BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GEN İFADE VERİTABANLARINDA İÇERİK TABANLI ARAMA

AHMET HAYRAN

YÜKSEK LİSANS TEZİ
2014

GEN İFADE VERİTABANLARINDA İÇERİK TABANLI ARAMA

CONTENT BASED SEARCH IN GENE EXPRESSION DATABASES

AHMET HAYRAN

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

2014

Gen İfade Veritabanlarında İçerik Tabanlı Arama başlıklı bu çalışma, jürimiz tarafından, 14/08/2014 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI 'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan (Danışman) : Doç. Dr. Hasan OĞUL

Üye : Yrd. Doç. Dr. Emre SÜMER

Üye : Yrd. Doç. Dr. Yunus Kasım TERZİ

ONAY

.../.....

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…

Değerli arkadaşım ve doktora öğrencisi olan Esma Ergüner ÖZKOÇ'a tez aşamasında yürüttüğümüz ortak çalışmalarda verdiği destekleri için...

Bu tez çalışması TUBİTAK tarafından 113E527 nolu proje ile desteklenmiştir.

ÖZ

GEN İFADE VERİTABANLARINDA İÇERİK TABANLI ARAMA

Ahmet HAYRAN
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Büyük ölçekli gen ifade veritabanlarında zaman serisi mikrodizi deneylerinin içerik tabanlı aranması problemi ilk defa bu çalışmada araştırılmaktadır. Probleme bir bilgi geri getirim görevi olarak yaklaşılmış ve bir deneyin tamamı sorgu olarak ele alınıp önceki deneyler içerisinde aranmıştır. Metadata (üstveri) açıklamalarından daha ziyade içerik benzerliğine göre uygun deneylerin veri tabanı içerisinden bulunup getirilmesi gerekmektedir. Bu çalışmada, farklı parmak izi oluşturma yöntemleri ve uzaklık hesaplama şemalarının karşılaştırılması çeşitli zaman noktaları içerisindeki genlerin farklı ifade olma durumlarına dayalı geri getirim çatısı üzerinden sunulmuştur. Bizim oluşturduğumuz veri tabanı üzerinde yapılan tüm deneyler için, sonuçlar Pearson Bağıntı Katsayısı ve Tanimoto Uzaklığı'nın Öklid Uzaklığına göre farkı ifadeye dayalı parmak izlerinin karşılaştırılmasında yaklaşık %15 daha iyi olduğunu göstermektedir.

ANAHTAR SÖZCÜKLER: gen ifade veritabanı, mikrodizi, zaman yönlü veri, zaman serisi veri, içerik tabanlı arama, biyolojik bilgi geri getirimi
Danışman: Doç.Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

CONTENT BASED SEARCH IN GENE EXPRESSION DATABASES

Ahmet HAYRAN
Baskent University Institute of Science and Engineering
Department of Computer Engineering

The problem of content-based searching of time-series microarray experiments in large-scale gene expression databases, for the first time, is investigated in this study. The problem is examined as an information retrieval task where an entire experiment is taken as the query and searched through a collection of previous experiments. The relevant experiments are required to be retrieved based on the content similarity rather than their meta-data descriptions. A comparison of different fingerprinting and distance computation schemes is presented over a retrieval framework based on the differential expression of genes in varying time points. For all experiments carried out on database we create, results show that Pearson Correlation Coefficent and Tanimoto Distance present about 15% better performance than Euclidean Distance in comparison fingerprints based on differential expression.

KEYWORDS: gene expression database, microarray, time-course data, time-series profile, content-based search, biological information retrieval
Advisor: Assoc. Prof. Dr. Hasan OĞUL, Başkent University, Department of Computer Engineering.

İÇİNDEKİLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

GEO	*Gene Expression Omnibus*
DNA	Deoksiribonükleik asit
RNA	Ribonükleik asit
mRNA	Mesajcı RNA
tRNA	Taşıyıcı RNA
rRNA	Ribozomal RNA
cDNA	Bütünleyici DNA
RT	Ters transkriptaz (Reverse transkriptaz)
TÜBİTAK	Türkiye Bilimsel ve Teknolojik Araştırma Kurumu
NIH	*The National Institute of Health*
EBI	*European Bioinformatic Institute*
NLM	*National Library of Medicine*
FİO	Farklı İfade Olmuş
PDE	*Probability of Diffrentialy Expressed*
IF	*Intersection Fingerprint*
UF	*Union Fingerprint*

ŞEKİLLER LİSTESİ

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

1. GİRİŞ

Tüm dünyada, mikrodizi ve ilişkili deneylerden sağlanan gen ifade verilerinin toplanması ile halka açık biyolojik veri havuzları hızla büyümektedir [1,2]. Teorik, deneysel ve hesaba dayalı biyolojik bilim alanındaki birçok çalışma, yeni hipotezler bulmak, yeni deneyler tasarlamak veya yeni algoritmaların doğrulaması için karşılaştırma setleri kurmak için bu veri tabanlarından faydalanmaktadır. Bu büyük koleksiyon bilimsel araştırma ve klinik çalışmalar için hazine olarak değerlendirilirken, bu veri tabanları içerisinde arama yapmak için kullanılan mevcut araçlar anlamsal olarak güçlendirilmiş sorgular yapmak için yetersizdir. Bu araçlar sadece çok iyi yapılandırılmış üstveri sorgularına cevap vermektedirler. Bu sebeple, güncel yürütülen klinik çalışmaları bu veri havuzları içerisinde gizli olan değerli bilgiden gerçek anlamda faydalanamamaktadır. Geçen son birkaç yılda, gelişmiş biyolojik bilgi keşfinin ve biyomedikal karar destek sistemlerinin geliştirilmesine olanak sağlamak için gen ifade veri tabanlarından içerik tabanlı bilgi çıkarımı adına yapılan araştırmaların popülerliği artmıştır. Bu teknolojinin, basit üstveri veya bilgi notu tabanlı çıkarımın ötesine geçerek gelişmiş anlamsal aramaya imkan vermesi ve koleksiyondaki yararlı deneylerin bulunup getirilmesine imkan verebilecek yapısal olmayan sorguların kurulması için kullanıcılara olanak sağlaması beklenmektedir. Çünkü, gen ifade verisi üzerinden anlamsal sorgular tanımlamak basit bir iş değildir. Anlamsal aramanın makul bir yolu da tüm deneyi bir sorgu olarak almak ve içeriğe dayalı olarak en benzer olanlarını çekmek için tüm veritabanı içerisinde aramaktır. Bu yaklaşım, örneğin içerik tabanlı resim arama ve mırıldanarak müzik arama gibi, diğer alanlarda yaygın ve pratik uygulamalara sahiptir.

Hunter ve arkadaşları, 2001 yılında içerik tabanlı gen ifade veri tabanı araması için bir girişimde bulunmuşlardır. Çalışmalarında tek boyutlu iki mikrodizi deneyini karşılaştırmak için Bayes benzerlik ölçütünü tanıtmışlardır. Bu mikrodizi deneylerinin her biri deneyin gerçekleştiği tek bir durum için gen ifade profili oluşturmaktadır [3]. Yine benzer bir fikir, mevcut deney ile benzer önceki deneyleri bağdaştırarak yeni biyolojik varsayımlar üretmek için kullanılmıştır [4-9]. Spearman bağıntı katsayısı iki profili karşılaştırmak için en hızlı alternatif olarak

düşünülmektedir [10-11]. Benzer amaçla çeşitli web sunucuları bulunmaktadır. Bu sunucular tüm deney verisini almak yerine gen listesini sorgu olarak alarak farklı şekilde düzenlenmiş aynı gen listesine sahip olan deneyleri bulup getirmektedir [12-15]. Benzer hedefleri olsa da, bunlar tam bir örnek ile sorgulama (query-by-example) çıkarım sistemi değildirler. Engreitz ve arkadaşları, 2010 yılında farklı ifade profili ile gen ifade deneyini temsil etme fikrini sunmuşlardır. Veri tabanı aramasında verimliliği arttırmak için, parmak izinin oluşturulması sırasında gen altkümesi seçilerek boyut azaltma stratejisi uygulanmıştır [16]. ProfileChaser, bu yaklaşımın web uygulamasıdır. Bu uygulama çok kullanılan bir gen ifade veri tabanı olan GEO (Gene Expression Omnibus) arşivinin güncel versiyonuyla çevrimiçi arama yapabilmektedir [1] ve direk olarak ilişkili deneyleri havuzdan getirmektedir [17]. Daha hızlı arama için diğer bir öneri ise sadece genlerin aşağı ve yukarı regülasyonlarını ifade eden ikili parmak izlerinin kullanılmasıdır [18]. Farklı ifade profili yerine, Caldas et. al. deneyler arasındaki benzerlikleri bulmak için model tabanlı bir yaklaşım önermişlerdir. Onlar deneyi temsil etmek için özel genler yerine gen seti zenginleştirmelerini kullanmışlardır [19]. İçerik tabanlı mikrodizi bilgi çıkarımı için gen setlerini kullanan sadece literatürde birkaç örnek bulunmaktadır [20-21]. Bu yaklaşımın bir kısıtı ise üzerinde zenginleştirme analizinin yürütüleceği güvenilebilir gen seti koleksiyonunu bulma zorluğudur. Çünkü, araştırmacılar tarafından gen ifade veritabanlarına yüklenen verilerde deneyin kendisinden kaynaklı eksik ve hatalı veril olmasının yanı sıra araştırmacı tarafından düzgün tasarlanmamış ve formatlanmamış deneylerin içerik tabanlı bilgi geri getirimi çalışmalarında kullanılmadan önce çok fazla veri ön hazırlığına (data preparation) ihtiyaç duyması araştırmacıların işini bir hayli zorlaştırmaktadır.

Bir dizi kayda değer girişime rağmen, içerik tabanlı gen ifade araması problemi emekleme aşamasındadır. Birkaç büyük zorluk bulunmaktadır; örneğin tüm veri tabanında daha verimli arama, biyomedikal bakış açısında deneylerin daha iyi ifade edilmesi, ve sonuçların yorumlanması bu zorluklardan bazılarıdır. Bu genel motivasyonlardan ayrı olarak, çok önemli bir kısıt ise tüm mevcut metotların sadece bir veya iki (kontrol ile) ortam veya durumu hedef alması gerçeğidir. Diğer taraftan, mikrodizi veritabanları birkaç diğer tip girdi içerebilir; örneğin zaman serisi deneyleri, mevcut veritabanının önemli bir bölümünü kapsamaktadır.

Bu tezin önceki yöntemlerden farklı olarak tüm zaman serisi deneyini sorgu olarak alan ve mikrodizi deney koleksiyonunda arayan ilk girişim olduğuna inanıyoruz. Bu çalışma, içerisinde kullanılan model ve metotların zaman serisi mikrodizi verileri arasında içerik tabanlı arama için uygun olup olmayacağının düşünülmesine olanak sağlayacaktır. Yanı sıra, bu çalışmanın zaman serisi mikrodizi verileri arasında içerik tabanlı aramanın yaklaşım olarak çalışan ve etkili bir yöntem olup olmayacağının farklı açılardan araştırılmasına ve üzerinde tartışılmasına zemin hazırlayacak değerli bilgiler içereceğini düşünüyoruz. Ayrıca ileride bu alanda yapılacak diğer çalışmalara da ışık tutacağını umuyoruz. Bu amaçla, farklı ifade profillerine dayanan bir çatı kurulmuş ve deneyleri sunmak ve karşılaştırmak için birkaç alternatif şema değerlendirilmiştir. Deneysel çalışma GEO'dan alınan örnek veri kümeleri üzerinde yürütülmüştür.

Bu tez raporu dört bölümden oluşmaktadır. İlk bölümde bu alanda yapılmış önceki benzer çalışmalar, tezin motivasyonu, katkısı ve çalışma alanıyla ilgili temel bilgiler yer almaktadır. İkinci bölüm alan bilgisi içermektedir. Üçüncü bölümde kullanılan yöntemler ve veriler ile ilgili detaylı bilgi verilmektedir. Son bölümde deneysel hazırlık hakkında bilgi verilmekte, deney sonuçları anlatılmakta ve çalışma hakkında tartışma sunulmaktadır.

2. ALAN BİLGİSİ

2.1 DNA Mikrodizi

Deoksiribonükleik asit veya DNA insanlar ve hemen hemen tüm diğer canlılar için kalıtsal materyaldir. Bir insan vücudundaki her hücre yaklaşık olarak aynı DNA'ya sahiptir. Tüm organizmalar ve bazı virüslerin canlılık işlevleri ve biyolojik gelişmeleri için gerekli olan bilgileri içermesinden dolayı, DNA reçete veya şablona benzetilebilir. Bu genetik bilgileri içeren DNA parçaları gen olarak adlandırılır.

Her ne kadar insan vücudundaki tüm hücreler özdeş genetik materyal barındırsa da, aynı genler her hücrede aktif halde bulunmazlar. Farklı hücre tiplerinde hangi genlerin aktif hangi genlerin pasif olduğuyla ilgili yapılan çalışmalar, bilim adamlarına bu hücrelerin normal olarak fonksiyonlarını nasıl yerine getirdiklerini ve yine bu hücrelerin, türlü genlerin düzenli çalışmadıklarından dolayı nasıl etkilendiklerini anlamalarına yardımcı olmaktadır. Önceden, bu genetik analizler tek seferde sadece birkaç gen üzerinde yapılabilmekteydi. Fakat DNA mikrodizi (DNA microarray) teknolojisinin gelişmesiyle birlikte bilim adamları artık istenilen bir zamanda binlerce genin nasıl aktif olabildiklerini inceleyebilmektedirler.

Bir DNA mikrodizisi (DNA çip veya bioçip olarak da söylenmektedir) katı bir yüzeye tutturulmuş her biri bir geni temsil eden mikroskobik DNA beneklerinden (spot) oluşmaktadır. Her bir DNA beneği pikomol özel DNA dizisi (probe) içermektedir. Bunlar genin veya zorlu şartlar altında cDNA veya cRNA (anti-sense RNA) örneklerini melezleştirmek için kullanılan diğer DNA elementlerinin kısa bir bölümü olabilirler. DNA mikrodiziler aynı anda binlerce genin ifade seviyelerini (expression level) ölçmede veya karşılaştırmalı genomik hibridizasyon çalışmalarında kullanılmaktadır.

