon katsayısı iki değişken arasındaki istatistiksel bağımlılığın ölçülmesinde kullanılan parametrik olmayan bir ölçüttür. Başka bir ifade ile iki değişken arasındaki ilişkinin monoton bir fonksiyon kullanılarak ne kadar iyi bir şekilde tanımlanabileceğini hesaplamaktadır. İkili olmayan iki imzada bulunan eşli miRNA'ların farklı ifade olma olasılık değerlerinin sıraları arasındaki kuvvetin ölçümüne olanak sağlar. Spearman ölçütü kullanılarak elde edilen benzerlik değerleri [-1,1] aralığında bulunmaktadır. Spearman benzerlik ölçütü (2.11) eşitliğinde belirtildiği gibi hesaplanır.

$$s(X,Y) = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)} \tag{2.11}$$

X ve Y iki imza olmak üzere d_i ifadesi eşli miRNA'ların sıraları arasındaki farkı, N ise eşli miRNA sayısını temsil etmektedir.

2.3.6 Jaccard benzerlik katsayısı

Jaccard indeksi olarak da bilinen Jaccard benzerlik katsayısı örneklem kümelerinin benzerliğini veya farklılığını karşılaştırmak için kullanılan bir istatistik ölçütüdür. Jaccard benzerlik katsayısı ikili imzaların karşılaştırılmasında kullanılabilir. Farklı ifade olma olasılık değerlerinden oluşan imza ikili imzaya iki yöntemle çevrilebilir. Birincisi eşik değerini baz alan yöntemdir. Bu yönteme göre belirlenen eşik değerinden büyük olan farklı ifadeler 1, küçük olan ifadeler de 0 ile temsil edilir. İkincisi yüzdelik dilimi baz alan yöntemdir. Bu yöntemde imzadaki farklı ifadeler büyükten küçüğe doğru sıralanarak imzanın belirlenen yüzdelik dilimi 1 ile geri kalanı da 0 ile temsil edilir. Böylelikle imzadaki her miRNA için farklı ifade olup olmama durumu 0 ve 1 ile gösterilmektedir. Jaccard benzerlik katsayısı kullanılarak iki ikili imzanın benzerlik skoru (2.12) eşitliğinde belirtildiği gibi hesaplanır. Benzerlik skoru [0,1] aralığında bulunmaktadır.

$$s(X,Y) = \frac{\sum_i (x_i \cap y_i)}{\sum_i (x_i \cup y_i)}, x_i \neq 0, y_i \neq 0 \quad (2.12)$$

X ve Y iki tane ikili imza olmak üzere iki imzadaki eşli miRNA'lar arasındaki kesişimin ve birleşimin büyüklüğü dikkate alınır. Her iki ikili imzada farklı ifade olmayan eşli miRNA'lar göz ardı edilir.

2.3.7 Tanimoto benzerlik katsayısı

İkili olmayan imzaların 2.3.6 bölümünde anlatılan yöntemler kullanılarak ikili imzalara çevrilmesiyle aralarındaki benzerliğin hesaplanmasında kullanılan bir benzerlik ölçütüdür. İkili imzalardaki eşli miRNA kategorileri, örneğin işaretli (+ veya -) ikili farklı ifadeler, arasındaki örtüşmeleri hesaplamak için kullanılabilir. Tanimoto benzerliği kullanıldığında elde edilen değer [0,1] aralığında bulunmaktadır. Tanimoto benzerlik katsayısı (2.13) eşitliğinde belirtildiği gibi hesaplanır.

$$s(X,Y) = \frac{\sum_i (x_i \cap y_i)}{\sum_i (x_i \cup y_i)} \quad (2.13)$$

X ve Y iki tane ikili imza olmak üzere iki imzadaki eşli miRNA'lar arasındaki kesişimlerin ve birleşimlerin büyüklüğü hesaplanır. Jaccard benzerlik ölçütünden farkı, her iki ikili imzada farklı ifade olmayan eşli miRNA'ların da hesaba katılmasıdır.

2.3.8 Yeni karşılaştırma yöntemi (Ağırlıklandırılmış Tanimoto)

Tüm derlem içerisindeki miRNA'ların bilgi içeriği temel alınarak her bir miRNA'nın ayrı ayrı katkısını dikkate alan iki tane işaretli (+ veya -) ikili imzayı karşılaştırmak için yeni bir benzerlik ölçüsü tanımlanmıştır. Bir miRNA'nın bilgi içeriği, sorgu deneyi dışarıda bırakılmak suretiyle tüm mikrodizi deneyleri derlemi kullanılarak hesaplanan Laplace skoru ile ölçülmektedir. Laplace skorlama yerellik koruma

gücüne göre öznitelikleri değerlendiren bir öznitelik seçim yöntemidir. Bu skorlama yöntemi, sınıf bilgisinin olmadığı gözetimsiz anlayışta, öznitelikleri ağırlıklandırmak için birkaç içerikte başarıyla uygulanmıştır [40].

Derlemde, (i=1,…,n; j=1,…,m) olmak üzere, j. mikrodizi deneyi $\mathbf{X^j}$ ile, j. deneydeki i. miRNA'nın imza değeri de x_i^j ile gösterilsin. Bu durumda i. miRNA'nın Laplace skoru, L_i, (2.14) eşitliğinde verildiği gibi hesaplanmaktadır.

$$L_i = \frac{\sum_{j,k}\left(x_i^j - x_i^k\right)^2 S_{jk}}{\sum_{j}\left(x_i^j - \mu_i\right)^2 D_{ii}} \tag{2.14}$$

$D = diag(S)$ olmak üzere S matrisi $\mathbf{X^j}$ imza vektörleri, (j=1,…,m), arasındaki komşuluk ilişkisine göre tanımlanmaktadır. S matrisi de (2.15) eşitliğinde belirtildiği gibi hesaplanmaktadır.

$$S_{jk} = \begin{cases} e^{-\|\mathbf{x^j}-\mathbf{x^k}\|}, & \boldsymbol{X^j} ve\ \boldsymbol{X^k}\ komşu\ ise \\ 0, & diğer\ durum \end{cases} \tag{2.15}$$

Eğer bir imza uygulanmış bir benzerlik ölçütü üzerinden, bu durumda Tanimoto ölçütüdür, diğer bir imzanın en yakın komşuları arasında yer alıyorsa iki imza komşu imzalar olarak tanımlanmaktadır.

Düşük bir L_i değeri yüksek ayırt edici gücü belirttiğinden her bir miRNA için $\ell_i = 1 - L_i$ eşitliği ile hesaplanan ℓ_i ağırlığı kullanılmıştır. Hesaplanmış ağırlıklar kullanılarak n uzunluklu iki imza vektörü, X ve Y, arasındaki benzerlik (2.16) eşitliğinde belirtildiği gibi bulunmaktadır.

$$s(X,Y) = \frac{\sum_{i=1}^{n} \ell_i\left(x_i \cap y_i\right)}{\sum_{i=1}^{n} \ell_i\left(x_i \cup y_i\right)} \tag{2.16}$$

3. SONUÇLAR

3.1 Veri Kümesi

Tez çalışmasında kullanılan deneyler GEO veri tabanından elde edilmiştir. Veri tabanına araştırmacılar tarafından yüklenip kullanıma sunulan her bir deney GSE ön eki ile başlayıp bir numara ile biten bir karakter dizisi ile temsil edilmektedir. Çalışmada kullanılan mikrodizi deneyleri Çizelge 3.1'de verilmiştir. Veri tabanındaki miRNA mikrodizi deneyleri sadece iki farklı fenotipten oluşabildiği gibi birden fazla kontrol (normal, sağlıklı durumu belirtir) durumundan ve birden fazla tedavi (hastalıklı veya farklı bir durumu belirtir) durumundan da oluşabilmektedir. Birden fazla kontrol durumunun olduğu deneylerde kontrol durumlarının ortalaması alınarak kontrol durumları tek kontrol durumuna indirgenmiştir. Teke indirgenen kontrol durumu ilgili deneydeki tedavi durumlarıyla veya diğer durumlarla eşleştirilerek iki fenotipli miRNA mikrodizi deneyleri elde edilmiştir. GEO veri tabanı kullanılarak farklı hastalıklara ait miRNA mikrodizi deneylerinden iki fenotipe (kontrol ve tedavi) sahip 135 tane deney oluşturulmuştur. Tüm derlem ek materyal olarak www.baskent.edu.tr/~hogul/mirsearch adresinde sunulmuştur.

miRNA mikrodizi deneyleri farklı platformlarda (Affymetrix, Agilent ve Illumina mikrodizi çipi markalarıyla yapılan deneyler kullanılmıştır) yapıldığından ölçüm değerleri her bir platform için farklılık göstermekte ve deney sonuçları farklı aralıkları kapsamaktadır. Bu yüzden her bir deneydeki durumlar için yapılan ölçümlerin değerleri, (0 - 1) aralığına çekilerek normalize edilmiştir. Bu işlem (3.1) eşitliği kullanılarak yapılmaktadır.

$$f(x_i) = \frac{(b-a)(x_i - min)}{(max - min)} + a \qquad (3.1)$$

x_i mikrodizi deneyindeki i. miRNA'nın mevcut değerini, a ve b yeni değer aralığının en küçük ve en büyük değerlerini, max ve min ise mikrodizi deneyindeki ölçülen en küçük ve en büyük miRNA değerlerini temsil etmektedir.

Çizelge 3.1 Mikrodizi deneyleri GSE numaraları

GSE No
GSE2564
GSE27430
GSE29248
GSE27606
GSE55025
GSE21394
GSE43571
GSE47056
GSE43249
GSE49470

Oluşturduğumuz deney derleminde, bütün deneylerdeki farklı miRNA'ları saydığımızda, toplam 1633 miRNA bulunmaktadır. Deneylerin hepsi Homo sapiens üzerinde yapılmıştır.