Mikrodizi teknolojoisi, araştırmacılara farklı birçok hastalığın patofizyolojisini daha detaylı anlamaları için yardım etmektedir. Bu hastalıklar arasına kalp hastalıkları, zihinsel rahatsızlıklar ve bulaşıcı hastalıklar da yer almaktadır. *The National Institute of Health* (NIH)'de yoğun mikrodizi araştırma alanlarından biri kanser

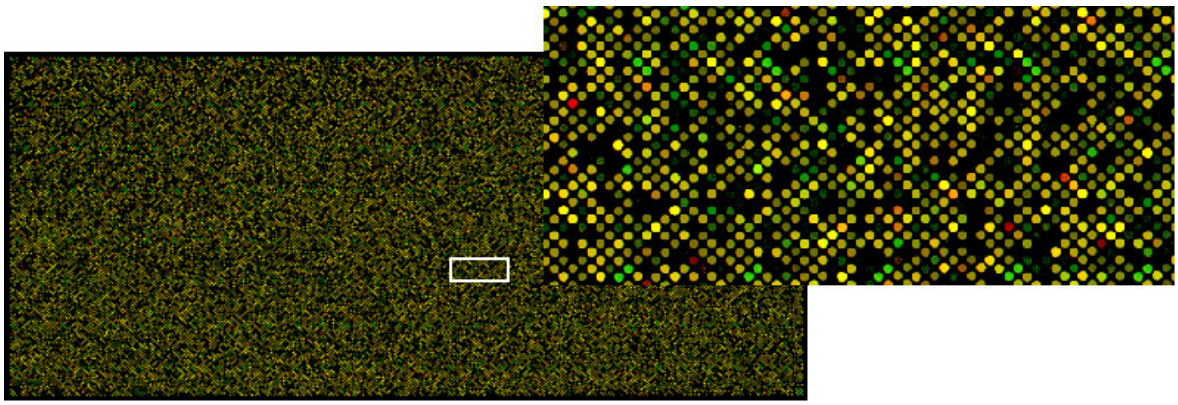
alanıdır. Geçmişte bilim adamları, kanseri üzerinde oluştuğu organa göre sınıflandırmaktaydı. Mikrodizi teknolojisinin gelişmesi ile birlikte kanseri, tümör hücrelerindeki gen aktivitelerinin örüntülerine göre daha ileri bir sınıflandırma yöntemiyle yapabileceklerdir. Böylece, araştırmacılar kanser tipine göre tedavi stratejisi tasarlayabiliyor olacaklardır. Buna ek olarak, bilim adamları tedavi uygulanmış ve uygulanmamış tümör hücreleri arasındaki gen aktivite farklılıklarını araştırarak tam olarak farklı ilaç uygulamalarının tümörleri nasıl etkilediğini anlayabilecek ve daha etkili tedavi yöntemleri geliştirebilecektir.

DNA mikrodizi teknolojisinin çalışma prensibi temelde basittir. Yukarıda bahsedildiği üzere DNA mikrodiziler, çok ufak boyutlardaki binlerce gen dizisinin tek bir mikroskobik parça üzerine robotik makinelerin düzenlemesi sayesinde oluşturulmuştur. Bu amaçla kullanılabilmesi için araştırmacılar yaklaşık 40.000'in üzerinde gen dizisine sahiptirler. Gen aktif olduğu zaman, hücresel mekanizma o genin yazılımını (transkripsiyon) gerçekleştirir. Sonuç ürün mesajcı RNA (mRNA) olarak bilinir. Bu ürün protein üretilmesi için gerekli şablon görevini görür.

Verilen hücre içerisinde hangi genin aktif ve hangi genin pasif olduğunu belirlemek için, araştırmacılar ilk olarak bu hücre içerisinde bulunan mRNA'ları toplarlar. Sonrasında toplanan bu mRNA'lar ters transkriptaz enzimler (RT) aracılığıyla etiketlenirler. Bu enzimler mRNA için tamamlayıcı cDNA üretirler. Bu etiketleme sürecinde floresan nükleotidler (fluorescent nucleodites) cDNA'ya bağlanırlar. Tümör ve normal örnekler farklı floresan boyalar ile etiketlenirler. Sonra, araştırmacılar etiketlenmiş olan cDNA'ları DNA mikrodizi üzerine yerleştirirler. Hücre içerisinden mRNA'ları temsil eden etiketlenmiş cDNA'lar mikrodizi üzerinde bulunan suni tamamlayıcı DNA'ları ile melezleşirler veya diğer bir anlamıyla bağlanırlar. Böylelikle floresan etiketlerini bırakırlar. Bu sayede araştırmacılar özel bir tarayıcı ile mikrodizi üzerindeki noktaları tarayarak floresan yoğunluğunu ölçerler.

Eğer ilgili gen çok aktif ise bu çok fazla mRNA molekülü ürettiği ve böylelikle fazla etiketlenmiş cDNA olacağı anlamına geldiğinden çok parlak floresan noktaları olacaktır. Eğer pasif ise daha az işaretlenmiş cDNA olacağından daha kısık

floresan noktaları olacaktır. Eğer floresan yok ise bu hiçbir mesajcı molekülün DNA ile melezleşmediğini anlamına gelmektedir ve genin pasif olduğunu göstermektedir. Araştırmacılar farklı zamanlarda değişik genleri test etmek amacıyla bu tekniği sıkça kullanmaktadırlar. Tümör (kırımızı boya) ve normal (yeşil boya) örnekler birlikte melezleşme safhasında mikrodizi üzerindeki suni tamamlayıcı DNA'lar için yarışırlar. Sonuç olarak, eğer nokta kırımızı ise ilgili gen tümör içerisinde normal hücreye göre daha fazla ifade olmuş (up-regulated), yeşil ise daha az ifade olmuş (down regulated), sarı ise eşit olarak ifade olmuş anlamına gelmektedir. Şekil 2.1'de yukarıda bahsetmiş olduğumuz yöntemlerle belirli bir firma tarafından üretilen platform üzerinde yapılan çalışma sonucunda ortaya çıkan mikrodizi floresan görüntüsü bulunmaktadır.



Şekil 2.1 Mikrodizi floresan görüntüsü[1]

2.2 Gen İfadesi

DNA, tüm bilinen yaşayan organizmaların ve bir çok virüs türünün gelişiminde ve yaşamsal faaliyetlerinin yerine getirebilmesi için kullanılan genetik bilginin kodlandığı bir moleküldür. DNA'nın ana rolü bilginin uzun süreli olarak saklanmasıdır. DNA bir şablona veya reçeteye benzetilebilir. Çünkü, protein ve RNA gibi hücrenin diğer bileşenlerinin oluşturulabilmesi için gerekli olan bilgiyi taşımaktadır. Genetik bilgiyi taşıyan bu DNA parçalarına gen denilmektedir. Özetle, her gen özel olarak belirli bir proteinin kodlanması için gerekli olan bilgi setini içermektedir. Proteinler de hücre fonksiyonlarını belirlerler. Bu bağlamda gen

[1] http://learn.genetics.utah.edu/content/labs/microarray

her hücrede bulunan, nesilden nesile aktarılabilen, canlı bireylerin kalıtsal özelliklerini taşıyıp ortaya çıkışını sağlayabilen kalıtım birimidir. İnsan genomunda yaklaşık 30.000 genin bulunduğu varsayılmaktadır [22].

Gen ifadesi ise genetik bilgilerin kullanılarak gen ürünlerinin sentezlenmesi sürecine denilmektedir. Gen ifadesi süreci, transkripsiyon (yazılım), RNA üretimi, taşıma, translasyon (çevirim) ve mRNA yıkımı olmak üzere bir dizi aşamaları içermektedir [23]. Sentezlenen bu gen ürünleri çoğunlukla önemli fonksiyonları yerine getiren enzim, hormon ve alıcı gibi proteinlerdir. Belirli bir hücre içerisindeki ifade olmuş binlerce gen o hücrenin ne yapabileceğini belirler. Buna ek olarak, DNA'dan RNA'ya RNA'dan da proteine olan bilgi akışı içerisindeki her aşama o hücrenin ürettiği protein miktarını ve tipini belirleyebilmesi ile kendi fonksiyonlarını, kendisinin ayarlayabilmesi için hücreye potansiyel kontrol noktası sağlar.

2.3 Mesajcı RNA (mRNA)

Mesajcı RNA veya kısaca mRNA (messenger ribosomal nucleic acid) bir RNA molekülüdür. Bu molekül hücre çekirdeğinde bulunan DNA'daki genetik bilgiyi sitoplazma içerisinde bulunan ribozoma iletir. Sentezlenecek bir proteinin amino asit dizesine karşılık gelen kimyasal şifreyi taşıyan bu molekül sayesinde ribozomda protein sentezlenir. mRNA'ya ek olarak protein sentezinde görev alan iki ana RNA türü bulunmaktadır. Bunlar ribozomal RNA (rRNA) ve taşıyıcı RNA'lar (tRNA) dır.

DNA'daki bilgi direk olarak proteine çevrilemez. İlk olarak DNA'daki bu bilgi mRNA'ya aktarılır (yazılım) ve protein sentez yeri olan ribozomlara taşınır. Sonrada burada mRNA'daki koda uygun olarak amino asit zinciri veya polipeptit sentezi süreci gerçekleşir (çevirim). Her bir mRNA molekülü bir protein bilgisi kodlar ve mRNA'nın amino asit zincirine karşılık gelen bölgelerindeki her üç baz, proteindeki bir amino aside karşılık gelir. Bu üçlülere kodon denir. Ayrıca mRNA tarafından kodlanmayan bitiş kodonu ise protein sentezini durdurur. Bu süreçte ribozom, mRNA zincirine bağlanır ve üretilecek ilgili proteinin doğru amino asit sırası için mRNA'yı şablon olarak kullanır. tRNA tarafından kodonlar tanınır ve

gerekli amino asitler toplanarak ribozoma taşınır. Ribozomun belirli bir bölümüne giren amino asitler birbirine eklenir. Daha sonra üretilen amino asit zinciri veya diğer bir anlamıyla polipeptit katlanarak (protein folding) etkin üç boyutlu protein halini alır.

2.4 GEO (Gene Expression Omnibus)

Ulusal Biyoteknoloji Bilgi Merkezi (NCBI - The National Center for Biotechnology Information), Sağlık Ulusal Enstitülerinin (National Institutes of Health) bir kolu olan Birleşmiş Devletler Ulusal Tıp Kütüphanesi'nin (NLM – United States National Library of Medicine) bir birimidir. *Gene Expression Omnibus* (GEO) veri merkezi NCBI'da bulunmaktadır. GEO, mikrodizi ve diğer bilimsel topluluklar tarafından üretilen yüksek işlem hacimli veri formatlarını destekleyen halka açık fonksiyonel genomik veri deposudur. Ağırlıklı olarak gen ifade verileri DNA mikrodizi teknolojisi ile üretilmiştir. GEO, şu anda tamamıyla halka açık en büyük gen ifade veri kaynağıdır [24]. Bu veri deposu, dizi (array), sıralama tabanlı (sequence-based) ve MIAME (Minimum Information About a Microarray Experiment) uyumlu formatta veri yüklemeyi desteklemektedir. GEO diğer kategorilerde de yüksek işlem hacimli fonksiyonel genomik verilere ev sahipliği yapmaktadır. Bunlar genom kopya sayısı değişimi, kromatin yapısı, metilasyon durumu ve transkripsiyon (yazılım) faktör bağlanması çalışmalarını içermektedir. Bu veriler, mikrodiziler gibi yüksek işlem hacimli teknolojiler tarafından üretilmektedir.

GEO'da kaydedilen veriler basit ve standart bir düzenlemeye sahiptir. Tüm gönderilen veriler özgün GEO örnekleri (GEO samples, GSM) olarak tutulur. Aynı deneyden kaydedilen örnekler GEO serileri (GEO data series, GSE), birbirine benzer ve aynı aktiviteleri içeren yüksek kaliteli (curated) bazı seriler GEO veri kümeleri olarak tutulur (GEO DataSet, GDS). GDS, çalışma ile ilgili detaylı bilgi içerir.

Haziran 2010'da GEO'da 495,422 örnek veri, dünya genelinde 5,000 den fazla laboratuvar ve 500 organizma için yapılmış yaklaşık 16 milyar özgün ölçüm bulunmaktadır [25-27]. Haziran 2014 itibariyle, yaklaşık 1,165,000 örnek veriye ev

sahipliği yapmaktadır. Ayrıca veri merkezinin web sitesi olan http://www.ncbi.nlm.nih.gov/geo/ adresinden sunulan bazı araçlara ulaşılabilmektedir. Bu araçlar kullanıcılara sorgulama, deneyleri indirme ve gen ifade profillerini analiz etmede yardımcı olmaktadır.

GEO'dan farklı olarak Avrupa Biyoenformatik Enstitüsü (European Bioinformatic Institude - EBI) tarafından kurulan *ArrayExpress* ve Stanford *Microarray* veritabanları da diğer halka açık veritabanlarıdır [28]. Bu çalışmada analiz edilen gen ifade verileri GEO veri merkezinden sağlanmıştır.

2.5 İçerik Tabanlı Arama

İçerik tabanlı (content-based) kelimesi arama işleminin öz bilgi dediğimiz örneğin anahtar kelime, etiketler ve açıklamalar yerine, arama yapılan şeyin içeriğinin analiz edilerek yapılması mantığına dayanmaktadır. Örnek olarak içerik tabanlı resim arama işleminde resim ile ilişkilendirilmiş anahtar kelimeler, etiketler veya açıklamalar yerine resmin içeriğinin analiz edilmesi verilebilir. Bu bağlamda içerik (content) renkler, şekiller, dolgular veya resmin kendisinden edinilen diğer bilgiler olabilir. Bu çalışmada ise gen ifade profilleri içerisinde arama yapıldığından içerik gen ifade profilleri içerisindeki ifade seviyelerini gösteren değerler veya bu değerlerden belirli yöntemler ile dönüştürülmüş veya özetlenmiş çıktılardır.