3.2 Deneysel Kurulum ve Değerlendirme

İçerik-tabanlı bir veri tabanı arama platformu, toplanmış veri kümesinin tüm bir veri tabanı olarak ele alınarak her deneyin sorgu olarak kabul edilip veri tabanında sorgulanması suretiyle, sorgu deneyi dışarıda bırakılarak, simüle edilmektedir. Sistem veri tabanında bulunan bütün deneyleri getirerek benzerlik skorlarına göre azalan bir şekilde sıralamaktadır. Bir deney ne kadar yüksek bir sıraya sahipse sorgu ile o kadar ilgili olması beklenmektedir. Böyle bir senaryoda geri getirim performansını değerlendirmek için yaygın bir yol Receiver Operating Characteristic (ROC) eğrilerini kullanmaktır.

Bir ROC grafiği; sınıflandırıcıların, performansları temel alınarak, görselleştirilmesi, organize edilmesi ve seçilmesi yöntemidir. ROC grafikleri sinyal algılama teorisinde sınıflandırıcıların isabet oranı ve yanlış alarm oranı arasındaki dengeyi resmetmek üzere uzun süredir kullanılmaktadır. İki sınıflı sınıflandırma problemlerinde her bir nesne pozitif ve negatif sınıf etiketlerinden {p, n} biriyle eşleştirilir. Bir sınıflandırma modeli veya sınıflandırıcı da nesnelerden tahmin edilen sınıflara eşleşmeyi yapar. Model tarafından üretilen sınıf tahminlerinde nesnenin gerçek sınıfı ile tahmin edilen sınıfını ayırt etmek için {**E**vet, **H**ayır} gibi etiketler kullanılır. Eğer nesnenin sınıfı pozitifse ve pozitif olarak sınıflandırılmışsa doğru pozitif (TP) olarak değerlendirilir. Eğer nesnenin sınıfı negatifse ve pozitif

olarak sınıflandırılmışsa yanlış pozitif (FP) olarak değerlendirilir [41]. Bu bilgiler ışığında her pozitif örneklem (ilgili bir deney) için hesaplanacak skor, deneyle ilişkilendirilen ROC eğrisinin altında kalan alan, Area Under Curve (AUC), ile hesaplanmaktadır. Sorgu deneyi için AUC skorunun hesaplanması aşağıdaki algoritma ile verilmektedir:

- Getirilen deneyleri benzerlik skorlarına göre sırala
- İlgililik durumlarına göre etiketle (0 veya 1)
- TP=0 /* doğru pozitiflere ilk değer atama */
- FP=0 /* yanlış pozitiflere ilk değer atama */
- AUC=0 /* AUC skoruna ilk değer atama */
- Her bir sıralı etiket için
    - Eğer (etiket==1)
        - TP=TP+1
    - Değilse
        - FP=FP+1
        - AUC=AUC+TP
- Eğer (TP==0)
    - AUC=0
- Eğer(FP==0)
    - AUC=1
- Değilse
    - AUC=AUC/(TP*FP)

Arama platformunun genel performansı, mikrodizi deneylerinin sayısına karşılık deneylerin ulaştığı minimum AUC skoru çizilerek tarif edilmektedir. Bu raporlama yöntemi bir geri getirim sisteminin genel performansını belirlemede ayrı AUC skorlarının dikkate alındığı çalışmalarda kullanılmıştır [42; 43; 44]. Belirtilen bütün benzerlik ölçütleri için de ortalama AUC skorları hesaplanmıştır. Yüksek AUC skoru daha iyi geri getirim performansını belirtirken, AUC skorunun 1 olması mükemmel durumu göstermektedir.

İki deneyin ilgililiği iki yolla tanımlanmaktadır. İlk durumda eğer iki deney aynı hastalık ile etiketlenmiş tedavi örneklemine sahip ise iki deneyin birbiriyle ilgili olduğu söylenir. Veri kümesindeki hastalık ilişkileri ve karşılık gelen mikrodizi

deney sayıları Çizelge 3.2'de gösterilmektedir. İkinci durumda hastalıklı dokunun embriyonik kökeni göz önünde bulundurulur ve deneyler embriyonik germ hücresi tabakasına bağlı olarak endoderm, mezoderm ve ektoderm olarak etiketlenir. Embriyonik kökene göre mikrodizi deney dağılımı Çizelge 3.3'te gösterilmektedir. İlk durumdaki temel kabul spesifik bir hastalıkla etiketlenmiş bir deneyin kullanılarak veri kümesinin sorgulanmasıdır. Böylece aynı hastalık etiketine sahip diğer deneyler getirilen listenin en tepesinde, farklı hastalık etiketine sahip olanlar da listenin en altında yer alacaklardır. Sistemin belirli bir hastalık açısından ilgililiği ortaya koymak zorunda olmadığı fakat farklı hastalıklar arasında dokuya özgün bir ilişkiyi açıklamasının beklendiği iddia edilebilir. Bu konuya dikkat çekmek için ikinci durumda deneylerde kullanılan dokuların embriyonik kökeni temel alındığında dokuya özgü bir ilişkinin yakalanabileceği varsayılmıştır. miRNA'ların embriyojenezin anahtar düzenleyicileri olduğunu gösteren en son keşifler bu savı desteklemektedir [45; 46]. Bu yüzden ikinci ilgililik yolu bağlamında ikinci bir performans kriteri ayrıca değerlendirilmiştir.

3.3 Deneysel Sonuçlar

3.3.1 Geri getirim performansı

Sistemin ilgili deneyleri geri getirim kabiliyeti iki ayrı ilgililik yöntemi tarafından değerlendirilmektedir. Bunlar hastalık ilişkilendirmesi ve embriyonik germ hücresi tabakası veya basitçe embriyonik kökendir. Performans her bir deney için ayrı ayrı hesaplanan AUC skoru ile ölçülmektedir. Deney imzalarını elde etmek için ayarlanan parametre global *K* değeridir. Bu değer imzaya farklı ifade olmuş olarak eklenecek miRNA'ların yüzdesini belirtmektedir. En iyi *K* değerini seçme aşamasında, her iki ilgililik yöntemi için bütün deneylerin AUC skorlarının maksimize edilmesinde iteratif olarak geniş kapsamlı bir arama gerçekleştirilmiştir.

Çizelge 3.2 Hastalıklara göre deney dağılımları

Hastalık	Deney Sayısı
Mesane Kanseri	7
Beyin Kanseri	4
Meme Kanseri	6
Kolon Kanseri	10
ILD	32

Çizelge 3.2 devam ediyor

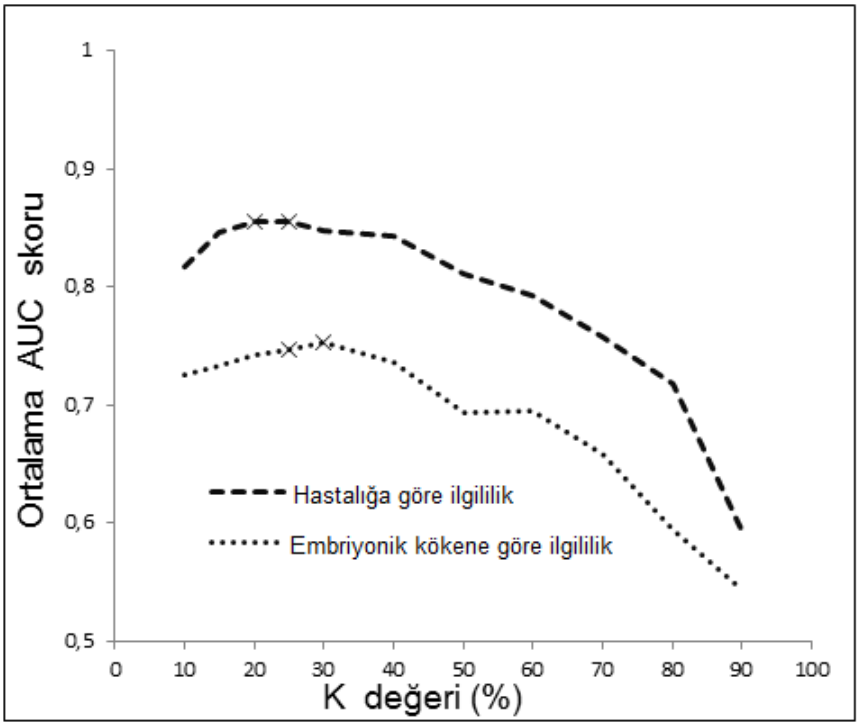
Böbrek Kanseri	5
Kan Kanseri	6
Akciğer Kanseri	25
Pankreas Kanseri	9
Prostat Kanseri	6
Schwannoma	15
Uterus Kanseri	10

Çizelge 3.3 Embriyonik kökene göre deney dağılımları

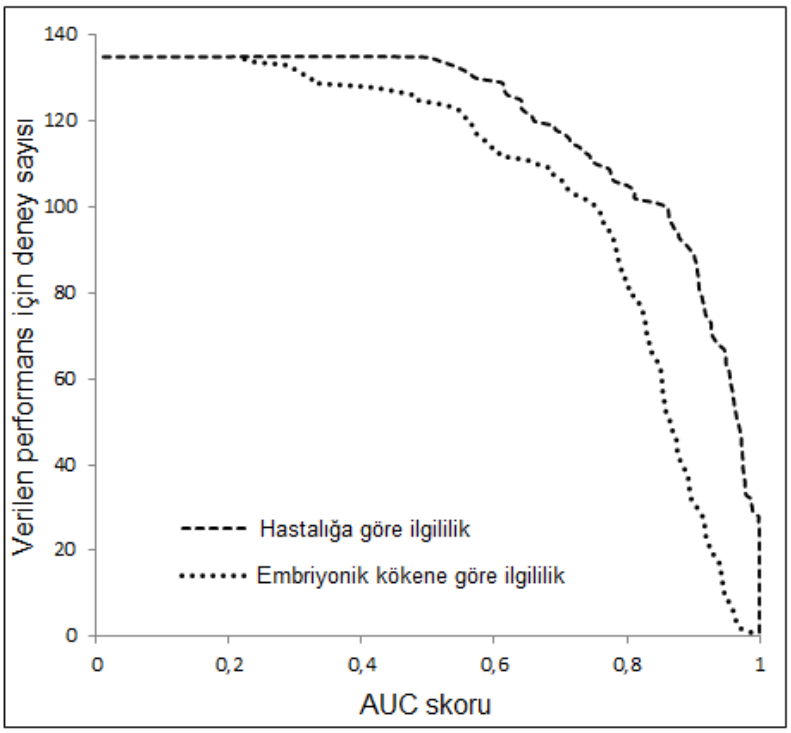
Embriyonik Köken	**Deney Sayısı**
Endoderm	53
Mezoderm	63
Ektoderm	19

Sistemin en iyi performansı *K* parametresinin 25 olduğu zaman sunduğu görülmektedir (Şekil 3.1). Deneylerin karıştırılarak dağıtılmasıyla oluşturulan alt kümeler için bu değerlendirmenin tekrar edilmesiyle optimal *K* değerinin 20 ve 30 arasında değiştiği saptanmıştır. Bu yüzden takip eden bütün deney karşılaştırmaları *K*=25 parametresi temel alınarak yapılmıştır.

Bütün deneyler için ortalama bir AUC skorunun bildirilmesi yerine sistemin genel performansının görselleştirilmesi için daha iyi bir yol, sistemin verilen AUC skorundan daha iyi sonuçlar aldığı deney sayılarının gösterilmesidir. Bu durumda yüksek bir eğri etkili bir geri getirim performansını belirtmektedir. Şekil 3.2'de görüleceği üzere sistem her ilgililik yönteminde birçok sorgu için ilgili deneyleri başarıyla getirmiştir. Fakat sistemin aynı hastalık ile ilişkilendirilen deneyleri geri getirme performansı embriyonik köken ile ilişkilendirilen deneyleri geri getirme performansından daha iyidir.



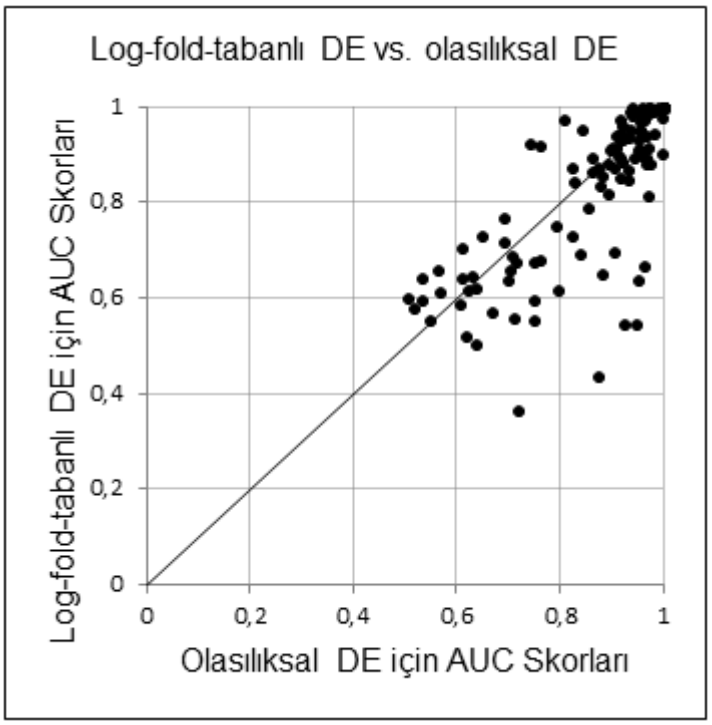
Şekil 3.1 Optimal *K* değerinin bulunması



Şekil 3.2 İki ayrı ilgililik yöntemine göre geri getirim performansları

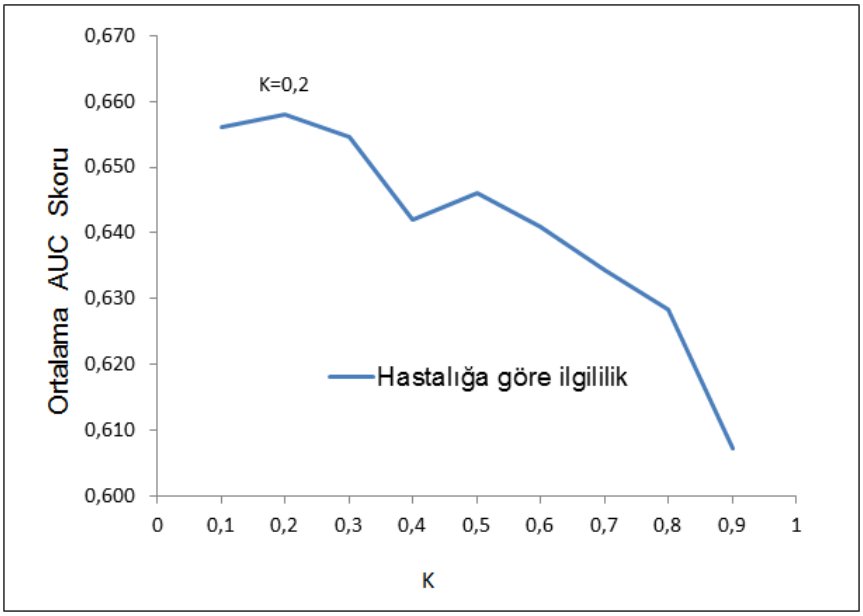
3.3.2 İmza çıkarım tekniğinin doğrulanması

İmza tasarımı, özgün tasarımı etkileyen faktörleri değiştirme yoluyla diğer birkaç alternatif ile karşılaştırılarak değerlendirilmiştir. Burada değerlendirme amaçları için sadece hastalık-ilişkili deney ilgililiği yöntemi göz önüne alınmıştır. İmza tasarımındaki ilk fark gözeten faktör tek bir deney içindeki her bir miRNA için farklı ifade olma (DE) değerini çıkarmak için kullanılan tekniktir. Bu görev için olasılıksal normal-tekdüze karışım modeli kullanılmıştır. Bu seçimin doğrulanması için iki biyolojik koşul (kontrol-tedavi) arasındaki log-fold değişimlerini temel alan DE değerleri hesaplanarak deneyler tekrar derlenmiştir [47]. Şekil 3.3'teki grafiğe göre veri kümesindeki mikrodizi deneylerinin çoğunluğu için olasılıksal DE değerleri kullanılarak elde edilen AUC skorlarının, log-fold değişimlerini temel alan model kullanılarak elde dilen AUC skorlarından daha iyi olduğu görülmektedir.



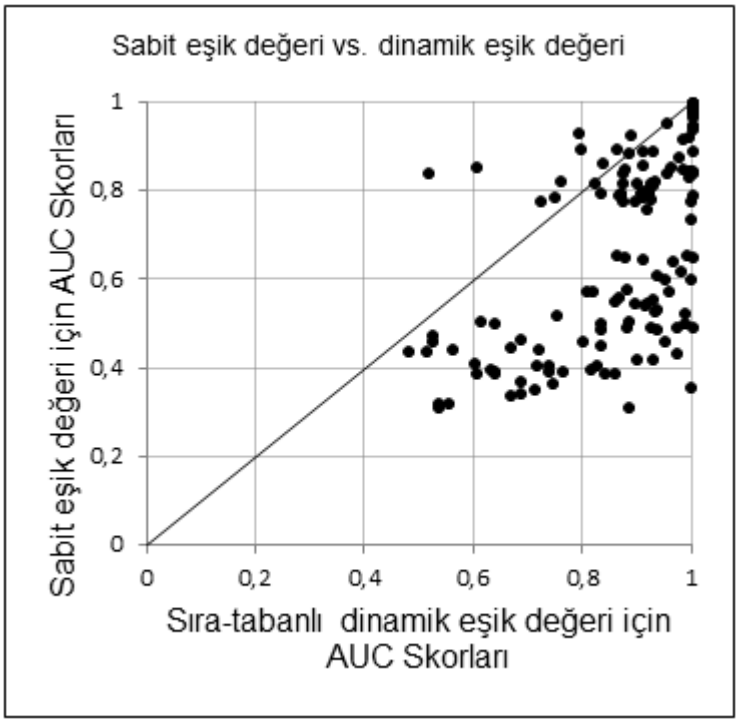
Şekil 3.3 Log-fold-tabanlı DE ve olasılıksal DE tekniğinin karşılaştırılması

İmza tasarımındaki ikinci faktör, ikili imzanın oluşturulmasında her bir deney için sıra-tabanlı bir dinamik eşik değerinin seçilmesidir. Bütün deneyler için sabit bir eşik değeri olarak kullanılmak üzere optimal bir değerin belirlenmesi için, dinamik eşik değeri ile karşılaştırılmak üzere, geniş kapsamlı bir arama yapılmıştır. Sabit eşik değeri 0.2 olarak tespit edilmiştir (Şekil 3.4). Sabit eşik değeri için sistem, sıra-tabanlı dinamik eşik değeri yöntemiyle karşılaştırıldığında 135 mikrodizi deneyinden sadece 11'i için daha iyi bir geri getirim performansı sergileyebilmiştir (Şekil 3.5). Böylece imza tasarımındaki sıra-tabanlı dinamik eşik değeri yöntemi doğrulanmaktadır.