Çalışmamızın 2.4 GEO bölümünde bahsettiğimiz gibi, GEO sık kullanılan güncel halka açık genomik veri deposudur. GEO gibi *ArrayExpress* ve Standford *Microarray* diğer halka açık veritabanlarıdır. Mikrodizi deneyi sonunda araştırmacılar kendi sonuçlarını analiz ettikten sonra genellikle bu gibi halka açık veritabanlarındaki diğer çalışmalarla karşılaştırmak istemektedirler. Tek tek veritabanında arama yapmak oldukça zahmetli ve zaman alıcıdır. Mikrodizi veritabanlarında arama için ortak yaklaşım metinsel açıklama*,* dizi ve genlerin tanımlamaları gibi öz bilgi (metadata) tabanlıdır. GEO, bu parametreler ile aramaya olanak sağlamaktadır. Ancak bu tür aramalarda tamamlanmamış sınıflandırma, eksik açıklama ve örneklerin etiketlenmemesi gibi nedenlerden dolayı tüm ilgili deneyler bulunamamaktadır. GEO'nun arama kapasitesini

arttırmak için *GEOmetadb* geliştirilmiştir. GEOmetadb, GEO'da arama seçeneği sunan bir SQL veritabanı versiyonudur. GEOmetadb, konu bulmak için verimlidir fakat benzer sonuçlar vermiş çalışmaları bulamaz. Benzer sonuçlar vermiş çalışmaları bulmaya "İçerik tabanlı arama (*Content based search*)" denir [29].

Benzer içerikli deneyleri bulmak için kullanılan yöntemlerden biri farklı ifade olmuş (FİO, Differentially Expressed) genlerin belirlenmesidir. FİO geni tanımlamanın en basit yolu farklı örneklerdeki veya farklı koşullar altındaki aynı örneğin ifade oranına bakmaktır. Eğer oran belirlenen değişiklik katsayısını (fold change) geçerse, gene veya proteine FİO gen denir. Bu katsayı genellikle 2 olarak kullanılır. Eğer bu değer yüksek olursa mesela 10, bulunan sonuçların kesinlikle FİO gen olduğundan emin olunabilir fakat bu durumda birçok FİO gen de gözden kaçabilir. Eğer düşük bir değer atanırsa FİO gen olmayanlar da FİO olarak tanımlanabilir. Bu nedenle sadece değişiklik katsayı oranına göre değil aynı zamanda istatistiksel hesaplamalar yapılarak gerçek FİO genler bulunabilir [30].

Mikrodizi deneylerinin analizinde, koşullar arasında farklı ifade olmuş genlerin tespit edilebilmesi için birçok yöntem mevcuttur [31]. Bu genlerin tanımlanabilmesi için uygun yöntemin seçimi tanımlanacak gen setini fazlasıyla etkileyebilir [32]. Birçok mevcut yöntem olmasına rağmen, biyologlar önceki yaklaşımlardan olan değişim katsayısı (fold-change) ve t-istatistik (t-statistic) metotlarına yakınlık göstermektedirler [33].

FİO genler bulunduktan sonra sorgulanan deney ile içerik olarak benzer deneyler listelenir. Bunun için literatürde farklı yaklaşımlar ve yöntemler mevcuttur. Ancak GEO'nun henüz içerik tabanlı arama yapan bir fonksiyonu bulunmamaktadır.

2.6 Zaman Serisi Deneyler

DNA mikrodizi deneyleri genel olarak deneylerde kullanılan dizilerin tiplerine (cDNA ve oligonucletid diziler) veya profili çıkarılan organizmaya göre sınıflandırılmaktadır. Statik ifade deneylerinde farklı örneklerde ki gen ifadelerinin anlık görüntü karesi alınırken zaman serisi deneylerde zaman serisi süreç ölçülür. Bu iki veri türü arasında diğer önemli fark ise örnek popülasyondan (örnek;

yumurtalık kanseri hastaları) alınan statik veriler bağımsız olarak aynen dağılmışken zaman serisi veriler başarılı noktalar (points) arasında güçlü öz ilinti sergilemesidir [34].

Mikrodizi deneyleri statik ve zaman serisi olmak üzere iki ana gruba ayrılmaktadır. Statik deneylerde gen ifade ölçümleri her bir örnek için bir defa alınır. Örneğin, statik mikrodizi deneyinin genel türünde, belirli bir hastalığın işleyişini araştıran araştırmacılar kişilerinden alınan normal ve hasta dokuların gen ifade seviyelerini ölçer ve karşılaştırırlar [35].

Diğer taraftan zaman serisi (time series veya temporal) deneylerde tek bir örneğin zaman içerisinde bir kaç noktada ifade seviyeleri ölçülür. Zaman serisi mikrodizi deney çalışmalarının çoğu dört yaygın kategorinin birinde yer almaktadır. Bunlardan birincisi çeşitli biyolojik sistemlerin arkasında yatan dinamikleri keşfetmektir. Bu sistemlere örnek olarak hücre döngüsü veya günlük saat (circadian clock) verilebilir [34, 36]. İkinci kategori ise gelişimdir. Araştırmacılar belirli gelişim sürecinde zaman serisi verileri toplayarak ve analiz ederek bu sureci kontrol eden genler hakkında bilgi edinebilirler. Bu yolla çalışılan ilginç bir örnek ise sinir sistemi gelişimi ve kök hücre farklılaşımıdır [35, 36]. Üçüncüsü, zaman serisi mikrodizi deneyler gözlemlenen semptomlar arkasında yatan genetik değişiklikleri ortaya çıkararak hastalığın ilerlemesine ilişkin ışık tutabilir [34, 36]. Araştırmacılar mikrodizi teknolojisini örneğin Alzheimer [37], HIV [38] ve kanser [39] gibi hastalıklar üzerinde araştırma yapmak için uygulamaktadırlar. Dördüncüsü ve sonuncusu, araştırmacılar merak uyandıran çeşitli durumlarda gerçekleşen genetik tepkiyi belirlemek için zaman serisi deneyleri kullanabilmektedir. Bu durumlara örnek büyük darbe, stres durumları ve ilaç uygulamasıdır [40, 35].

Zaman serisi ifade verileri büyük miktarda biyolojik bilgi üretmek için açıkça iyi potansiyele sahiptir. Ham zaman serisi mikrodiziler ile başlamayan ve yukarıda bahsedilen faydalı sonuç türlerine ulaşmaya çalışan araştırmacalar için veri analizi en zorlu aşamayı oluşturmaktadır.

Yukarıdaki paragraflarda bahsedildiği gibi zaman serisi mikrodiziler devam eden hücresel sürecin farklı zaman noktalarındaki (örneğin; dakika, saat ve gün) çoklu ifade profillerini tutmaktadır. Bu veriler, zaman fonksiyonu olarak farklı gen ifadeleri şeklinde karmaşık dinamikleri ve regülasyonları nitelendirebilmektedirler. Farklı disiplinlerden kaynaklanan birçok zaman serisi analiz metodu (örneğin; sinyal işleme, dinamik sistem teorisi, bilgisayar öğrenme ve bilgi teorisi) farklı ifade olmuş genlerin belirlenmesi, ifade örüntülerinin tanımlanması ve gen ağının kurulması için kullanılmaktadır [41 – 44]. Fakat yine de zorluklar sürmektedir.

Zaman serisi veriler ile uğraşmanın en önemli zorluğu kısıtlı örnek ve alınan zaman noktası sayısıdır ki, kısa zaman serisi verilere sebep olmaktadır. Zaman serisi mikrodizi veri setlerinin genişleyen havuzu içinde genellikle zaman serisi kayıtları 10 zaman noktasından (time-points) daha azdır [45]. Zaman serisi verilerin en yaygın türü kısa zaman serisi verilerdir. Bunun sebebi birçok zaman noktası için örneklerin elde edilmesinde yaşanan zorluklardan doğmaktadır ki, özellikle hayvan veya klinik çalışmalarında ki çoğu kez dizilerin yüksek maliyeti veya kısıtlı biyolojik örneklerden dolayıdır [46, 47]. “Kısa” zaman serisi, zaman ölçeği (time-scale) veya birkaç farklı zaman noktası anlamına gelebilir. Genellikle, ikincisini ima etmektedir ki, bunun için en uygunu seyrek zaman serisi verisi (sparse time-series) olmalıdır.

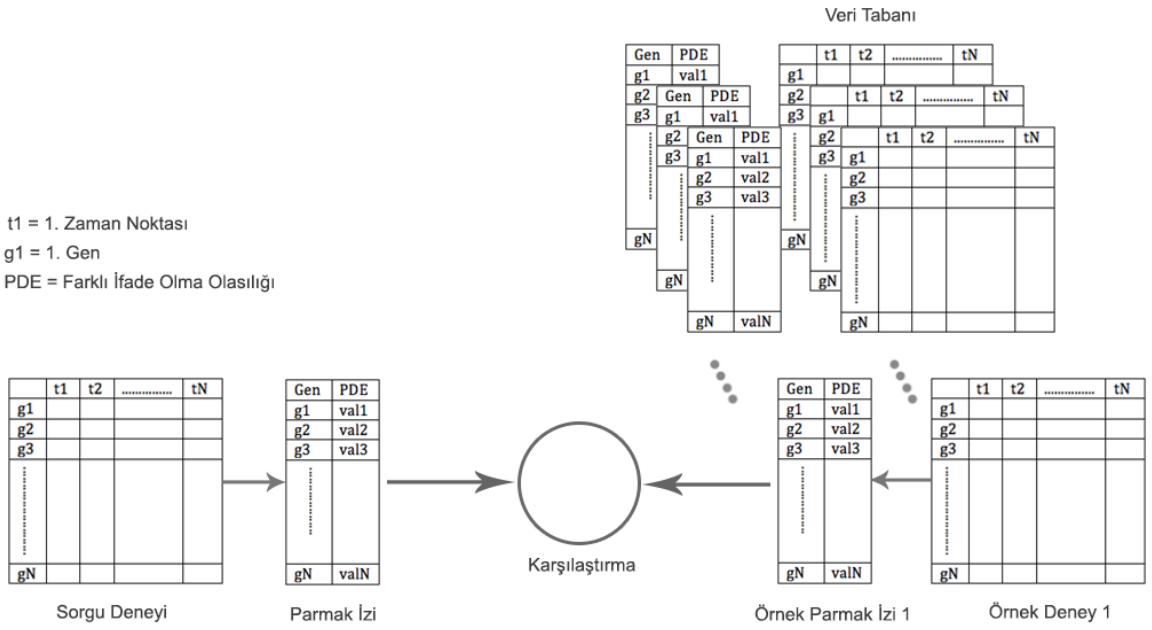
Kısıtlı örnekleme, statik veya standart zaman serisi analizlerindeki zorluğu şiddetlendirmektedir. İlk olarak, statik ve uzun zaman serisi mikrodizi verilerinin analiz edilmesinde ki problemler yüksek boyutluktan kaynaklanmaktadır ki, buna tekil matris (matrix singularity) ve model uyumluluğu gibi az örnek sayısı eşlik etmektedir. Bu durum kısa zaman serisi verilerinde daha fazla göze çarpmaktadır [48]. İkincisi, kaçınılmaz gürültüler uzun zaman serilerine göre kısa zaman serisi verilerin analizi üzerinde daha fazla etkisi vardır ve bu rastgele örüntülerden gerçeklerin ayırt edilmesinde ki zorluğu artırmakta ve yanıltıcı analizlerin potansiyelini arttırmaktadır [49].

3. YÖNTEMLER

3.1 Bilgi Çıkarım Modeli

İlişkili mikrodizi deneylerinin içerik-tabanlı geri getirimi için kurulacak bir alt yapı, her bir deney için deney içeriğini ifade eden parmak izlerinin (fingerprints) oluşturulmasını gerektirmektedir. Buradaki, parmak izi ifadesi ile geleneksel bilgi geri getirimi için analojik olarak içerik (index) terimi kastedilmektedir. Veri tabanı arama sırasında, sorgu deneyinin karşılaştırılması onun parmak izi ve veri tabanında önceden bulunan diğer parmak izleri ile yapılmaktadır. Bundan dolayı, başarılı uygulanmış bir içerik tabanlı arama stratejisi ilk olarak belirlenmiş deneylerden nasıl temsilci parmak izi elde edilebileceğinin ve ikinci olarak etkili ve verimli bir yolla nasıl iki parmak izinin karşılaştırılacağının üstesinden gelmelidir.

Verilen E gen ifade matrisinde e_{ij}, j^{th} koşulunda i^{th} genin ifadesini simgelemektedir. Bu ifadede ilişkili deneylerin geri getirilmesi problemi mikrodizi deposu içerisinde $\{M_1, M_2, \dots, M_t\}$ matrisleri arasında M_k matrisinin bulunması olarak tanımlanmaktadır. Burada k en kısa mesafe $d(E, M_k)$'yı elde etmektedir. Başka bir ifadeyle, geri getirim işlemi sorgu matrisi ve veri tabanları içerisinde bulunan diğer matrisler arasında mesafelerin karşılaştırıması ve en az skoru elde edenin raporlanması görevini içermektedir. Karşılaştırılan deney matrisleri arasındaki tüm satırların (gen listelerinin) eşleştirilmesi mantıklı bir varsayım olduğundan bilgi geri getirim modeli iki matrisin baştan başa tüm gen ifade profillerinin eşleştirilmesi üzerine kurulabilir. O zaman, $d(E, M_k)$ uzaklığı E nin parmak izlerinin ve M_k'nın tüm matrislerinin yerine onların parmak izleri üzerine tanımlanabilir. Şekil 3.1'de zaman serisi verilerde içerik tabanlı aramanın genel çalışma modeli gösterilmektedir.



Şekil 3.1 Zaman serisi verilerde içerik tabanlı arama

3.2 Parmak İzi Çıkarma

Parmak izi ile mikrodizi deneyinin ifade edebilmenin genel ve başarılı yöntemi parmak izini deneydeki ölçülen tüm genlerin farklı ifadelerinin (differential expressions) vektörü olarak tasarlamaktır. Farklı ifade profili parmak izi, deneydeki tüm genlerin farklı ifadelerinin tek bir vektöre birleştirilmesi ile kolayca elde edilebilmektedir. Bu çalışmada, farklı ifade (differential expression), genin iki deneysel koşul arasında farklı ifade olma olasılığı değeri ile tanımlanmaktadır. Bu farklı ifade olmuş gen i nin olasılığını gizli değer z_i ile göstermekteyiz. Ayrıca, zaman serisi ifade verisi için farklı ifadenin iki tanımı tartışılmaktadır. Birinci alternatif, ilk (genellikle tedavisiz kontrol – untreated control) ve son zaman noktaları arasında farklı ifade olasılığının hesaplanmasıdır ki bu LAST_DE olarak çalışmada isimlendirilmiştir. İkinci alternatif ise zaman serisi mikrodizi deneyi içerisindeki ilk zaman noktası ile aynı mikrodiziye ait diğer tüm zaman noktaları arasındaki farklı ifade olma olasılığını hesaplamakta ve en yüksek değeri farklı ifade olmuş olarak belirlemektedir. Bu parmak izi şeması MAX_DE olarak isimlendirilmiştir. İki zaman noktası arasında farklı ifade olmuş genlerin olasılıklarını hesaplamak için Dean ve Raftery tarafından tanıtılan bir metot benimsenmiştir [50]. Bu metot z_i yi verinin düz ve farklı ifade olmuş genlerin

normal tekdüze bileşiminin içine sırasıyla sığdırarak hesaplamaktadır. Bu model Eşitlik 3.1'de verilmiştir.