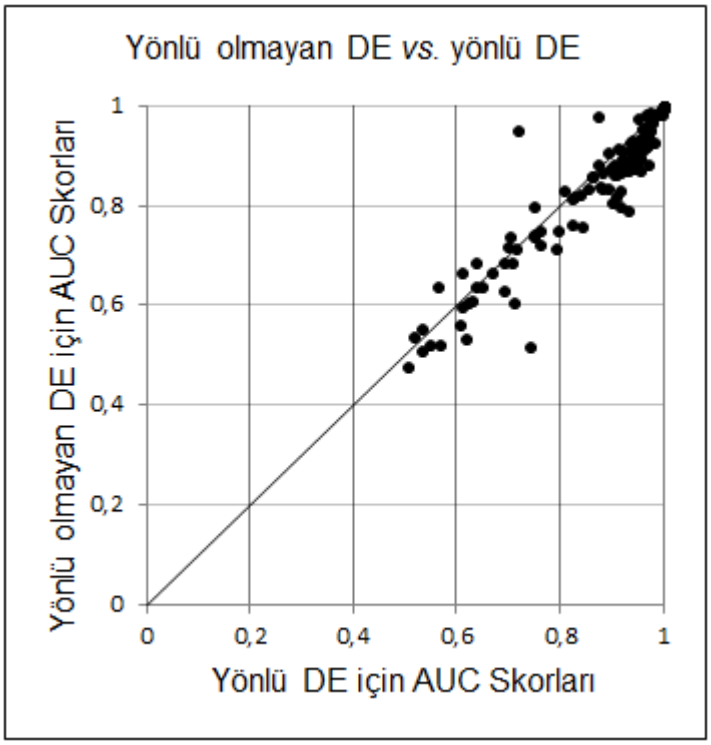


Şekil 3.4 Optimal sabit eşik değerinin bulunması

İmza tasarımındaki değerlendirilmesi gereken üçüncü faktör farklı ifade olma değerinin yönüdür (yukarı veya aşağı regülasyon). 135 deneyin 122'si için yönlü imzaların yönlü olmayan imzalardan daha iyi geri getirim performansına neden olduğu ortaya çıkarılmıştır (Şekil 3.6). Bu sonuç farklı ifade olma değerinin yönünün mikrodizi deneyleri arasında benzerliği belirlemede önemli bir etkiye sahip olduğunu açıkça ortaya koymaktadır.

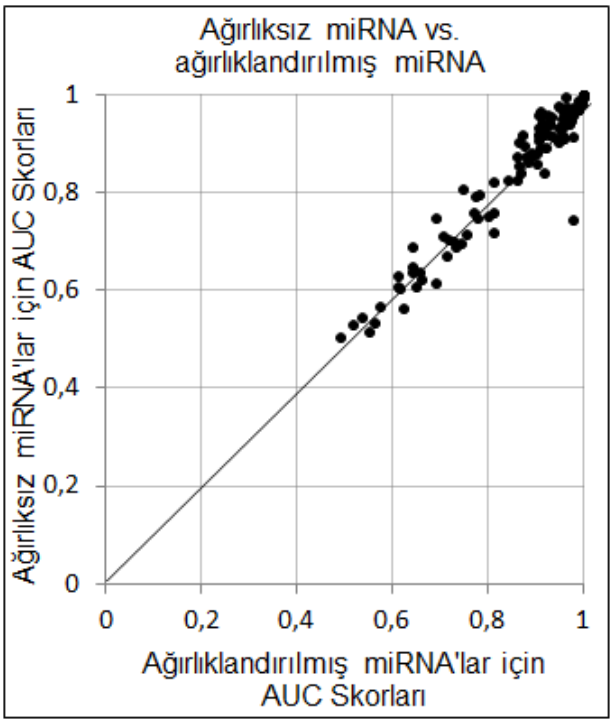


Şekil 3.5 Sabit ve dinamik eşik değeri tekniğinin karşılaştırılması



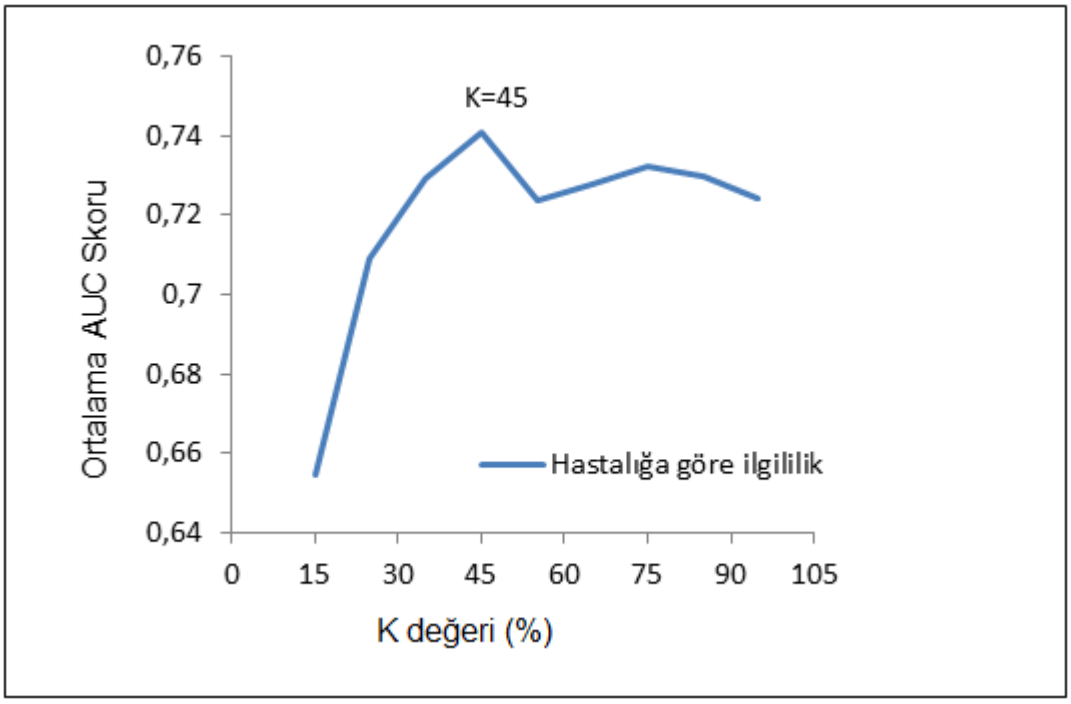
Şekil 3.6 Yönlü ve yönlü olmayan DE tekniklerinin karşılaştırılması

İmza tasarımındaki dördüncü faktör, tüm karşılaştırma veri tabanı içinde miRNA'ların bilgi içeriğine dayalı benzerlik hesaplamasında ağırlıklandırılmış miRNA'ların kullanılmasıdır. 135 deneyin 76'sında, ağırlıklandırılmamış yöntem yerine ağırlıklandırılmış yöntemin kullanılması, geri getirim performansını iyileştirebilmiştir (Şekil 3.7). Ağırlıklandırılmış miRNA'lar ile ortalama AUC skoru 0.885 iken ağırlıksız miRNA'lar ile 0.876 değeri elde edilmiştir. Bu sonuç farklı mikrodizi deneylerinin ayırt edilmesinde bazı miRNA'ların nispeten daha fazla değere sahip olduğunu ifade etmektedir.

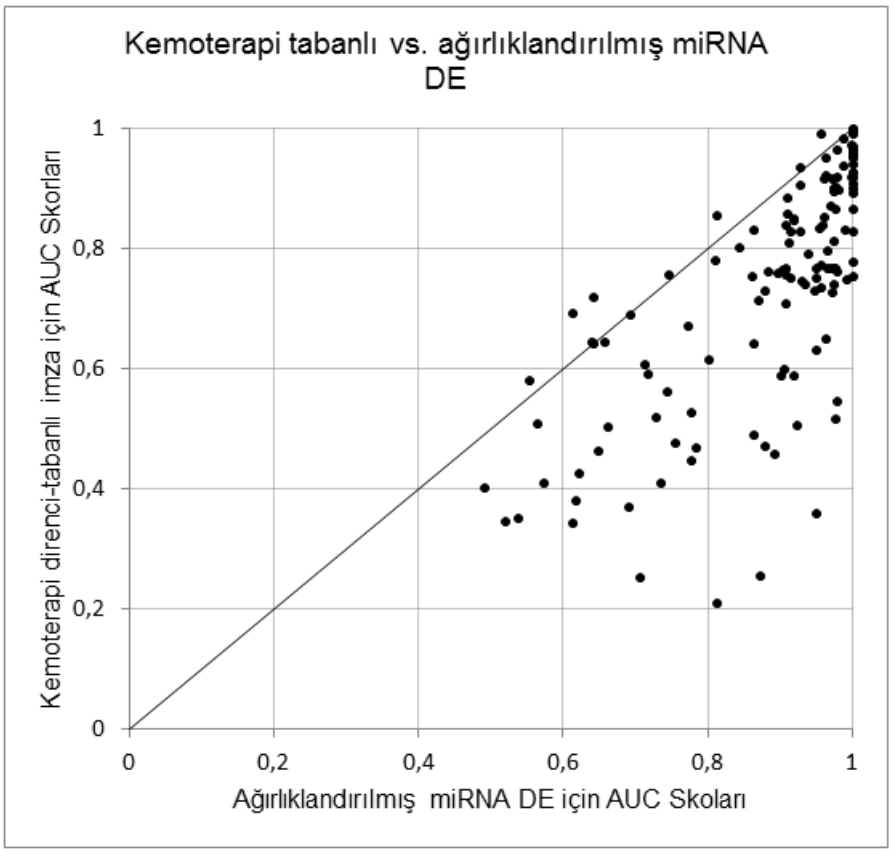


Şekil 3.7 Ağırlıklandırılmış ve ağırlıksız miRNA tekniklerinin karşılaştırılması

Kemoterapi direnci-tabanlı (miRNA kümeleri tabanlı) imza çıkarım yöntemi için ortalama AUC skorunun maksimum olduğu *K* değeri kapsamlı bir arama sonucunda 45 olarak bulunmuştur (Şekil 3.8). Bu noktada, kemoterapi direncine bağlı miRNA kümeleri tabanlı yöntemle elde edilebilen en iyi geri-getirim performansı 0.74 ortalama AUC skoru olarak gözlemlenmiştir. Ağırlıklandırılmış miRNA yöntemi ile karşılaştırıldığında bu imza çıkarım yönteminin olasılıksal DE imza çıkarım yöntemine göre daha düşük sonuçlar verdiği ortaya çıkmaktadır (Şekil 3.9).