3.3 Farklı İfade Olmuş Genlerin Çıkartılması

Normal Tekdüze Diferansiyel Gen İfadesi (Normal Uniform Differential Gene Expression, NUDGE) cDNA mikrodizilerde farklı ifade olmuş genlerin belirlenmesinde kullanılan bir metottur. Bu metot basit tek değişkenli normal-tekdüze karışık model üzerine kurulmuştur. Birden fazla testi hesaba katarak çıktısının bir bölümü olarak farklı ifade olasılığı verir. Tek parçalı veya tekrarlı deneyler üzerinde uygulanabilir ve hızlıdır [50].

Genler farklı ifade olmuş ve farklı ifade olmamış olarak iki farklı grup olarak modellenir. Her bir grup kendi yoğunluğuna göre modellenir ve bu nedenle tüm veri bu yoğunlukların ağırlıklandırılmış karışımına göre modellenmiş olur. Burada ağırlık iki grubun her birinde öncelikli olma olasılığına karşılık gelmektedir. Bu iki parçalı karışık model olarak neticelenir. İlk grup farklı ifade olmayan genler grubudur ve logaritma oranları sıfırdır. Gözlenen log oranları bu genler için uygun dönüşümlerden sonra Gaussian yoğunluğu ile modellenir. İkinci grup ise farklı ifade olmuş genlerdir, log oranları diğer gruptan oldukça uzaktır ve uygun aralıkta uniform dağılımlı olarak modellenir. Bu model şöyle tanımlanmaktadır (3.1):

$$x_i \sim \pi N(x_i|\mu, \sigma^2) + (1-\pi)\, \mathrm{U}_{[a,b]}\, x_i \;, i = 1, \ldots, N \qquad (3.1)$$

Model içerisindeki belirtilen parametreler şöyledir;

1. x_i → i geni için gözlenen log değeri
2. π → öncelik olasılığı
3. $\mathrm{N}(x_i|\, \mu, \sigma 2)$ → μ mean ve σ^2 varyanslı Gaussian dağılımı
4. $\mathrm{U}_{[a,b]}\, x$ → [a,b] aralığı için uniform dağılım
5. N → gen sayısı

Model, EM (Expectation-Maximization) algoritması kullanarak maksimum likelihood ile tahmin edilir. Öncelikle tanımlanmamış etiketler z_i , i $= 1, \dots,$ N belirlenir. Eğer gen farkı ifade olmuş gen ise 1 değilse 0 olarak etiketlenir.

Algoritma içerisinde iki aşama vardır. Bunlar;

1. E Adımı (Expectation) : Mevcut verilen parametreye göre etiketler tahmin edilir. k yineleme için algoritma şöyledir;

$$\hat{z}_i^{(k)} = \frac{(1-\hat{\pi}^{(k-1)}) \ \cup_{[\hat{a},\widehat{b}]}(x_i)}{\hat{\pi}^{(k-1)} N(x_i|\hat{\mu}^{(k-1)}, (\hat{\sigma}^{(k-1)})^2) + (1-\hat{\pi}^{(k-1)}) \ \cup_{[\hat{a},\widehat{b}]}(x_i)} \ , i = 1, \dots, N.$$

2. M Adımı (Maximization) : Verilen etiket tahminleri ile π, μ, σ^2 parametreleri tahmin edilir. Yine k yineleme için algoritma şöyledir;

   - $\hat{\pi}^{(k)} = \frac{\sum_{i=1}^{N}(1-\hat{z}_i^{(k)})}{N}$
   - $\hat{\mu}^{(k)} = \frac{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right).x_i}{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)}$
   - $(\hat{\sigma}^{(k)})^2 = \frac{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right).(x_i-\hat{\mu}^{(k)})^2}{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)}$

   Verilen parametre tahminleri ile k. yinelemede olasılık (likelihood) şöyledir;

$$L\left(X: \hat{\pi}^{(k)}, \hat{\mu}^{(k)}, (\hat{\sigma}^{(k)})^2\right) = \prod_{i=1}^{N} \left\{ \hat{\pi}^{(k)} N\left(x_i \, ; \, \hat{\mu}^{(k)}, (\hat{\sigma}^{(k)})^2\right) + \left(1-\hat{\pi}_i^{(k)}\right) \bigcup_{[\hat{a},\widehat{b}]} (x_i) \right\}.$$

$\hat{a} = \min\{x_i : i = 1, \dots, N\}$ ve $\hat{b}$= $\max\{x_i : i = 1, \dots, N\}$ değerleri algoritma süresince değişmez.

Yukarıdaki adımlar yakınsayana (convergence) kadar tekrar eder. Yakınsama, parametre tahminlerini, etiketleri ve olasılık logaritmalarını her adımda kontrol ederek bulabilir. Miktarlardaki değişim adımlar arasında yeterince küçüldüğü zaman algoritma yakınsadı denilebilir. z_i'nin başlangıç değeri eğer i. geninin gözlenen log değerinden tüm genlerin ortalama değeri çıkarılıp standart sapmaya

bölündüğünde 2 den büyükse 1, değilse 0'dır. Gen i için son etiket tahmini $\hat{z}_i$, genin farklı ifade olup olmadığı ile ilgili son olasılık (posterior probability) değerini vermektedir.

Bu çalışmada Bölüm 3.2'de belirtildiği gibi zaman serisi mikrodiziler ile çalışırken ilgili genin farklı ifade olma olasılığını hesaplamak için Dean ve Raftery tarafından tanıtılan NUDGE metodu kullanılmıştır [50]. Bu metot ile veriler üzerinde çalışabilmemiz için istatistiksel hesaplama ve grafikler için geliştirilmiş olan R ücretsiz yazılım ortamı kullanılmıştır. NUDGE metotlarını içeren hazır bir kütüphane R için mevcuttur. R üzerinde geliştirdiğimiz bir betik yardımıyla NUDGE metotlarını da kullanarak önceden organizasyonunu yaptığımız veriler üzerinde çalışarak her bir örnek için Şekil 3.2'de içeriği görülen parmak izi dosyaları oluşturulmuştur.

fp_nudge_intersec_GSE2241.txt

	GENESYM	PDE-LAST	LAST-SIGN	PDE-MAX	MAX-SIGN
2	AADAC	0.827461938520494	+	0.999754479268718	+
3	AAK1	0.00364047871768258	-	0.0176316220935381	-
4	AAMP	0.00381425081675157	-	0.0160019350904402	-
5	AANAT	0.00500156828097298	-	0.994151191222366	-
6	AARS	0.00360075085918854	-	0.017150893746724	-
7	AASDHPPT	0.00360541931904146	-	0.017369351196434	-
8	AASS	0.793086438035566	+	0.793086438035566	+
9	AATF	0.00360260739264795	-	0.0158760688520576	-
10	AATK	0.00847033826792221	-	0.0332087338667708	-
11	ABAT	0.00474516825781679	-	0.0158736276755379	-
12	ABCA1	0.00360723031233912	-	0.0164584555101391	-
13	ABCA12	0.0350746219075689	-	0.0651771332386043	-
14	ABCA2	0.00357705421601295	-	0.0458558892395672	-
15	ABCA3	0.00435281514661512	-	0.0489156346654525	-
16	ABCA4	0.00524176976322677	-	0.0293394994416055	-
17	ABCA5	0.0107959202453866	+	0.0217981131991156	-
18	ABCA6	0.0148925722973766	+	0.0247757883015223	-
19	ABCA8	0.00541958634586259	-	0.428569954962363	-
20	ABCB1	0.00377785218404048	-	0.0159874494564749	-
21	ABCB11	0.00432191806067728	-	0.0336779801944467	-
22	ABCB4	0.00428350745773531	-	0.0168179264424817	-
23	ABCB6	0.00408719740666552	-	0.0254415517603119	-
24	ABCB7	0.00582603936910031	-	0.0593080415765779	-
25	ABCB8	0.00485233333110036	-	0.0162493265344983	-
26	ABCB9	0.00602522731312805	-	0.0263003527724867	-

Şekil 3.2 Her örnek için oluşturulan parmak izi dosyası

Şekil 3.2'de sırasıyla, GENESYM ilgili probeID'nin veya diğer bir ifadeyle gen sembolü; PDE-LAST zaman serisi mikrodizi içerisinde ilk ve son noktalar alınarak NUDGE ile hesaplanmış farkı ifade olma olasılığı; LAST-SIGN, ilk ve son nokta arasındaki değişimin yönü; PDE-MAX ilk nokta ile zaman serisi mikrodizi içerisindeki tüm noktalar karşılaştırıldıktan sonra alınan en yüksek farklı ifade olma olasılığı; ve son olarak MAX-SIGN yine PDE-MAX'da en yüksek değeri veren noktaların değişim yönünü göstermektedir. Burada, artı (+) ifadesi ilk noktaya göre diğer noktanın farkı ifade olma olasılığının yüksek eksi (-) ise düşük olduğunu göstermektedir.

Bölüm 3.2'de anlatıldığı üzere yapacağımız deneyde karşılaştırmalar her örnek için oluşturulan parmak izi dosyası içerisinde belirtilen LAST_DE (PDE-LAST) ve MAX_DE (PDE-MAX) değerlerine göre teker teker yapılmaktadır. Bunun yanında, her bir örnek için iki farkı parmak izi dosyası üretilmiştir. Bölüm 3.5.1'de bahsedileceği üzere bu parmak izi dosyaları birleşim ve kesişim ismiyle adlandırılmış gen sembol veritabanı içeriğine göre oluşturulmuştur.

3.4 Benzerlik Ölçümleri

Tüm deneyler, parmak izi vektörüyle gösterildiğinde, aralarındaki benzerliği modelleme bu ilişkili vektörler arasındaki uzaklığı hesaplayarak yapılmaktadır. Bu çalışmada, dört uzaklık metriği tartışılmaktadır: Öklid Uzaklığı (Euclidean Distance), Pearson Bağıntı Katsayısı (Pearson Correlation Coefficient), Spearman'ın Derece Bağıntı Katsayısı (Spearman's Rank Correlation Coefficent) ve Tanimoto Uzaklığı (Tanimoto Distance). Bunlar kısaca aşağıda açıklanmaktadır.

3.4.1 Öklid Uzaklığı

Matematikte, Öklid uzaklığı veya Öklid metriği iki nokta arasındaki doğrusal uzaklık olarak tanımlanmaktadır. Bu uzaklık cetvel ile ölçülebilir. Pisagor formülü ile tanımlanmaktadır. Bu formül uzaklık olarak kullanıldığında Öklid uzayı bir metrik uzayı olur. İlgili standart Öklid standardı olarak isimlendirilir. Eski literatür, bu

metrikten Pisagor metriği olarak söz etmektedir. Z ve Y olmak üzere n uzunluğundaki iki parmak izi vektörünün birbirine olan uzaklığı aşağıdaki gibi tanımlanmaktadır (3.2):

$$d(Z,Y) = \sqrt{\sum_{i=1}^{N}(z_i - y_i)^2} \tag{3.2}$$

Diğer bir ifadeyle, Öklid uzaklığı iki vektörün ilgili elementlerinin arasındaki farkların karelerinin toplamının karekökü'dür.

3.4.2 Pearson Bağıntı Katsayısı

0 ve 1 arasında derecelenen Öklid uzaklığına benzememektedir. Sayısal olarak, Pearson bağıntı katsayısı doğrusal regresyonda kullanılan bağıntı katsayısı ile aynı şekilde ifade edilir. -1 ila +1 arasında değer almaktadır. +1 değeri iki veya daha fazla değişken arasında mükemmel pozitif ilişkinin sonucudur. Aksine, -1 değeri ise mükemmel negatif bir ilişki olduğunu göstermektedir. 0 değeri ise ilişki olmadığını gösterir. Pearson bağıntı katsayısı (Pearson correlation coefficeint veya Pearson product-moment correlation coefficient) bilim alanında iki değişken arasındaki doğrusal bağımlılığının derecesini bulmak için yaygın şekilde kullanılmaktadır. Bu benzerlik ölçümü şöyle tanımlanmaktadır (3.3):

$$d(Z,Y) = \frac{N\sum z_i\, y_i - \sum z_i \sum y_i}{\sqrt{N\sum z_i^2 - (\sum z_i)^2} - \sqrt{N\sum y_i^2 - (\sum y_i)^2}} \tag{3.3}$$

Bu eşitlikte, (Z,Y) veri neslerini ve N ise özelliklerin toplam sayısını ifade etmektedir.

3.4.3 Spearman'ın Derece Bağıntı Katsayısı

Spearman'ın derece bağıntı katsayısı (Spearman's rank correlation coefficient veya Spearman's rho) iki değişken arasındaki ilişkinin gücünü ölçmeye izin verir. İstatistikte, iki değişken arasındaki istatistiksel bağımlılığın parametrik

olmayan (nonparametric) ölçümü olarak tanımlanmaktadır. Sıkça Yunan harfi olan ρ veya r_s ile gösterilir. Tekdüze ilişki (monotonic relationship) Spearman derece-sıra (rank-order) bağıntısının önemli temel varsayımıdır. Eğer tekrarlayan değerler yoksa, her bir değer diğerinin mükemmel tekdüze fonksiyonel olduğu zaman -1 ve +1 olan mükemmel Sperman bağıntısı oluşur. Spearman derece bağıntı katsayısı aşağıdaki formül ile tanımlanmaktadır (3.4).

$$p = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.4}$$

Bu eşitlikte d ikili dereceler (paired ranks) arasındaki fark ve n ise durum sayısıdır.

3.4.4 Tanimoto Uzaklığı

Tanimoto katsayısı ikilik (binary) parmak izleri için en yaygın kullanılan metriktir. Bu çalışmada belirli boyuttaki vektör içerisindeki her bir bit ilgili genin farklı ifade olup olmadığını göstermektedir. Parmak izleri veri setleri içerisindeki diğer bir ifadeyle vektör içerisindeki her bir PDE (Probability Value for Differentialy Expressed Gene) değeri reel sayısıdır. Bu sebepten, tüm PDE değerleri ikilik değere çevrilmiştir. Bu çevirme işlemi için iki metot kullanılmıştır. Bu metotlar eşik değeri tabanlı (threshold-based) metot ve yüzde tabanlı (percent-based) metotlardır.

Eşik değeri tabanlı metotta, PDE reel sayılar PDE>=eşik değeri veya PDE<eşik değeri durumuna göre 1 veya 0 ile değiştirilmiştir. Yüzde tabanlı metotta, ilk olarak veri kümesi içerisindeki genler PDE değerlerine göre azalan yönde sıralanmıştır. Sonra, belirlenen belirli bir yüzde değerine göre bu sıralanan veri kümesi içerisinde en üstteki yüzde içerisindeki kalan kısım 1 olarak, geriye kalanlar ise 0 olarak işaretlenmiştir. Ek olarak, en iyi sonucu elde etmek amacıyla eşik ve yüzde parametreleri farklı değerler ile incelenmiştir. Bu metodun matematiksel ifadesi şöyledir (3.5):

$$t(Z, Y) = \frac{\sum_i (z_i \cap y_i)}{\sum_i (z_i \cup y_i)} \tag{3.5}$$

Verilen eşitlikte Z ve Y biteşlemler, z_i Z'nin i inci bitidir ve ∩,∪ bit bit operatörlerdir.

3.5 Veri Kümeleri ve Organizasyonu

Çalışmanın bu bölümünde deneyde kullanılan veriler hakkında bilgi verilmekte ve veriler üzerinde deneysel çalışmaya geçilmeden önce yapılan veri organizasyonu anlatılmaktadır. Ayrıca, indirilen ham GEO verilerinden çalışmamız için gerekli özet veya ön işlenmiş veri kümelerinin nasıl ve hangi yöntemlerle yapıldığı hakkında detayı bilgi içermektedir.

3.5.1 Veri kümeleri

Bu çalışmada, örnekler Gene Expression Omnibus (GEO) deposundan alınmıştır. Deneyde kullanılan örnek veritabanı 111 mikrodizi profilinden oluşmaktadır. Verilerden ilk 46 tanesi meme kanseri ile ilgili yapılmış zaman serisi deneylerine aitken diğer kalan 65 veri rahim kanseri, prostat kanseri, beyin kanseri, kolon kanseri, pankreas kanseri, lösemi ve diyabet gibi farklı hastalıklar ile ilgili yapılmış zaman serisi deneylerine aittir. Her bir örnekte yaklaşık 20,000 ve 40,000 arasında probe bulunmaktadır. Örnekler zaman serisidir ve genellikle her bir zaman noktası için en azından 2 kopyadan (replicate) oluşmaktadır.

Ek olarak, tüm örnekler insan genomuna ve farklı farklı platformların değişik versiyonlarına aittir. Fakat, veri tabanında en fazla Affymetrix platformundan veri mevcuttur. Platformlar ile ilgili bilgi Çizelge 3.1'de verilmiştir. Parmak izi çıkarma işleminden önce, tüm örneklerin probeID'leri onların karşılık geldiği özgül Gen Sembolleri ile eşlenmiştir. Böylece, platformda bağımsız olarak karşılaştırma yapılabilmektedir.

Çizelge 3.1 Veri kümelerinin alındığı platformların listesi

No	**Platform Adı**
1	Affymetrix
2	Illumina
3	Agilent Technologies

Her bir örnek için, Kesişim Parmak İzi (Intersection Fingerprint - IF) ve Birleşim Parmak İzi (Union Fingerprint - UF) olarak adlandırılan iki tür parmak izi veri tabanı oluşturulmuştur. IF tüm örnek veri kümelerinde bulunan gen sembollerinden oluşmaktadır. Diğer taraftan, UF her bir örnek veri kümesinde olan gen sembollerinden oluşmaktadır. Her bir IF veri kümesi 7,076 adet aynı gene sahiptir.

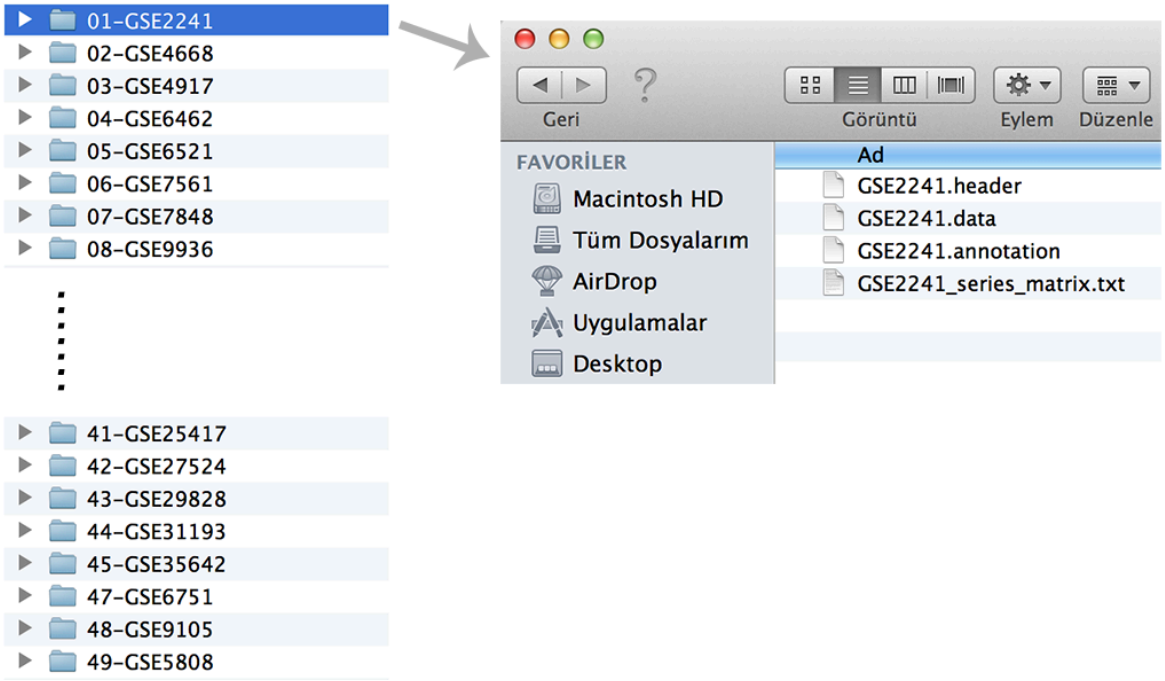
Çalışmada kullanılan verilerin ait oldukları Platform Adı ve GEO ID'si Çizelge 3.2'de verilmiştir.

Çizelge 3.2 Veri kümelerinin ait olduğu platform ve GEO ID'si

No	GEO ID	PLATFORM
1	GPL91	Affymetrix
2	GPL96	Affymetrix
3	GPL570	Affymetrix
4	GPL571	Affymetrix
5	GPL887	Agilent Technology
6	GPL6102	Illumina
7	GPL6883	Illumina
8	GPL6884	Illumina
9	GPL8300	Affymetrix
10	GPL9419	Affymetrix

3.5.2 Veri organizasyonu

Çalışmada yapılacak olan deneyin istatistiksel ve matematiksel modeli belirlendikten sonra bu işi gerçekleştirecek olan bilgisayar algoritmalarının ve kullanılacak programların özelliklerine göre veriler uygun bir format ve yapıya çevrilmiştir. GEO'dan indirilen veriler ilk önce ham veri olarak bilgisayar sabit diskinde bir klasör altına toplanmıştır. Sonra, indirilen veriler parçalanarak seriden alınan her bir örnek grubu için ".data", ".annotation" ve ".header" dosyası oluşturulmuştur. Oluşturulan bu dosyalar yine belirli bir standart içerisinde sabit disk üzerinde oluşturulan bir klasör altında depolanmıştır. Şekil 3.3'te bu verilerin klasör yapısı altındaki organizasyonu gösterilmektedir.



Şekil 3.3 Veri organizasyonu

Bu organizasyonda her bir örnek için oluşturulan dosyalardan ".header" içerisinde deneyin ne zaman ve kimin tarafından yapıldığı, başlığı, durumu, özeti gibi deney ile ilgili açıklayıcı bilgiler bulunmaktadır. ".data" içerisinde ilgili deneye ait genlerin örnek numaralarına (GEO Sample, GSM) ve probeID'lerine göre ifade değerleri yer almaktadır (Şekil 3.4). ".annotation" içerisinde ise deneyde kaç zaman noktasının olduğu, her zaman noktası için kaç tekrar alındığı ve ilgili deneyin hangi hastalık için yapıldığının bilgisini içermektedir (Şekil 3.5).

GSE2241.data

	"ID_REF"	"GSM41162"	"GSM41163"	"GSM41164"	"GSM41165"	"GSM41166"	"GSM41167"
2	"1007_s_at"	669.1	173.9	78.1	159.4	244.2	144.4
3	"1053_at"	294.8	224.7	38.3	123.3	141.1	165.6
4	"117_at"	47.9	35.3	8.6 19.4	21.3	45.7	
5	"121_at"	823.5	419.5	109.9	376.8	277.2	344.3
6	"1255_g_at"	25.6	13.1	12.6	12.3	11.4	9.3
7	"1294_at"	218.5	77.7	24.2	56.8	95.3	86.3
8	"1316_at"	69.7	23 16	21.8	37.2	28.9	
9	"1320_at"	10.2	34.1	13.2	21.1	4 19	
10	"1405_i_at"	3.1 1.8	1.2 13.2	2.2	1.1		
11	"1431_at"	26.5	16.3	16.6	27.1	21.2	35.3
12	"1438_at"	17.4	17 2.2	23.9	7 62.8		
13	"1487_at"	442.4	261.6	62.8	244.7	158.8	196.1
14	"1494_f_at"	172.1	86.1	19.4	97.5	74.9	90.8
15	"1598_g_at"	600.1	305.6	73.4	213.9	202.7	232.5
16	"160020_at"	350.8	225.1	47.3	205.8	161.1	157.7
17	"1729_at"	247.9	40.8	7.7 9.7	71.5	10.9	
18	"1773_at"	70.4	49.3	17.6	69.1	39.3	32.4
19	"177_at"	86.6	23.9	12.3	41.7	57.1	31.2
20	"179_at"	699.2	421.4	274.7	725.1	394.5	1369.1

Şekil 3.4 “.data” dosyası içeriği

Şekil 3.4’te görüldüğü gibi yapılan meme kanseri mikrodizi deneyi sonucunda tek tek GSM numaraları verilen örneklerin ilgili platformdaki gene verilen probeID’leri ve genin ifade değerli yer almaktadır. Bu örnekte sırasıyla 0saat-0saat-6saat-6saat-9saat-9saat olmak üzere 3 zaman noktası ve her zaman noktası için 2 tekrar bulunmaktadır.

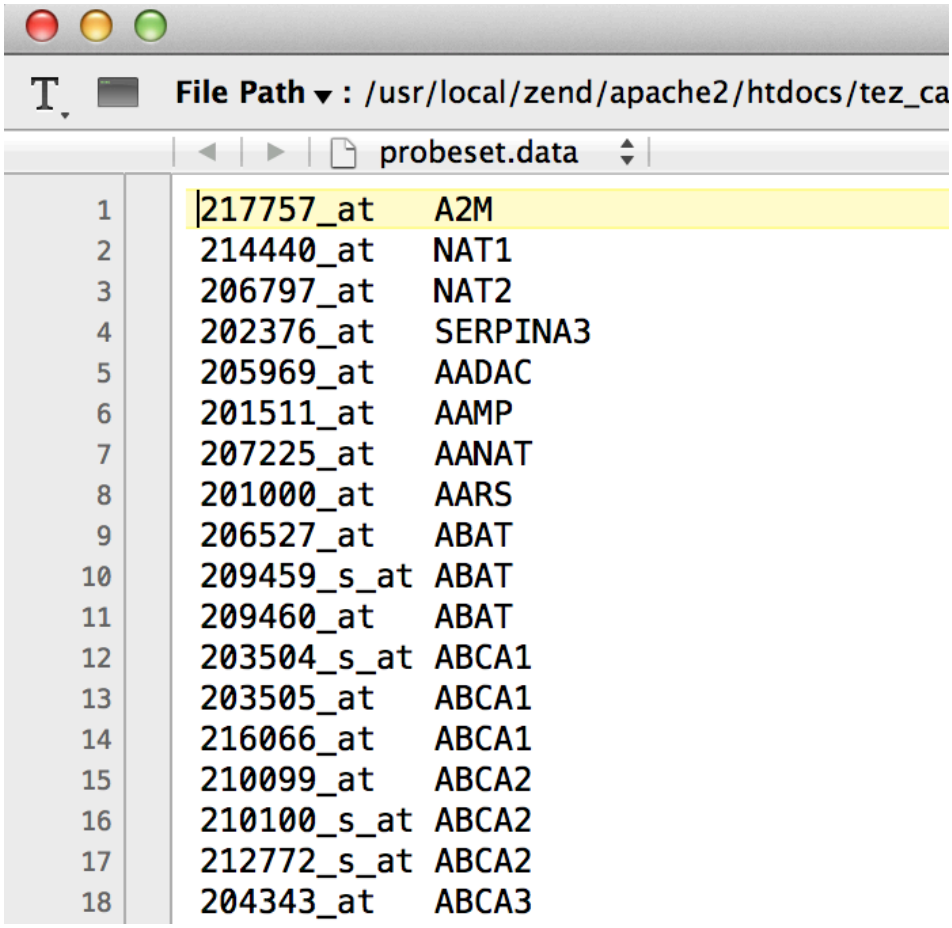
GSE2241.annotation

```
1  >timepoints
2  0h  6h  9h
3  >rep
4  2
5  >disease
6  breast
```

Şekil 3.5 Oluşturulan “.annotation” dosyası içeriği

Veri organizasyonunu yaptıktan sonra diğer önemli çalışma ise probeID’lere karşılık gelen gen sembollerinin bulunup bu tüm veriler için tek bir gösterimin sağlanmasıdır. Çünkü, probeID’ler platformlara özgü olan bir gen kodlama biçimidir ve bu gösterim platformdan platforma her gen için değişiklik

göstermektedir. Bu sebeple Şekil 3.6'da gösterildiği gibi tüm platformlara ait olan probeID'ler "probeset.data" dosyası içerisinde karşılıklarına gen sembolleri eklenerek gösterilmiştir. Bu süreçte yine GEO üzerinden indirilen GEO ID'leri Çizelge 3.4'te verilmiş olan GPL dosyaları kullanılmıştır. MAC OS X işletim sistemi üzerinde geliştirilen bir küçük uygulama sayesinde bu GPL dosyaları birleştirilerek tek bir dosya haline getirilmiştir.