Şekil 3.8 Kemoterapi direnci-tabanlı imza tekniği için optimal K değeri



Şekil 3.9 Kemoterapi direnci-tabanlı ve olasılıksal DE tekniklerinin karşılaştırılması

Bahsi geçen dört faktörü temel alan geri-getirim performansındaki gözlemlenmiş iyileştirmelerin istatistiksel olarak anlamlı olup olmadığının belirlenmesinde iki ayrı test yöntemi kullanılarak ikili AUC skorları arasındaki farklar için p-değeri hesaplanmıştır. Bu testler eşli t-test ve (paired t-test) ve parametrik olmayan Wilcoxon işaretli sıra testidir (Wilcoxon signed rank test). Yapılan testler sonucunda, Wilcoxon işaretli sıra testindeki log-fold-tabanlı DE ve olasılıksal DE arasındaki karşılaştırma hariç, her iki istatistiksel değerlendirme testinde bütün p-değerlerinin 0.05'in altında olduğu bulunmuştur (Çizelge 3.4). Bu sonuç imza tasarımında yapılan seçimler ile başarılan iyileştirmelerin istatistiksel olarak anlamlı olduğunun güçlü bir kanıtıdır.

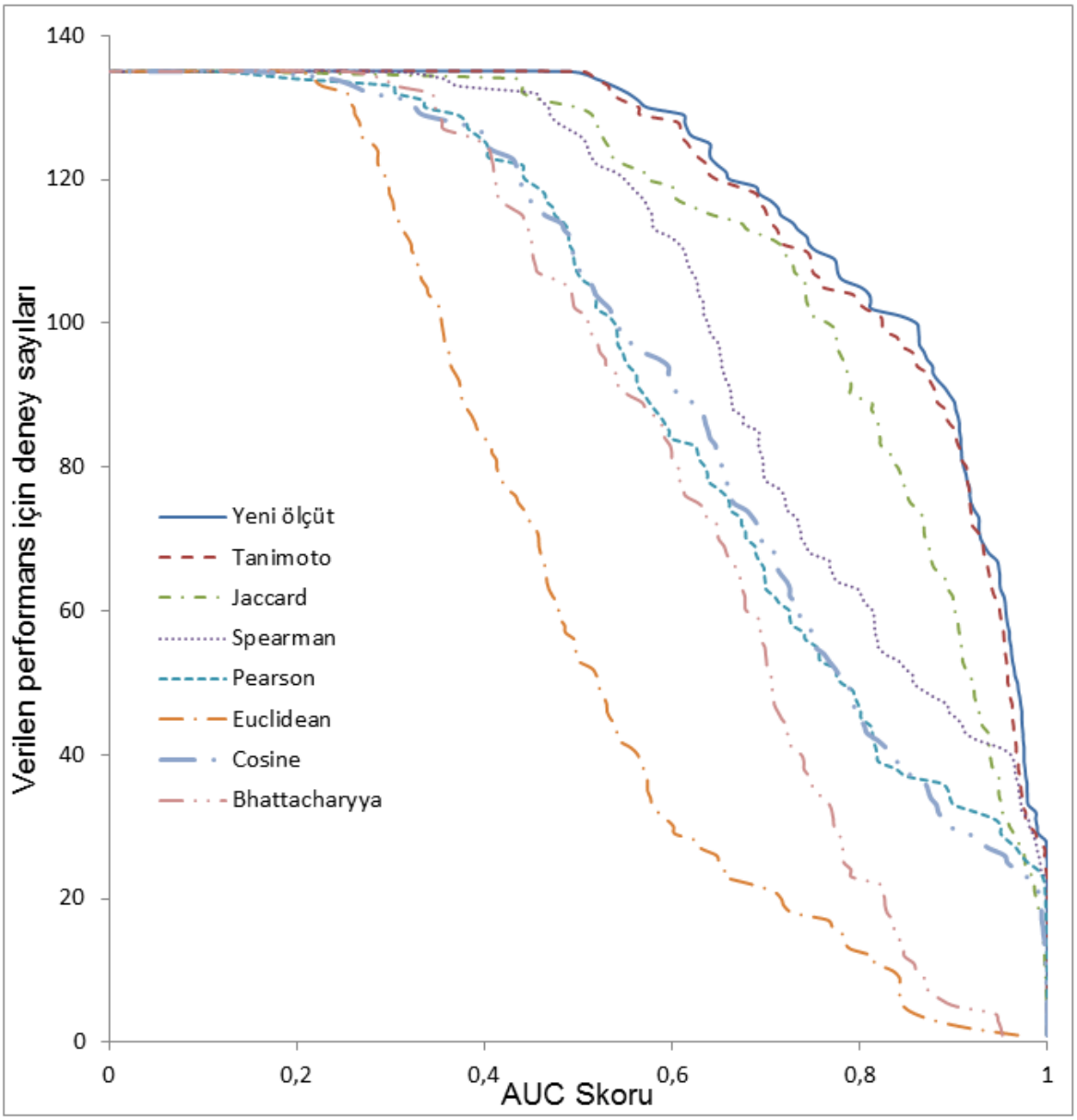
Çizelge 3.4 İmza tasarımı için istatistiksel anlamlılık testleri

Karşılaştırma	**p-değeri**	
	Eşli t-test	**Wilcoxon işaretli sıra testi**
Log-fold-tabanlı DE - olasılıksal DE	0.0053	0.077
Sabit eşik - dinamik eşik	2.88×10^{-25}	1.22×10^{-19}
Yönlü olmayan DE - yönlü DE	3.09×10^{-7}	1.35×10^{-10}
Ağırlıksız miRNA - ağırlıklı miRNA	0.019	8.17×10^{-4}

3.3.3 Benzerlik ölçütünün doğrulanması

Geri getirim sisteminde kullanılan benzerlik ölçütünün doğrulanması için farklı benzerlik ölçütlerinin performansı karşılaştırılmıştır. İkili imzaları (Tanimoto ve Jaccard ölçütleri) karşılaştıran ölçütlerin bütün sürekli ölçütlerden üstün olduğu görülmektedir (Şekil 3.10). Euclid, Pearson, Bhattacharyya ve Cosine ölçütlerinin düşük doğruluk (accuracy) değerleri; kesin DE değerinin deney tasarımı, platform tipi ve veri işlemeyi içeren birçok teknik faktör tarafından belirlenmesi gerçeğine ve bunun sonucunda da deney ilgililiği çıkarımında yanlış yönlendirmeye dayandırılabilir. Spearman ölçütü nispeten daha iyi sonuçlar üretmiştir. Bunun sebebi ise kesin DE değeri yerine DE sırasını göz önünde bulundurmasıdır. İkili imzalarda kullanılan ölçütlerden Tanimoto ölçütünün Jaccard ölçütüne üstün geldiği görülmektedir. Bu sonuç, deneylerde farklı ifade olmayan miRNA'ların da

deney benzerliğinin belirlenmesinde bir etkisi olduğunu göstermektedir. Ayrıca yeni benzerlik ölçütünde miRNA'ların ağırlıklandırılması ile daha iyi sonuçlar elde edilerek iyileştirme sağlanmıştır. Bütün benzerlik ölçütleri için AUC skorları EK-1'de verilmiştir.



Şekil 3.10 Benzerlik ölçütleri için performans karşılaştırması

Yeni benzerlik ölçütünün kullanılarak geri-getirim performansında gözlemlenmiş iyileştirmelerin istatistiksel olarak anlamlı olup olmadığını belirlemek için eşli t-test ve Wilcoxon işaretli sıra testi yardımıyla p-değeri hesaplanmıştır (Çizelge 3.5).

Bütün p-değerleri 0.05'ten küçük olarak bulunmuştur. Bu sonuç da yapılan iyileştirmelerin istatistiksel olarak anlamlı olduğunu belirtmektedir.

Çizelge 3.5 Benzerlik ölçütleri için istatistiksel anlamlılık testleri

Karşılaştırma	**p-değeri**	
	Eşli t-test	**Wilcoxon işaretli sıra testi**
Yeni ölçüt - Tanimoto	0.0019	8.17×10^{-4}
Yeni ölçüt - Jaccard	1.10×10^{-10}	2.31×10^{-12}
Yeni ölçüt - Spearman	1.25×10^{-14}	5.26×10^{-13}
Yeni ölçüt - Pearson	2.91×10^{-21}	3.51×10^{-19}
Yeni ölçüt - Euclid	1.95×10^{-56}	8.87×10^{-24}
Yeni ölçüt – Cosine	$6.08\text{x}10^{-20}$	$1.81\text{x}10^{-18}$
Yeni ölçüt - Bhattacharyya	$4.82\text{x}10^{-34}$	$4.9\text{x}10^{-22}$

4. TARTIŞMA VE ÖNERİLER

Bu tez çalışmasında deneysel veri tabanlarında bilgi geri getirimine ilişkin önemli bir problem ele alınmıştır. Büyük veri tabanlarında aranan deneylerin içeriği geri-getirim çalışmalarında asıl merak konusudur. Yapılan sorguyu temel alan ilgili girdilerin (kayıtların) geri getirimini sağlayan bir alt yapının geliştirilmesi için miRNA ifade profili mikrodizi deneyleri üzerine odaklanılmıştır. Özelleştirilmiş bir imza tasarımı ve iyileştirilmiş bir benzerlik ölçütü ile bir alt yapı önerilmiştir. Optimize edilmiş parametreler kullanılarak farklı ilgililik tanımları için yalnızca deneylerin ham içeriği temel alınarak ilgili deneylerin geri-getirimi sağlanmıştır. Alt yapının geri-getirim performansını test etmek için GEO deneysel veri kümeleri kullanılmıştır. Çalışmanın ana hedefi miRNA deneylerini de içeren büyük gen ifade profili veri havuzlarından ilgili bilgiyi geri getirmek için uyarlanabilecek pratik bir çözüm sunmaktır. GEO veri tabanının temsili bir alt kümesi üzerinde, deney benzerlik ölçütünde kullanılabilecek model parametreleri düzenlenmiştir. Bu parametreler, farklı ifade olmuş olarak kabul edilecek miRNA'ların yüzdelik kısmı ve atanacak miRNA ağırlıkları gibi faktörlerdir. Bu parametrelerin, önerilen modelin pratik uygulamalarında, daha büyük veri kümeleri üzerinde tekrar ayarlanması önerilmektedir.

Deney imzası çıkarım yöntemi olarak olasılıksal farklı ifade tabanlı yöntemin kullanılması, deney geri-getirim performansında üstünlük sağlamıştır. Bu yöntem, log-fold tabanlı farklı ifadeyi ve kemoterapi-direnci tabanlı miRNA kümelerini temel alan imza çıkarım yöntemlerine göre deney geri-getiriminde daha başarılı olmuştur. Farklı ifadelerin ikili imzaya dönüştürülmesinde dinamik eşikleme yönteminin sabit eşik yöntemine göre performansa daha olumlu etki ettiği gözlenmiştir. İkili imzaların oluşturulması aşamasında farklı ifadelerin yön bilgisinin kullanılması performans artışı sağlamıştır. Sürekli benzerlik ölçütleri arasında, deney geri-getirim performansı açısından, Spearman sıra korelasyon katsayısı diğer ölçütlere üstünlük sağlamıştır. Bu sonuç, miRNA'ların sıralarının deney geri-getirim performansını olumlu yönde etkilediğini göstermektedir. Ayrık benzerlik ölçütleri göz önünde bulundurulduğunda Tanimoto benzerlik katsayısı ölçütünün Jaccard ölçütüne üstünlük sağladığı gözlemlenmiştir. Bu sonuç da Tanimoto ölçütünde kullanılan farklı ifade olmayan miRNA'ların da geri-getirim

performansında etkili olduğunu göstermektedir. Tüm derlemdeki her bir miRNA'nın bilgi içeriğinin aynı olamayacağı düşüncesinden yola çıkılarak miRNA'lar ağırlıklandırılmış ve geri-getirim performansında iyileştirme sağlanarak önerilen yeni yöntemin üstünlüğü ortaya konmuştur.

Bilindiği kadarıyla miRNA mikrodizileri bağlamında deney geri getirim problemi ilk kez bu çalışmada araştırılmıştır. İçerik-tabanlı geri getirim yöntemleri üzerine yapılacak ilerdeki çalışmalarda kullanılması için bu çalışmada kullanılan karşılaştırma veri kümesi www.baskent.edu.tr/~hogul/mirsearch web adresinde ve EK-2'de verilmiştir. Sonuçlar genel modelin miRNA mikrodizi deneylerinden ilgili bilginin geri getirimine uygulanabileceğini gösterirken önerilen yöntemler mRNA ifade profili veya RNA-Seq deneyleri gibi diğer içeriklere de uygulanabilir durumdadır. Önerilen modelin hasta hakkında teşhis yapmak veya hastalık seyrinde öngörülerde bulunmak için, hasta-ilişkili veri ile birlikte miRNA ifade profillerinin analizinde özellikle yararlı olması beklenmektedir.

KAYNAKLAR LİSTESİ

[1] BARTEL, D.P., MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, vol.116, no.2, s.281-297, 2004.

[2] LEE, R.C., FEINBAUM, R.L. and AMBROS V., The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell, vol.75, no.5, s.843-854, 1993.

[3] VAN DER KROL, A.R., MUR, L.A., BELD, M., MOL, J.N. and STUITJE, A.R., Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression, Plant Cell, vol.2, no.4, s.291-299, 1990.

[4] NISHIDA-AOKI, N. and OCHIYA, T., Interactions between cancer cells and normal cells via miRNAs in extracellular vesicles, Cellular and Molecular Life Sciences, vol.72, no.10, 2015.

[5] OĞUL, H., UMU, S.U., TUNCEL, Y.Y. and AKAYA, M.S., A probabilistic approach to microRNA-target binding, Biochemical and Biophysical Research Communications, vol.413 no.1, s.111-115, 2011.

[6] JANSSON, M.D. and LUND, A.H., MicroRNA and cancer, Molecular Oncology, vol.6, no.6, s.590-610, 2012.

[7] O'DONNELL, K.A., WENTZEL, E.A., ZELLER, K.I., DANG, C.V. and MENDELL, J.T., c-Myc-regulated microRNAs modulate E2F1 expression, Nature, vol.435, no.7043, s.839-843, 2005.

[8] BOLON-CANEDO, V., SANCHEZ-MARONO, N., ALONSO-BETANZOS, A., BENITEZ, J.M. and HERRERA, F., A review of microarray datasets and applied feature selection methods, Information Sciences, vol.282, s.11-135, 2014.

[9] HERNANDEZ, Bort J.A., HACKL, M., HÖFLMAYER, H., JADHAV, V., HARREITHER, H., KUMAR, N., ERNST, W., GRILLARI, J. and BORTH, N., Dynamic mRNA and miRNA profiling of CHO-K1 suspension cell cultures, Biotechnology Journal, vol.7, no.4, s.500-515, 2012.

[10] D'AURIA, S., ROSSI, M., MALICKA, J., GRYCZYNSKI, Z. and GRYCZYNSKI, I., Topics in Fluorescence Spectroscopy, Kluwer Academic/Plenum Publishers: New York, NY, USA, s.213-237, 2003.

[11] WANG, B. and XI, Y., Challenges for MicroRNA Microarray Data Analysis, Microarrays, vol.2, no.2, s.34-50, 2013.

[12] PRITCHARD, C.C., CHENG, H.H. and TEWARI, M., MicroRNA profiling: approaches and considerations, Nature Reviews Genetics, vol.13, s.358-369, 2012.

[13] PARKINSON, H., KAPUSHESKY, M., SHOJATALAB, M., ABEYGUNAWARDENA, N., COULSON, R., FARNE, A., HOLLOWAY, E., KOLESNYKOV, N., LILIJA, P., LUKK, M., MANI, R., RAYNER, T., SHARMA,

A., WILLIAM, E., SARKANS, U. And BRAZMA, A., ArrayExpress—a public database of microarray experiments and gene expression profiles, Nucleic Acids Res., vol.35 (Database issue), s.747-750, 2007.

[14] BARRETT, T., WILHITE, S.E., LEDOUX, P., EVANGELISTA, C., KIM, I.F., TOMASHEVSKY, M., MARSHALL, K.A., PHILIPPY, K.H., SHERMAN, P.M., HOLKO, M., YEFANOV, A., LEE, H., ZHANG, N., ROBERTSON, C.L., SEROVA, N., DAVIS, S. And SOBOLEVA, A., NCBI GEO: archive for functional genomics data sets—update, Nucleic Acids Res., vol.41 (Database issue), s.991-995, 2013.

[15] RHEE, S.Y., BEAWIS, W., BERARDINI, T.Z., CHEN, G., DIXON, D., DOYLE, A., GARCIA-HERNANDEZ, M., HUALA, E., LANDER, G., MONTOYA, M., MILLER, N., MUELLER, L.A., MUNDODI, S., REISER, L., TACKLIND, J., WEEMS, D.C., WU, Y., XU, I., YOO, D., YOON, J. and ZHANG P., The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community, Nucleic Acids Res., vol.31, no.1, s.224-228, 2003.

[16] BETEL, D., WILSON, M., GABOW, A., DEBORA, S.M. and SANDER, C., The microRNA.org resource: targets and expression, Nucleic Acids Res., vol.36 (Database issue), s.149-153, 2008.

[17] KAYA, K.D., KARAKÜLAH, G., YAKICIER, C.M., ACAR, A.C. and KONU, Ö., mESAdb: microRNA Expression and Sequence Analysis Database, Nucleic Acids Res., vol.39 (Database issue), s.170-180, 2011.

[18] BARRETT, T. and EDGAR, R., Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO), Methods Mol Biol., vol.338, s.175-190, 2006.

[19] IVLIEV, A.E., HOEN, P.A.C.'t, VILLERIUS, M.P., DUNNEN, J.T.den and BRANDT, B.W., Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data, Nucleic Acids Res., vol.36 (Issue suppl 2), s.327-331, 2008.

[20] SEGOTA, I., BARTONICEK, N. and VLAHOVICEK, K., MADNet: microarray database network web server, Nucleic Acids Res., vol.36 (Issue suppl 2), s.332-335, 2008.

[21] HUNTER, L., TAYLOR, R.C., LEACH, S.M. and SIMON, R., GEST: a gene expression search tool based on a novel Bayesian similarity metric, Bioinformatics, vol.17, s.S115-S122, 2001.

[22] TANAY, A., STEINFELD, I., KUPIEC, M. and SHAMIR, R., Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium, Mol Syst Biol., vol.1, 2005.

[23] LAMB, J., CRAWFORD, E.D., PECK, D., MODELL, J.W., BLAT, I.C., WROBEL, M.J., LERNER, J., BRUNET, J.P., SUBRAMANIAN, A., ROSS, K.N., REICH, M., HIERONYMUS, H., WEI, G., ARMSTRONG, S.A., HAGGARTY, S.J., CLEMONS, P.A., WEI, R., CARR, S.A., LANDER, E.S. and GOLUB, T.R., The Connectivity Map: using gene-expression signatures to

connect small molecules, genes, and disease, Science, vol.313, no.5795, s.1929-1935, 2006.

[24] HASSANE, D.C., GUZMAN, M.L., CORBETT, C., LI, X., ABBOUD, R., YOUNG, F., LIESVELD, J.L., CARROLL, M. and JORDAN C.T., Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data, Blood, vol.111, no.12, s.5654-5662, 2008.

[25] SUTHRAM, S., DUDLEY, J.T., CHIANG, A.P., CHEN, R., HASTIE, T.J. and BUTTE, A.J., Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets, PLoS Comput Biol., vol.6, no.2, e1000662, 2010.