Şekil 3.6 ProbeID ve karşılık gelen gen sembol listesi

4. DENEYSEL SONUÇLAR

Çalışmamızın bu bölümünde, organizasyonu ve ön işlemesi tamamlanarak deneye hazırlanmış örnekler üzerinde yapılan çalışmalar ve sonuçları yer almaktadır. Örneklerin kesişim ve birleşim parmak izleri veri tabanları ile bu veri tabanlarının LAST_DE ve MAX_DE olarak hesaplanan farkı olma olasılıkları değerlerinin ve ayrıca bu değerler ışığında farkı yöntemlerle genin farklı ifade olmuş veya olmamış olarak işaretlenmesi detaylı olarak anlatılmaktadır. Belirlenen FİO genler üzerinden benzerlik metrikleri uygulanmış, bu sonuçların Alıcı İşletim Karakteristiği (Receiver Operation Characteristic - ROC) ile performans ölçümleri yapılmış ve sonuçlar değerlendirilmiştir.

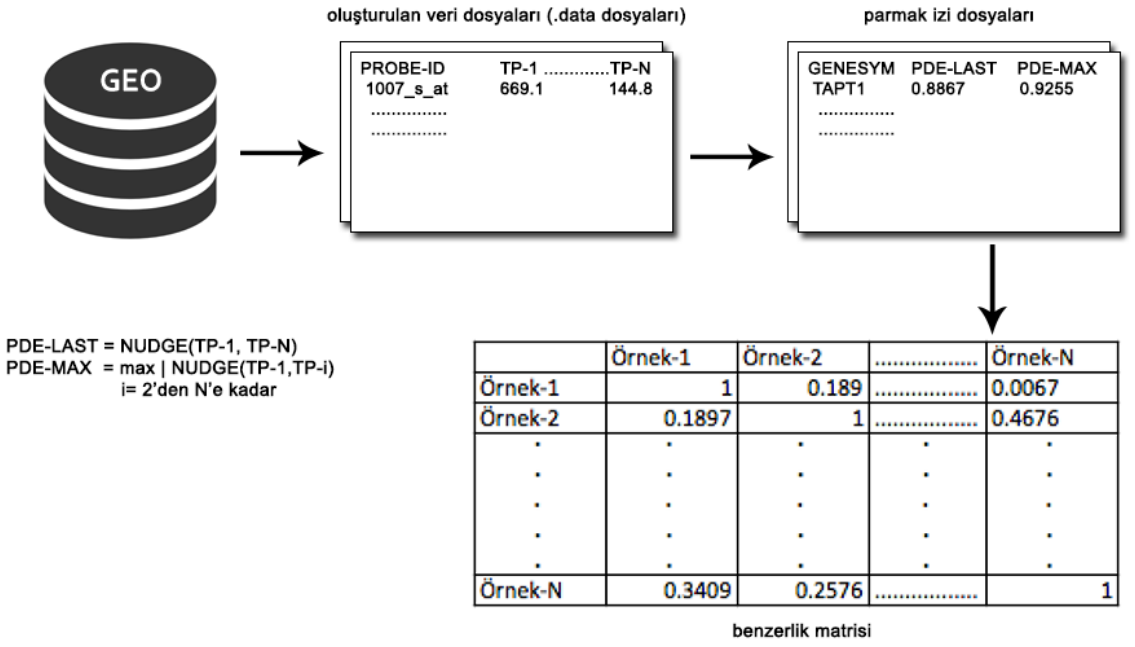
4.1 Deneysel Hazırlık

Bu çalışmada, tüm deneylerden toplu bir veri kümesi oluşturulmuş ve her bir meme kanseri deneyi kendisi hariç diğer tüm deneyler ile karşılaştırılarak içerik tabanlı veri tabanı arama simülasyonu yapılmıştır. Bu çalışmadaki varsayımımız bir meme kanseri deneyi sorgulandığında, yakınlık derecesine göre büyükten küçüğe sıralanmış olarak çekilen deneyler listenin en tepesinde başka meme kanseri örneklerinin olmasıyken diğer kanser tiplerine ait örneklerinin listenin en altında olmasıdır. Bilgi çıkarım performansı Alıcı İşletim Karakteristiği eğrisi (ROC) ile değerlendirilmiştir. Her bir pozitif örnek (meme kanseri mikrodizi deneyi) için ROC skoru, ilişkili ROC eğrisi altındaki alan hesaplanarak bulunmaktadır. Genel performans tüm mikrodizi deneyler için çizilmiş diğer bir eğri tarafından gösterilmektedir. Tüm alternatif yaklaşımlar için Ortalama ROC skorları ayrıca çalışmada raporlanmıştır. ROC değerinin yüksek olması geri getirim performansının daha iyi olduğunu göstermektedir. ROC skorunun 1 olması ise mükemmel bir sistem olduğunu gösterirken, 0 olması herhangi bir pozitif bulunamadı anlamına gelmekte yani sistemin çok kötü olduğunu göstermektedir.

4.1.1 Benzerlik matrisi

Benzerlik matrisi veri noktaları arasındaki benzerliği ifade eden skor matrisidir. Matrisin her bir elementi iki veri noktası arasındaki benzerlik ölçümünün değerini içermektedir. Bu çalışmada Bölüm 3.4'te anlatılan benzerlik ölçümleri kullanılmıştır. İlk önce hedef örneklerimiz olan meme kanseri örnekleri 2 boyutlu listenin en başında olacak şekilde örneklerin diğer örneklere olan yakınlıkları veya tersten düşünecek olursak uzaklıkları hesaplanmıştır. Bu durumda çalışmanın bu bölümünde, her yöntem ve parametre kombinasyonu içi NxN'lik diyagonal matrisler elde edilmiştir.

Benzerlik ölçümlerinden Tanimoto, ikilik (binary) noktalar arasında hesaplama yapan bir metot olduğundan benzerlik matrisinin hesaplanmasından önce tüm PDE değerleri ikilik tabana çevrilmiştir. Yani, FİO genler bulunmuştur. Fakat diğer metotlar için bu işlemin yapılmasına gerek kalmamıştır. Çünkü, o metotlar her gen için hesaplanan PDE değerini olduğu gibi kabul etmektedir. Şekil 4.1'de benzerlik matrisi oluşturmasıyla ilgili temel süreç aşamaları gösterilmektedir.



Şekil 4.1 Benzerlik matrisinin oluşturulma aşamaları

4.1.2 Alıcı İşletim Karakteristiği (ROC)

Alıcı işletim karakteristiği (ROC) analizi, sınıflandırma problemlerinin değerlendirilmesinde çok önemli bir araçtır. Bu sınıflandırmada kesin referans (ground truth) bir ikili referans standardıdır (örneğin; hastalığın olup olmaması), başka bir ifadeyle iki-sınıflı sınıflandırma problemleridir [51]. Kesin olmayan ortamlarda, ROC analizi özellikle kullanışlıdır. Çünkü ROC, operasyon kondisyon aralığında yarışan modellerin karşılaştırılmasında anlam sağlamaktadır. ROC altında kalan alan (Area Under Curve - AUC) operasyon kondisyonlarına göre değişmediğinden dolayı önemli performans ölçüm yöntemi olmuştur. Hem de, sınıf miktarlarındaki değişikliklerden ötürü performanstaki azalıp çoğalmalar analiz edilebilmektedir. Çünkü, bu dalgalanmalar ROC boyunca değişikliklere zorunludur [52].

Sinyal algılama teorisinde, ROC veya basitçe ROC eğrisi ikili sınıflandırma sisteminin performansını tanımlayan bir grafiksel eğridir. ROC eğrisi, ikili sınıflandırma sistemlerinde ayrım eşik değerinin farklılık gösterdiği durumlarda, hassasiyetin kesinliliğe olan oranıyla ortaya çıkmaktadır. ROC daha basit anlamda doğru pozitiflerin, yanlış pozitiflere olan kesri olarak da ifade edilebilir [53].

Eğer bir sınıflandırıcı ve bir örneğimiz varsa, mümkün olan temel dört çıktıdan bahsetmek mümkündür. Eğer örneğimiz pozitif ise ve sınıflandırıcı tarafından pozitif olarak sınıflandırılmışsa örneğimiz doğru pozitif (true positive); eğer negatif sınıflandırılmış ise bu durumda yanlış negatif (false negatif) olur. Eğer örneğimiz negatif ve negatif olarak sınıflandırılmışsa doğru negatif (true negative); eğer pozitif sınıflandırılmışsa ise yanlış pozitif olur. Verilen sınıflandırıcı ve örnek kümesi için, örnek setlerinin yanlış sınıflara yerleştirilmesini ifade edecek 2x2'lik bir karışıklık matrisi (confusion matrix) oluşturulabilir. Bu matris birçok genel ölçüm için temel oluşturur (Çizelge 4.1).

Çizelge 4.1 Karışıklık matrisi

	Öngörülen Sınıf			
Gerçek Sınıf		0	1	**Toplam**
	0	DN	YP	P
	1	YN	DP	N

Çizelge 4.1'de verilen karışıklık matrisi üzerindeki ifadelerden temel olarak aşağıdaki ölçümler çıkarılmaktadır (Eşitlik 4.1; Eşitlik 4.2; Eşitlik 4.3; Eşitlik 4.4; Eşitlik 4.5).

$$dp\ oranı\ (sensitivity = duyarlılık) = \frac{DP}{N} \tag{4.1}$$

$$yp\ oranı = \frac{YP}{P} \tag{4.2}$$

$$dn\ oranı\ (specificity = özgüllük) = \frac{DN}{P} \tag{4.3}$$

$$yn\ oranı = \frac{YN}{N} \tag{4.4}$$

$$kesinlik\ (precision) = \frac{DP}{YP+DP} \tag{4.5}$$

Deneyin ilk aşamasında Bölüm 4.1.1'de bahsedildiği üzere parmak izi dosyaları üzerinde gerçek zamanlı olarak çalışan ve verilen parametreler göre tüm veri kümesi üzerinde gezerek benzerlik matrisini oluşturan bir algoritma geliştirilmiştir. Uygulamış olduğumuz içerik tabanlı arama modelinde temel olarak performansı etkileyen temel iki ana unsur bulunmaktadır. Bunlar FİO genlerin belirlenmesi ve benzerlik matrisinin oluşturulmasıdır. Bu çalışmada önerilen yöntemler FİO için zaman serisi verilerin yorumlanması ve benzerlik matrisi için önerdiğimiz benzerlik metrikleridir. Bu ana unsurların farklı kombinasyonlarıyla çalışılarak oluşturulan farklı benzerlik matrislerinin diğer bir ifadeyle farklı modellerin performansı ROC ile değerlendirilmiştir. ROC çalışması aşamasında temel iki algoritma kullanılmıştır. Biricisi her bir hedef örnek için ROC skorun bulunması; ikincisi bu ROC skor

değerlerinden oluşan grafiğin altında kalan alanı hesaplayarak (Area Under Curve - AUC) çalışmada farklı modellerin karşılaştırılması için kullandığımız nihai deney performans değerinin bulunmasıdır. Sonrasında bu AUC değerleri karşılaştırılarak sonuçları Bölüm 4.2'de verilmiş ve yorumlanmıştır. Farklı parametreler ve yöntemler ile oluşturulmuş benzerlik matrislerine aşağıdaki algoritma her bir deney için tek tek uygulanmıştır.

Algoritma 1 : *ROC Skor*

Girdi : *L*, bir meme kanseri örneğine en yakın tahmin edilen sınıf değeri (etiket) dizisi. (1 = meme kanseri, 0 = meme kanseri değil)

Çıktı : R, ROC skoru

```
/*Doğru Pozitiflere Başlangıç Değerinin Atanması*/
tp=0

/*Yanlış Pozitiflere Başlangıç Değerinin Atanması*/
fp=0

/*ROC Skor İçin Başlangıç Değerinin Atanması*/
R=0

/*Sıralanmış Her bir Etiket İçin Doğru Pozitif ve Yanlış Pozitif Değer Hesabı*/
for L as etiket
  if etiket=1
      tp=tp+1
  else
      fp=fp+1
      R=R+tp
  end if
end for

/*Doğru Pozitif ve Yanlış Pozitif Değerlerine Göre ROC Skor Hesabı*/
if tp=0
   R=0
else
```

```
 if fp=0
     R=1
 else
     R=R/tp*fp
 endif
endif
```

Her bir yöntem için farklı yüzde ve eşik değerleri ile ROC skorları yine her bir örnek için ayrı ayrı hesaplanmıştır. Örneğin, Tanimoto benzerlik metriği kullanılarak yapılacak bir deneyde PDE_MAX yöntemi (maximum NUDGE) kullanılarak hesaplanmış farklı ifade olma olasılıklarına göre FİO olarak etiketlenecek genlerin seçiminde 0.5 eşik değeri kullanılması. Aynı şekilde FİO genlerin işaretlenmesinde sıralama yöntemi kullanarak küme içerisinde %5'in FİO olarak kabul edilmesi. Tanimoto için yüzde ve eşik değelerine göre en iyi sonuç elde edilene kadar yukarıdaki örnekte verildi gibi oranlar değiştirilmiş ve sonuçlar analiz edilmiştir.

Her deneyde her bir meme kanseri örneği için ayrı ayrı hesaplanan ROC skorlarından elde ettiğimiz değerler ile bir dizi oluşturulmaktadır. İlgili deneyin veya diğer bir değişle ilgili sınıflandırıcının performansını ölçmek ve bunları karşılaştırabilmek için tek bir sayısal değere ihtiyaç duyulmaktadır. Bunu için kullanılan genel metot ROC eğrisi altına kalan alanı kısaca AUC'yi hesaplamaktadır [54]. Çünkü, AUC birim karenin alanıdır ve değeri her zaman 0 ile 1 arasındadır. Ancak, hiçbir gerçekçi sınıflandırıcının AUC değeri 0.5 altında olmamalıdır. Çünkü, rastgele tahminleme (0,0) ve (1,1) arasında çapraz çizgi oluşturmaktadır ve bunun alanı 0.5'dir. Çalışmamızda AUC'nin hesaplanması için aşağıdaki algoritma kullanılmıştır.

Algoritma 2 : *AUC Hesaplanması*

Girdi : *K*, tüm önceden etiketi bilinen meme kanserlerinin Algoritma 1 kullanılarak bulunmuş ROC değerlerini büyükten küçüğe sıralanmış olarak içeren dizidir.

Çıktı : *A*, ROC eğrisi altındaki alan

```
/*Toplam ROC Skora Başlangıç Değerinin Atanması*/
tr=0

/*Örnek Sayısına Başlangıç Değerinin Atanması*/
c=0

/*AUC Değeri İçin Başlangıç Değerinin Atanması*/
A=0

for K as r
   tr=tr+r
   c=c+1
end for
A=tr/c
```

Yukarıdaki bölümlerde bahsedildiği üzere Tanimoto ikilik değerler ile çalıştığından PDE değerlerinin ikilik sisteme çevrilmesi aşamasındaki yöntemler bu metot ile oluşturulan benzerlik matrisinin elemanlarını etkileyecektir. Çizelge 4.2 ve Çizelge 4.3'te birleşim/kesişim PDE_MAX ve PDE_LAST değerleri için Tanimoto metriği ile oluşturulan benzerlik matrislerinin hesaplanan ROC değerleri listelenmektedir. Bu çizelgede PDE_MAX ve PDE_LAST ile oluşturulan temsilci parmak izlerinin yüzde ve eşik olmak üzere verilen değerler ile yapılan deneylerden en iyi sonucu vermiş olanlarının ROC listeleri gösterilmektedir. Örneğin Çizelge 4.2'nin GSE No (Örnek GEO Numarası) ve ROC Değeri'ni içeren ilk sütununda, PDE_LAST parmak izi vektörüne ikilik çeviri için yüzde yaklaşımı uygulanarak elde edilen en iyi sonucun %5 değeri ile alındığını göstermektedir.

Çizelge 4.2 Kesişim PDE_LAST ve PDE_MAX için Tanimoto ile hesaplanmış benzerlik matrisinden 46 meme kanseri örneğinin ROC Değerleri

Tanimoto							
%5 Last		**0.2 Last**		**3% Max**		**0.2 Max**	
GSE No.	**ROC D.**	**GSE No.**	**ROC D.**	**GSE No.**	**ROC D.**	**GSE No.**	**ROC D.**
GSE7848	0,781	GSE13009	0,759	GSE7848	0,801	GSE6521	0,725
GSE7848	0,781	GSE6521	0,714	GSE7848	0,755	GSE6521	0,702

GSE7848	0,774	GSE6462	0,713	GSE6521	0,713	GSE7848	0,698
GSE6521	0,744	GSE7848	0,711	GSE9936	0,699	GSE6462	0,690
GSE9936	0,727	GSE6521	0,711	GSE6521	0,696	GSE7848	0,688
GSE9936	0,724	GSE6462	0,708	GSE18494	0,690	GSE9936	0,688
GSE6521	0,720	GSE6462	0,697	GSE9936	0,689	GSE6462	0,679
GSE9936	0,702	GSE7848	0,691	GSE6462	0,680	GSE6462	0,675
GSE6462	0,696	GSE9936	0,690	GSE6462	0,679	GSE6462	0,667
GSE18494	0,695	GSE6521	0,687	GSE9936	0,677	GSE9936	0,660
GSE9936	0,688	GSE6462	0,686	GSE9936	0,670	GSE6521	0,659
GSE9936	0,687	GSE9936	0,685	GSE9936	0,663	GSE9936	0,648
GSE28305	0,687	GSE9936	0,685	GSE18494	0,655	GSE18494	0,646
GSE6462	0,678	GSE6462	0,683	GSE6521	0,651	GSE7848	0,646
GSE9936	0,672	GSE9936	0,681	GSE9936	0,651	GSE9936	0,646
GSE9936	0,670	GSE6462	0,681	GSE9936	0,649	GSE9936	0,644
GSE7561	0,663	GSE9936	0,676	GSE9936	0,644	GSE9936	0,640
GSE9936	0,663	GSE9936	0,673	GSE13009	0,637	GSE9936	0,638
GSE18494	0,662	GSE9936	0,672	GSE28305	0,634	GSE9936	0,636
GSE9936	0,660	GSE9936	0,668	GSE9936	0,634	GSE28305	0,633
GSE9936	0,659	GSE18494	0,663	GSE7561	0,632	GSE13009	0,628
GSE6462	0,656	GSE7848	0,661	GSE9936	0,630	GSE9936	0,627
GSE18494	0,648	GSE9936	0,661	GSE6521	0,629	GSE9936	0,627
GSE6462	0,648	GSE9936	0,658	GSE9936	0,629	GSE18494	0,626
GSE9936	0,644	GSE6462	0,652	GSE9936	0,627	GSE6521	0,623
GSE6462	0,640	GSE9936	0,649	GSE6462	0,624	GSE6462	0,619
GSE9936	0,639	GSE9936	0,647	GSE9936	0,619	GSE6462	0,619
GSE9936	0,635	GSE18494	0,636	GSE6462	0,618	GSE9936	0,612
GSE6462	0,626	GSE9936	0,628	GSE11506	0,611	GSE9936	0,608
GSE4917	0,622	GSE28305	0,614	GSE4917	0,605	GSE4917	0,607
GSE6521	0,619	GSE6462	0,612	GSE6521	0,604	GSE9936	0,591
GSE6521	0,614	GSE13009	0,611	GSE7848	0,599	GSE6462	0,589

GSE13009	0,598	GSE6521	0,588	GSE6462	0,594	GSE18494	0,568
GSE6462	0,586	GSE11324	0,578	GSE6462	0,591	GSE6462	0,567
GSE6521	0,565	GSE4917	0,577	GSE6462	0,587	GSE7561	0,562
GSE11324	0,545	GSE6521	0,554	GSE6462	0,559	GSE11324	0,539
GSE4917	0,545	GSE2241	0,552	GSE11324	0,550	GSE6521	0,539
GSE20361	0,533	GSE11352	0,546	GSE18494	0,543	GSE4917	0,535
GSE6462	0,527	GSE4917	0,539	GSE4917	0,504	GSE20361	0,530
GSE11506	0,519	GSE18494	0,515	GSE13009	0,459	GSE4668	0,519
GSE2241	0,504	GSE4668	0,510	GSE20361	0,427	GSE11506	0,495
GSE7561	0,437	GSE7561	0,505	GSE7561	0,411	GSE2241	0,460
GSE13009	0,436	GSE20361	0,414	GSE4668	0,394	GSE11352	0,438
GSE4668	0,387	GSE11352	0,387	GSE2241	0,386	GSE13009	0,423
GSE11352	0,347	GSE11506	0,362	GSE11352	0,381	GSE11352	0,392
GSE11352	0,292	GSE7561	0,109	GSE11352	0,305	GSE7561	0,281

Çizelge 4.3 Birleşim PDE_LAST ve PDE_MAX için Tanimoto ile hesaplanmış benzerlik matrisinden 46 meme kanseri örneğinin ROC Değerleri

Tanimoto							
%0.9 Last		**0.2 Last**		**%0.9 Max**		**0.2 Max**	
GSE No.	**ROC D.**	**GSE No.**	**ROC D.**	**GSE No.**	**ROC D.**	**GSE No.**	**ROC D.**
GSE7848	0,819	GSE13009	0,765	GSE7848	0,770	GSE7848	0,733
GSE7848	0,762	GSE6462	0,729	GSE7848	0,756	GSE7848	0,733
GSE9936	0,759	GSE6521	0,720	GSE9936	0,738	GSE6521	0,706
GSE9936	0,752	GSE7848	0,716	GSE7848	0,738	GSE9936	0,703
GSE13009	0,750	GSE6521	0,713	GSE13009	0,714	GSE6462	0,701
GSE6462	0,739	GSE9936	0,706	GSE9936	0,709	GSE6462	0,689
GSE9936	0,731	GSE6521	0,706	GSE6462	0,701	GSE6462	0,686
GSE9936	0,728	GSE6462	0,702	GSE6462	0,698	GSE9936	0,681
GSE9936	0,723	GSE9936	0,702	GSE6462	0,693	GSE6462	0,680
GSE9936	0,721	GSE6462	0,700	GSE18494	0,689	GSE6521	0,679

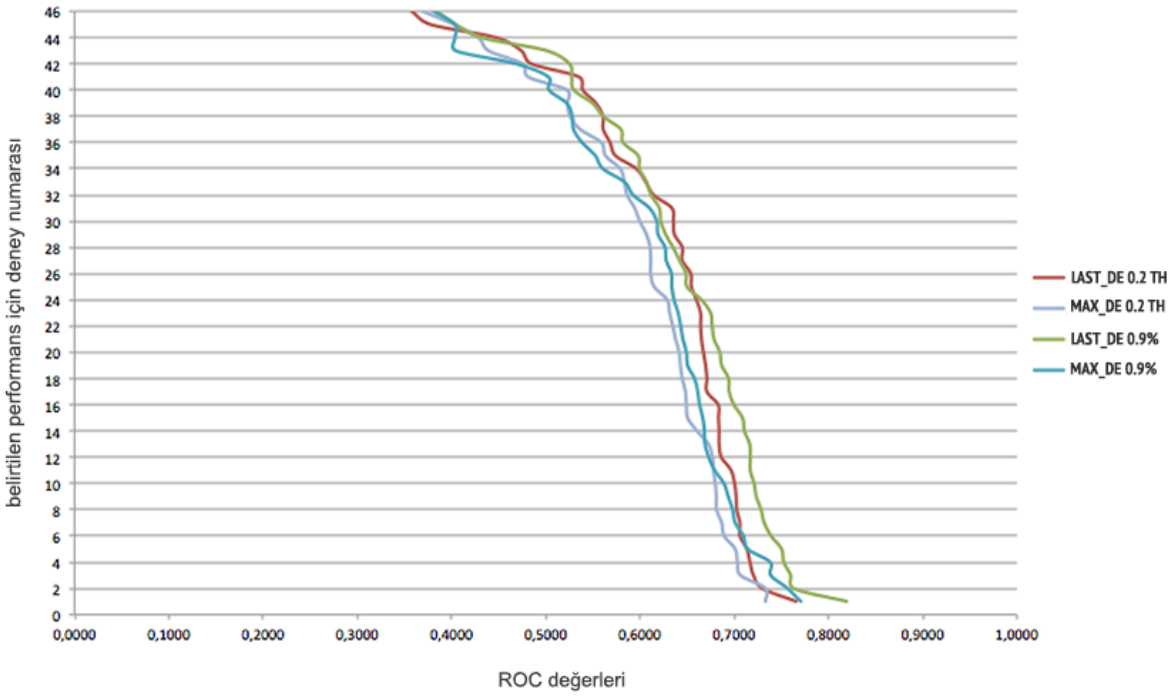
GSE9936	0,717	GSE9936	0,696	GSE9936	0,679	GSE9936	0,677
GSE9936	0,717	GSE13009	0,686	GSE9936	0,672	GSE6521	0,676
GSE9936	0,716	GSE6462	0,683	GSE9936	0,668	GSE7848	0,672
GSE7848	0,710	GSE7848	0,683	GSE9936	0,668	GSE9936	0,660
GSE6462	0,708	GSE9936	0,683	GSE9936	0,666	GSE9936	0,650
GSE6462	0,699	GSE9936	0,682	GSE9936	0,663	GSE9936	0,649
GSE6521	0,694	GSE9936	0,670	GSE6521	0,661	GSE6521	0,648
GSE9936	0,694	GSE9936	0,670	GSE9936	0,657	GSE18494	0,644
GSE9936	0,686	GSE6462	0,669	GSE6521	0,650	GSE9936	0,642
GSE9936	0,684	GSE9936	0,667	GSE9936	0,649	GSE9936	0,641
GSE28305	0,678	GSE9936	0,665	GSE9936	0,645	GSE9936	0,637
GSE6521	0,676	GSE6462	0,664	GSE6521	0,643	GSE9936	0,634
GSE9936	0,674	GSE9936	0,664	GSE6462	0,640	GSE9936	0,631
GSE6462	0,665	GSE6462	0,659	GSE4917	0,636	GSE18494	0,628
GSE18494	0,649	GSE18494	0,654	GSE28305	0,633	GSE28305	0,615
GSE6521	0,648	GSE7848	0,653	GSE9936	0,633	GSE13009	0,611
GSE7561	0,641	GSE9936	0,644	GSE18494	0,627	GSE6462	0,611
GSE13009	0,635	GSE18494	0,644	GSE9936	0,626	GSE9936	0,610
GSE6462	0,627	GSE9936	0,636	GSE6462	0,619	GSE4917	0,606
GSE6462	0,622	GSE28305	0,635	GSE6462	0,617	GSE9936	0,599
GSE18494	0,620	GSE9936	0,633	GSE6462	0,610	GSE6462	0,594
GSE4917	0,611	GSE6462	0,614	GSE13009	0,592	GSE6462	0,586
GSE6521	0,606	GSE6521	0,606	GSE6521	0,583	GSE6462	0,583
GSE6462	0,599	GSE6521	0,596	GSE11506	0,560	GSE6462	0,578
GSE6462	0,598	GSE4917	0,573	GSE6462	0,552	GSE7561	0,563
GSE4917	0,581	GSE11324	0,568	GSE11324	0,538	GSE4917	0,558
GSE11506	0,579	GSE18494	0,560	GSE7561	0,529	GSE20361	0,536
GSE11324	0,560	GSE2241	0,560	GSE4917	0,528	GSE6521	0,525
GSE2241	0,549	GSE7561	0,553	GSE4668	0,522	GSE11324	0,523
GSE18494	0,529	GSE4668	0,539	GSE18494	0,503	GSE4668	0,522

GSE4668	0,527	GSE4917	0,535	GSE6521	0,502	GSE7561	0,481
GSE7561	0,525	GSE11352	0,484	GSE7561	0,468	GSE2241	0,475
GSE6521	0,501	GSE11352	0,473	GSE20361	0,404	GSE13009	0,439
GSE20361	0,431	GSE20361	0,447	GSE2241	0,403	GSE11506	0,428
GSE11352	0,404	GSE11506	0,377	GSE11352	0,403	GSE11352	0,402
GSE11352	0,384	GSE7561	0,357	GSE11352	0,380	GSE11352	0,369

4.2 Deneysel Sonuç

Çalışma içerisinde Bölüm 4.1'de bahsedildiği gibi benzerlik matrisinin oluşturulması sürecinde Tanimoto metriği diğerlerinden farkı olarak uygulanmadan önce ek bir işleme daha gerek duymaktadır. Bu işlem örnek içerisindeki genlerin farklı ifade olma olasılığı değerlerinin ikilik bir yapıya çevrilmesidir. Bu nedenle tüm benzerlik metotları uygulanıp birbirleri ile karşılaştırılmadan önceden farklı parametre ve yöntemler için Tanimoto ile deneyler yapılmıştır.