[26] HORTON, P., KISELEVA, L. and FUJIBUCHI, W., RaPiDS: an algorithm for rapid expression profile database search, Genome Informatics, vol.17, no.2, s.67-76, 2006.

[27] FUJIBUCHI, W., KISEEVA, L., TANIGUCHI, T., HARADA, H. and HORTON, P., CellMontage: similar expression profile search server, Bioinformatics, vol.23, no.22, s.3103-3104, 2007.

[28] CHEN, R., MALLELWAR, R., THOSAR, A., VENKATASUBRAHMANYAM, S. and BUTTE, A.J., GeneChaser: Identifying all biological and clinical conditions in which genes of interest are differentially expressed, BMC Bioinformatics, vol.9, s.548, 2008.

[29] HIBBS, M.A., HESS, D.C., MYERS, C.L., HUTTENHOWER, C., LI, K. and TROYANSKAYA, O.G., Bioinformatics, vol.23, no.20, s.2692-2699, 2007.

[30] ENGREITZ, J.M., MORGAN, A.A., DUDLEY, J.T., CHEN, R., THATHOO, R., ALTMAN, R.B. and BUTTE, A.J., Content-based microarray search using differential expression profiles, BMC Bioinformatics, vol.11:603, 2010.

[31] ENGREITZ, J.M., CHEN, R., MORGAN, A.A., DUDLEY, J.T., MALLELWAR, R. and BUTTE, A.J., ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression, Bioinformatics, vol.27, no.23, s.3317-3318, 2011.

[32] Bell, F., Sacan, A., Content based searching of gene expression databases using binary fingerprints of differential expression profiles, Health Informatics and Bioinformatics (HIBIT) 7th International Symposium, 19-22 Nisan, Kapadokya-Türkiye, s.107-113, 2012.

[33] CALDAS, J., GEHLENBORG, N., FAISAL, A., BRAZMA, A. and KASKI, S., Probabilistic retrieval and visualization of biologically relevant microarray experiments, Bioinformatics, vol.25, no.12, s.i145-i153, 2009.

[34] SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., MUKHERJEE, S., EBERT, B.L., GILLETTE, M.A., PAULOVICH, A., POMEROY, S.L., GOLUB, T.R., LANDER, E.S., MESIROV, J.P., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci USA., vol.102, no.43, s.15545-15550, 2005.

[35] LE, H., OLTVAI, Z.N. and BAR-JOSEPH, Z., Cross-species queries of large gene expression databases, Bioinformatics, vol.26, no.19, s.2416-2423, 2010.

[36] GEORGII, E., SALOJARVI, J., BROSCHE, M., KANGASJARVI, J., KASKI, S., Targeted retrieval of gene expression measurements using regulatory models, Bioinformatics, vol.28, no.18, s.2349-2356, 2012.

[37] CHEN, J., ZHAO. P., MASSARO, D., CLERCH, L.B., ALMON, R.R., DUBOIS, D.C., JUSKO, W.J., HOFFMAN, E.P., Nucleic Acids Res., vol.32 (Database issue), s.D578-D581, 2004.

[38] DEAN, N. and RAFTERY, A.E., Normal uniform mixture differential gene expression detection for cDNA microarrays, BMC Bioinformatics, vol.6:173, 2005.

[39] DAI, E., LV, Y., MENG, F., YU, X., ZHANG, Y., WANG, S., LIU, X., LIU, D., WANG, J., LI, X. And JIANG, W., CREAM: a database for chemotherapy resistance-associated miRSNP, Cell Death and Disease, vol.5, e1272, 2014.

[40] He, X., Cai, D., Nyogi, P., Laplacian Score for Feature Selection, Conference: Advances in Neural Information Processing Systems 18, 5-8 Aralık, Vancouver, British Columbia, Canada, s. 507-514, 2005.

[41] FAWCETT, T., An introduction to ROC analysis, Pattern Recognition Letters - Special issue: ROC analysis in pattern, vol.27, no.8, s.861-874, 2006.

[42] LIAO, L. and NOBLE, W.S., Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships, J Comput Biol.,vol.10, no.6, s.857-868, 2003.

[43] BEN-HUR, A. and BRUTLAG, D., Remote homology detection: a motif based approach, Bioinformatics, vol.19, s.i26-i33, 2003.

[44] OĞUL, H. and MUMCUOĞLU, E.Ü., A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets, Biosystems, vol.87, no.1, s.75-81, 2007.

[45] PAULI, A., RINN, J.L. and SCHIER, A.F., Non-coding RNAs as regulators of embryogenesis, Nature Reviews Genetics, vol.12, s.136-149, 2011.

[46] COLAS, A.R., MCKEITHAN, W.L., CUNNINGHAM, T.J., BUSHWAY, P.J., GARMIRE, L.X., DUESTER, G., SUBRAMANIAM, S. and MERCOLA, M., Whole-genome microRNA screening identifies let-7 and mir-18 as regulators of germ layer formation during early embryogenesis, Genes Dev., vol.26, no.23, s.2567-2579, 2012.

[47] CUI, X. and CHURCHILL, G.A., Statistical tests for differential expression in cDNA microarray experiments, Genome Biol., vol.4, no.4, 2003.

EKLER LİSTESİ

Sayfa

EK-1

BENZERLİK ÖLÇÜTLERİ İÇİN AUC SKORLARI

Sürekli benzerlik ölçütleri için ortalama AUC skorları

Benzerlik ölçütü	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
Euclid	0,49	0,39
Bhattacharrya	0,63	0,61
Pearson	0,69	0,54
Cosine	0,70	0,58
Spearman	0,77	0,57

Tanimoto (sabit eşik değeri) için ortalama AUC skorları

Eşik değeri	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
0,001	0,49	0,47
0,005	0,51	0,48
0,01	0,61	0,54
0,1	0,66	0,61
0,2	0,66	0,61
0,3	0,65	0,60
0,4	0,64	0,59
0,5	0,65	0,60
0,6	0,64	0,59
0,7	0,63	0,59
0,8	0,63	0,59
0,9	0,61	0,57

Tanimoto (dinamik eşik değeri) için ortalama AUC skorları

Yüzde (%)	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
10%	0,82	0,73
15%	0,85	0,73
20%	0,86	0,74
25%	0,86	0,75
30%	0,85	0,75
40%	0,84	0,74
50%	0,81	0,69
60%	0,79	0,69
70%	0,76	0,66
80%	0,72	0,59
90%	0,59	0,54

Tanimoto (sabit eşik değeri ve yönlü imza) için ortalama AUC skorları

Eşik değeri	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
0,001	0,78	0,69
0,005	0,72	0,66
0,01	0,74	0,66
0,1	0,66	0,61
0,2	0,66	0,60
0,3	0,65	0,59
0,4	0,64	0,59
0,5	0,65	0,60
0,6	0,64	0,59
0,7	0,64	0,59
0,8	0,63	0,59
0,9	0,61	0,58

Tanimoto (dinamik eşik değeri ve yönlü imza) için ortalama AUC skorları

Yüzde (%)	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
10%	0,83	0,74
15%	0,86	0,74
20%	0,87	0,76
25%	0,87	0,77
30%	0,86	0,78
40%	0,86	0,76
50%	0,85	0,75
60%	0,85	0,75
70%	0,85	0,75
80%	0,84	0,73
90%	0,82	0,72

Yeni yöntem (Ağırlıklandırılmış Tanimoto) için ortalama AUC skorları (dinamik eşik değeri ve yönlü imza)

Yüzde (%)	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
10%	0,85	0,75
15%	0,88	0,76
20%	0,88	0,77
25%	0,89	0,78
30%	0,88	0,79
40%	0,88	0,78
50%	0,86	0,75
60%	0,86	0,76
70%	0,86	0,75
80%	0,84	0,73
90%	0,82	0,72

Jaccard (dinamik eşik değeri) için ortalama AUC skorları

Yüzde (%)	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
10%	0,78	0,63
15%	0,81	0,64
20%	0,81	0,63
25%	0,80	0,64
30%	0,79	0,64
40%	0,79	0,64
50%	0,78	0,64
60%	0,77	0,66
70%	0,74	0,63
80%	0,70	0,58
90%	0,59	0,54

Jaccard (dinamik eşik değeri ve yönlü imza) için ortalama AUC skorları

Yüzde (%)	Ortalama AUC skoru (Hastalık)	Ortalama AUC skoru (Embriyonik köken)
10%	0,81	0,66
15%	0,83	0,66
20%	0,83	0,67
25%	0,82	0,67
30%	0,80	0,67
40%	0,80	0,67
50%	0,79	0,68
60%	0,79	0,71
70%	0,78	0,70
80%	0,76	0,68
90%	0,72	0,66

Log-fold tabanlı imza için ortalama AUC skorları (Tanimoto, dinamik eşik değeri ve yönlü imza)

Yüzde (%)	Ortalama AUC skoru (Hastalık)
10%	0,84
15%	0,84
20%	0,85
25%	0,85
30%	0,85
40%	0,85
50%	0,84
60%	0,82
70%	0,81
80%	0,81
90%	0,82

MikroRNA kümeleri tabanlı imza için ortalama AUC skorları (Tanimoto ve dinamik eşik değeri)

Yüzde (%)	Ortalama AUC skoru (Hastalık)
15%	0,65
25%	0,71
35%	0,73
45%	0,74
55%	0,72
65%	0,73
75%	0,73
85%	0,73
95%	0,72

EK-2

VERİ KÜMESİ İÇERİĞİ

No	GPL_ID	Platform Tanımı	GSE_ID	miRNA sayısı	Karşılaştırma	Hastalık	Embriyonik Köken	Kontrol Mikrodizi Sayısı	Deneysel Mikrodizi Sayısı
1	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_1	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
2	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_2	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
3	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_3	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
4	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_4	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
5	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_5	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
6	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_6	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
7	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_7	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
8	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_8	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
9	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_9	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
10	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_1_10	160	normal vs. tumor	COLON CANCER	Endoderm	5	1
11	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_1	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
12	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_2	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
13	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_3	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
14	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_4	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
15	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_5	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
16	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_6	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
17	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_7	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
18	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_8	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
19	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_2_9	160	normal vs. tumor	PANCREAS CANCER	Endoderm	1	1
20	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_3_1	160	normal vs. tumor	KIDNEY CANCER	Endoderm	3	1
21	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_3_2	160	normal vs. tumor	KIDNEY CANCER	Endoderm	3	1

22	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_3_3	160	normal vs. tumor	KIDNEY CANCER	Endoderm	3	1
23	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_3_4	160	normal vs. tumor	KIDNEY CANCER	Endoderm	3	1
24	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_3_5	160	normal vs. tumor	KIDNEY CANCER	Endoderm	3	1
25	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_4_1	160	normal vs. tumor	BLADDER_CANCER	Endoderm	2	1
26	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_4_2	160	normal vs. tumor	BLADDER_CANCER	Endoderm	2	1
27	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_4_3	160	normal vs. tumor	BLADDER_CANCER	Endoderm	2	1
28	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_4_4	160	normal vs. tumor	BLADDER_CANCER	Endoderm	2	1
29	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_4_5	160	normal vs. tumor	BLADDER_CANCER	Endoderm	2	1
30	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_4_6	160	normal vs. tumor	BLADDER_CANCER	Endoderm	2	1
31	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_4_7	160	normal vs. tumor	BLADDER_CANCER	Endoderm	2	1
32	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_5_1	160	normal vs. tumor	PROSTATE CANCER	Endoderm	8	1
33	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_5_2	160	normal vs. tumor	PROSTATE CANCER	Endoderm	8	1
34	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_5_3	160	normal vs. tumor	PROSTATE CANCER	Endoderm	8	1
35	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_5_4	160	normal vs. tumor	PROSTATE CANCER	Endoderm	8	1
36	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_5_5	160	normal vs. tumor	PROSTATE CANCER	Endoderm	8	1
37	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_5_6	160	normal vs. tumor	PROSTATE CANCER	Endoderm	8	1
38	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_1	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
39	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_2	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
40	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_3	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
41	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_4	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
42	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_5	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
43	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_6	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
44	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_7	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
45	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_8	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1

46	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_9	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
47	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_6_10	160	normal vs. tumor	UTERUS CANCER	Endoderm	9	1
48	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_7_1	160	normal vs. tumor	LUNG CANCER	Mesoderm	4	1
49	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_7_2	160	normal vs. tumor	LUNG CANCER	Mesoderm	4	1
50	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_7_3	160	normal vs. tumor	LUNG CANCER	Mesoderm	4	1
51	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_7_4	160	normal vs. tumor	LUNG CANCER	Mesoderm	4	1
52	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_7_5	160	normal vs. tumor	LUNG CANCER	Mesoderm	4	1
53	GPL1986	Bead-based microRNA profiling platform version 1	GSE2564_7_6	160	normal vs. tumor	LUNG CANCER	Mesoderm	4	1
54	GPL1986	Bead-based microRNA profiling platform version 2	GSE2564_8_1	160	normal vs. tumor	BREAST CANCER	Endoderm	3	1
55	GPL1986	Bead-based microRNA profiling platform version 3	GSE2564_8_2	160	normal vs. tumor	BREAST CANCER	Endoderm	3	1
56	GPL1986	Bead-based microRNA profiling platform version 4	GSE2564_8_3	160	normal vs. tumor	BREAST CANCER	Endoderm	3	1
57	GPL1986	Bead-based microRNA profiling platform version 5	GSE2564_8_4	160	normal vs. tumor	BREAST CANCER	Endoderm	3	1
58	GPL1986	Bead-based microRNA profiling platform version 6	GSE2564_8_5	160	normal vs. tumor	BREAST CANCER	Endoderm	3	1
59	GPL1986	Bead-based microRNA profiling platform version 7	GSE2564_8_6	160	normal vs. tumor	BREAST CANCER	Endoderm	3	1
60	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_1	558	normal vs. IPF	ILD	Mesoderm	12	1
61	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_2	558	normal vs. IPF	ILD	Mesoderm	12	1
62	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_3	558	normal vs. IPF	ILD	Mesoderm	12	1
63	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_4	558	normal vs. IPF	ILD	Mesoderm	12	1
64	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_5	558	normal vs. IPF	ILD	Mesoderm	12	1

65	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_6	558	normal vs. IPF	ILD	Mesoderm	12	1
66	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_7	558	normal vs. IPF	ILD	Mesoderm	12	1
67	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_8	558	normal vs. IPF	ILD	Mesoderm	12	1
68	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_9	558	normal vs. IPF	ILD	Mesoderm	12	1
69	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_10	558	normal vs. IPF	ILD	Mesoderm	12	1
70	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_11	558	normal vs. IPF	ILD	Mesoderm	12	1
71	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_12	558	normal vs. IPF	ILD	Mesoderm	12	1
72	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE27430_13	558	normal vs. IPF	ILD	Mesoderm	12	1
73	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE29248_1	667	normal vs. tumor	LUNG CANCER	Mesoderm	6	1
74	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE29248_2	667	normal vs. tumor	LUNG CANCER	Mesoderm	6	1
75	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE29248_3	667	normal vs. tumor	LUNG CANCER	Mesoderm	6	1
76	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE29248_4	667	normal vs. tumor	LUNG CANCER	Mesoderm	6	1
77	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE29248_5	667	normal vs. tumor	LUNG CANCER	Mesoderm	6	1
78	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE29248_6	667	normal vs. tumor	LUNG CANCER	Mesoderm	6	1
79	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_1	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1

80	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_2	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
81	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_3	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
82	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_4	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
83	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_5	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
84	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_6	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
85	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_7	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
86	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_8	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
87	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_9	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
88	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_10	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
89	GPL8179	Illumina Human v2 MicroRNA expression beadchip	GSE27606_11	415	pre- vs. post-tumor	LUNG_CANCER	Mesoderm	11	1
90	GPL16384	[miRNA-3_0] Affymetrix Multispecies miRNA-3 Array	GSE55025_1	1539	cellular vs. exosomal	LEUKAMIA	Mesoderm	6	1
91	GPL16384	[miRNA-3_0] Affymetrix Multispecies miRNA-3 Array	GSE55025_2	1539	cellular vs. exosomal	LEUKAMIA	Mesoderm	6	1
92	GPL16384	[miRNA-3_0] Affymetrix Multispecies miRNA-3 Array	GSE55025_3	1539	cellular vs. exosomal	LEUKAMIA	Mesoderm	6	1
93	GPL16384	[miRNA-3_0] Affymetrix Multispecies miRNA-3 Array	GSE55025_4	1539	cellular vs. exosomal	LEUKAMIA	Mesoderm	6	1
94	GPL16384	[miRNA-3_0] Affymetrix Multispecies miRNA-3 Array	GSE55025_5	1539	cellular vs. exosomal	LEUKAMIA	Mesoderm	6	1
95	GPL16384	[miRNA-3_0] Affymetrix Multispecies miRNA-3 Array	GSE55025_6	1539	cellular vs. exosomal	LEUKAMIA	Mesoderm	6	1
96	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_1	554	normal vs. ILD	ILD	Mesoderm	6	1

97	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_2	554	normal vs. ILD	ILD	Mesoderm	6	1
98	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_3	554	normal vs. ILD	ILD	Mesoderm	6	1
99	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_4	554	normal vs. ILD	ILD	Mesoderm	6	1
100	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_5	554	normal vs. ILD	ILD	Mesoderm	6	1
101	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_6	554	normal vs. ILD	ILD	Mesoderm	6	1
102	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_7	554	normal vs. ILD	ILD	Mesoderm	6	1
103	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_8	554	normal vs. ILD	ILD	Mesoderm	6	1
104	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_9	554	normal vs. ILD	ILD	Mesoderm	6	1
105	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_10	554	normal vs. ILD	ILD	Mesoderm	6	1
106	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_11	554	normal vs. ILD	ILD	Mesoderm	6	1
107	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_12	554	normal vs. ILD	ILD	Mesoderm	6	1
108	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_13	554	normal vs. ILD	ILD	Mesoderm	6	1
109	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_14	554	normal vs. ILD	ILD	Mesoderm	6	1

110	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_15	554	normal vs. ILD	ILD	Mesoderm	6	1
111	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_16	554	normal vs. ILD	ILD	Mesoderm	6	1
112	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_17	554	normal vs. ILD	ILD	Mesoderm	6	1
113	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_18	554	normal vs. ILD	ILD	Mesoderm	6	1
114	GPL8936	Agilent-019118 Human miRNA Microarray 2.0 G4470B (Probe Name version)	GSE21394_19	554	normal vs. ILD	ILD	Mesoderm	6	1
115	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_1	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
116	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_2	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
117	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_3	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
118	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_4	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
119	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_6	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
120	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_7	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
121	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_8	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
122	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_9	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
123	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_10	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
124	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_11	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
125	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_12	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
126	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_13	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
127	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_14	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
128	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_15	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1
129	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE43571_16	677	normal vs. tumor	SCHWANNOMA	Ectoderm	3	1

130	GPL8786	[miRNA-1_0] Affymetrix miRNA Array	GSE47056_1	677	invasive vs. non-invasive	LUNG CANCER	Mesoderm	3	3
131	GPL14613	[miRNA-2_0] Affymetrix Multispecies miRNA-2_0 Array	GSE43249_1	914	cisplatin-sensitive vs. cisplatin-resistant	LUNG CANCER	Mesoderm	3	3
132	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE49470_1	558	normal vs. tumor	BRAIN CANCER	Ectoderm	2	1
133	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE49470_2	558	normal vs. tumor	BRAIN CANCER	Ectoderm	2	1
134	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE49470_3	558	normal vs. tumor	BRAIN CANCER	Ectoderm	2	1
135	GPL8227	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version)	GSE49470_4	558	normal vs. tumor	BRAIN CANCER	Ectoderm	2	1