Tanimoto uzaklığı ikilik formata dönüştürülecek olan iki farklı şema için uygulanmıştır. Parmak izi veri dosyalarının oluşturulması için ilk başta birleşim gen listesi kullanılmıştır. Şekil 4.2'de en iyi konfigürasyonlar ile uygulanmış bu metodun sonuç grafiği verilirken, Çizelge 4.4'te ortalama skorlar listelenmektedir. Bu deneydeki en iyi performans LAST_DE parmak izi verilerinin büyükten küçüğe doğru sıralanıp ilk %0.9'un farklı ifade olmuş olarak işaretlenmesiyle elde edilmiştir.

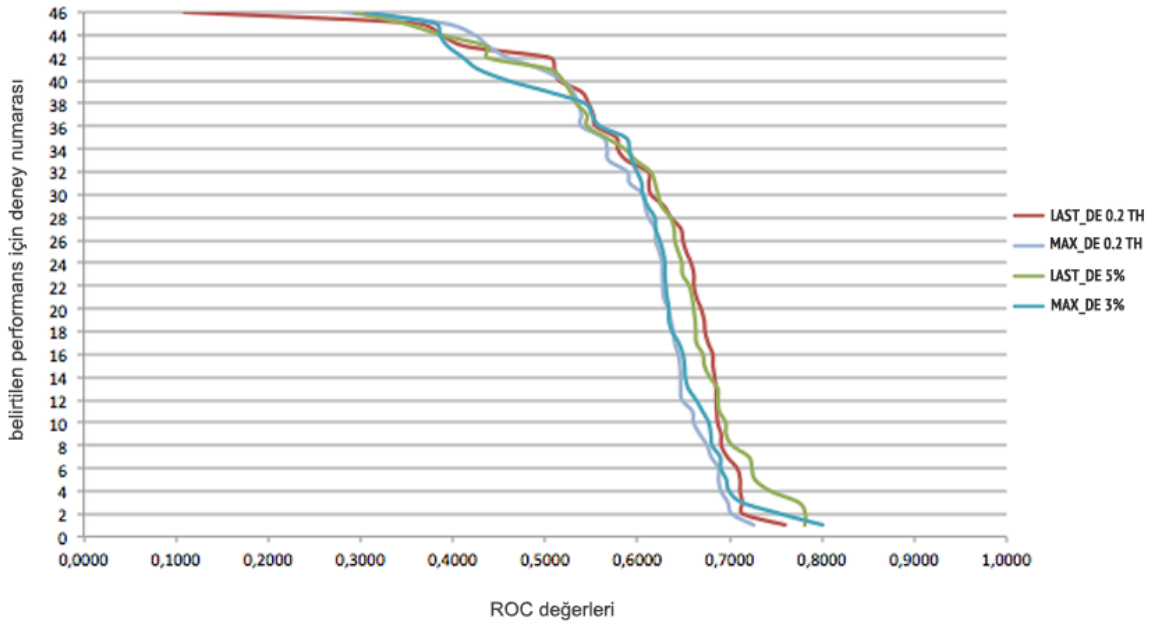


Şekil 4.2 Tanimoto Uzaklığının farklı parametreler ile birleşim gen listesi kullanılarak uygulanması sonucu elde edilen ROC sonuçları

Çizelge 4.4 Tanimoto Uzaklığının farklı parametreler ile birleşim gen listesi kullanılarak uygulanması sonucu elde edilen ortalama ROC sonuçları

Metot	**AUC Değeri**
LAST_DE 0.2 TH	0.629
MAX_DE 0.2 TH	0.621
LAST_DE 0.9%	0.644
MAX_DE 0.9%	0.614

Bu deneyde ise yukarıdaki deneyden farklı olarak birleşim yerine kesişim gen listesi kullanılmıştır. Şekil 4.3'te en iyi konfigürasyonlar ile uygulanmış bu metodun sonuç grafiği verilirken, Çizelge 4.5'te ortalama skorlar listelenmektedir. Bu deneydeki en iyi performans LAST_DE parmak izi verilerinin büyükten küçüğe doğru sıralanıp ilk 5%'nin farklı ifade olmuş olarak işaretlenmesiyle elde edilmiştir.



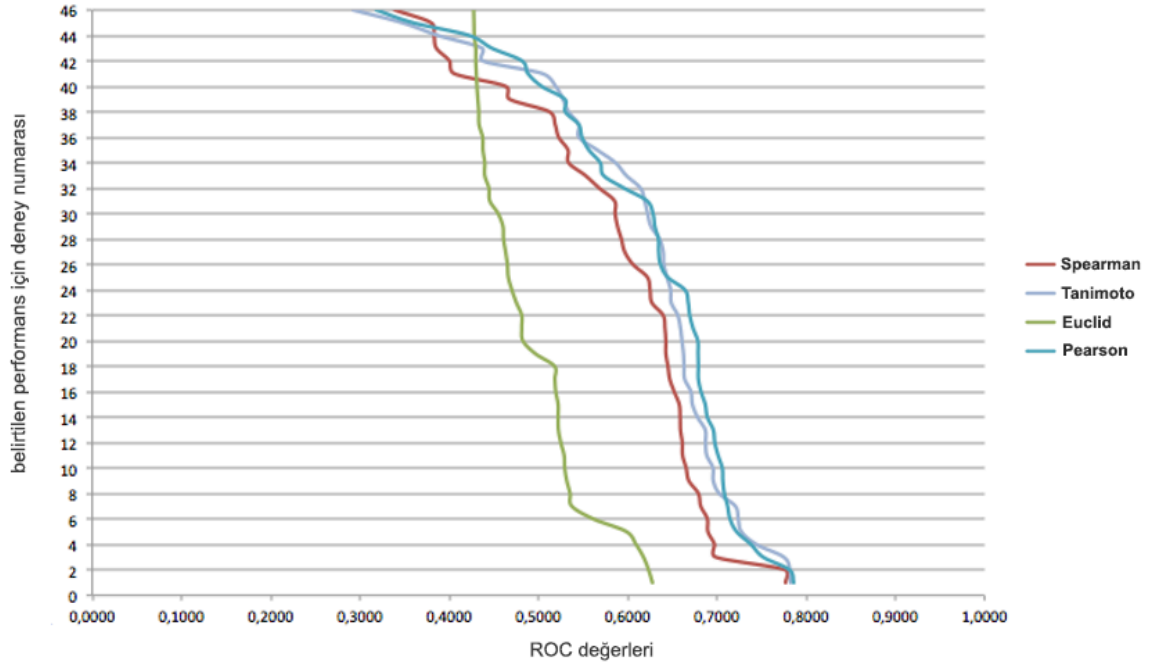
Şekil 4.3 Tanimoto Uzaklığının farklı parametreler ile kesişim gen listesi kullanılarak uygulanması sonucu elde edilen ROC sonuçları

Çizelge 4.5 Tanimoto Uzaklığının farklı parametreler ile kesişim gen listesi kullanılarak uygulanması sonucu elde edilen ortalama ROC sonuçları

Metot	**AUC Değeri**
LAST_DE 0.2 TH	0,615
MAX_DE 0.2 TH	0,598
LAST_DE 5%	0,621
MAX_DE 3%	0,602

Dört farklı benzerlik metriğinin farkı yöntemler ile vermiş olduğu en iyi sonuçların karşılaştırması Şekil 4.6'da gösterilmektedir. Bu deneyde geri getirim simülasyonu, veri kümesi içerisindeki tüm deneylerin ortak gen listesine sahip olan parmak izi verileri üzerinde çalıştırılmıştır. Ayrıca LAST_DE ve MAX_DE için ayrı ayrı yapılan deneyin sonuçları Şekil 4.4 ve Şekil 4.5'te altındaki mevcut karşılaştırma çizelgeleriyle birlikte gösterilmiştir (Çizelge 4.6 ve Çizelge 4.7). Aynı veri kümesi üzerinde farklı benzerlik metrikleri ile yapılan deney sonuçları Pearson Korelasyon

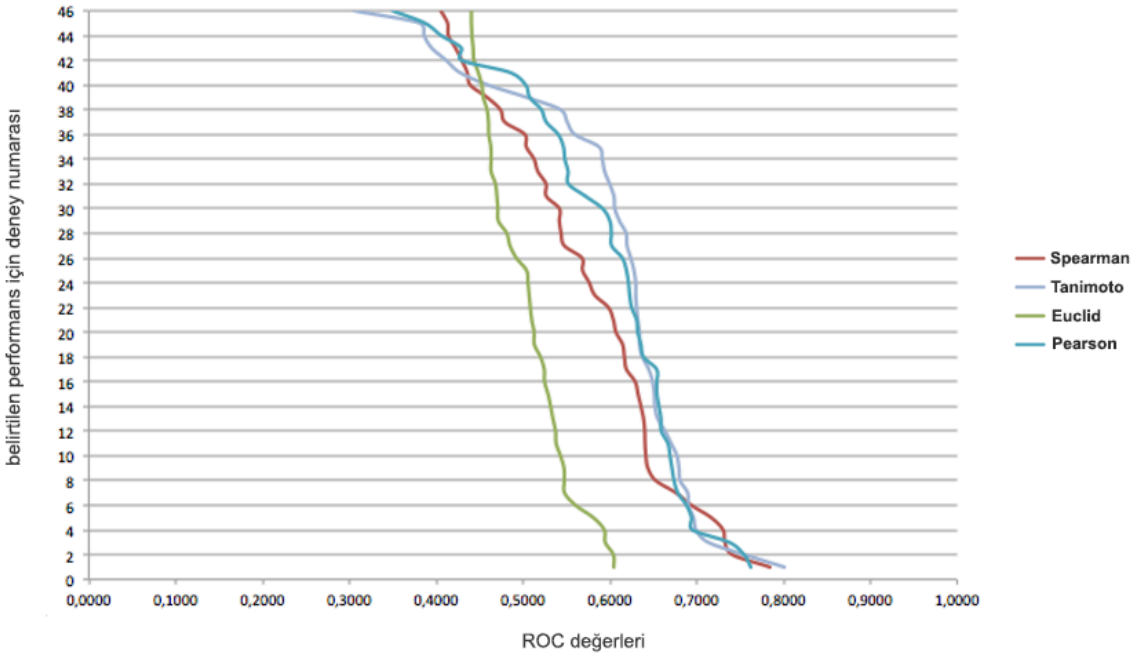
Katsayısı ve Tanimoto Uzaklığı'nın Öklid Uzaklığı'na göre yaklaşık %15, Spearman'ın Derece Bağıntı Katsayısı'na ise yaklaşık %4 daha iyi sonuç verdiğini göstermektedir. Ayrıca, Pearson Korelasyon Katsayısı bu deneyde Tanimoto Uzaklığı'na göre %0.4 daha iyi olduğu görülmektedir. Deneylerin en iyi sonuçları Çizelge 4.8'de verilmiştir.



Şekil 4.4 Farklı benzerlik metriklerinin kesişim gen listesi kullanılarak LAST_DE parmak izi verilerine uygulanması sonucu el edilen ROC sonuçları

Çizelge 4.6 Farklı benzerlik metriklerinin kesişim gen listesi kullanılarak LAST_DE parmak izi verilerine uygulanması sonucu el edilen ortalama ROC değerleri

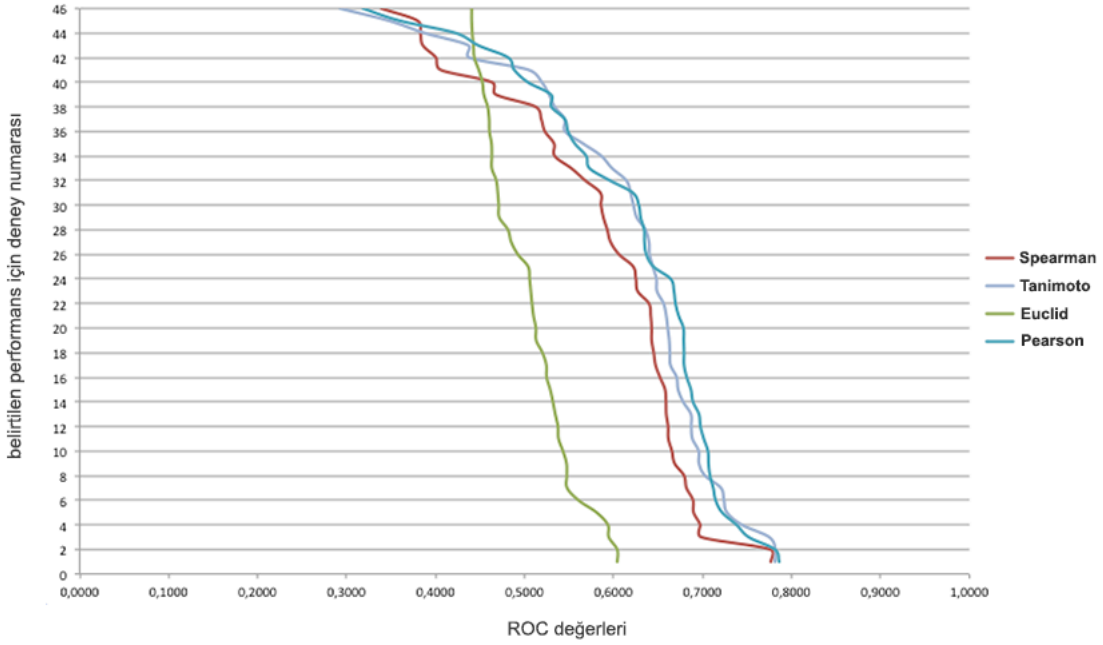
Metot	**AUC Değeri**
Öklid Uzaklığı	0,490
Spearman'ın Derece Bağıntı Katsayısı	0,591
Tanimoto Uzaklığı	0,621
Pearson Bağıntı Katsayısı	0,625



Şekil 4.5 Farklı benzerlik metriklerinin kesişim gen listesi kullanılarak MAX_DE parmak izi verilerine uygulanması sonucu el edilen ROC sonuçları

Çizelge 4.7 Farklı benzerlik metriklerinin kesişim gen listesi kullanılarak MAX_DE parmak izi verilerine uygulanması sonucu el edilen ortalama ROC değerleri

Metot	**AUC Değeri**
Öklid Uzaklığı	0,504
Spearman'ın Derece Bağıntı Katsayısı	0,574
Tanimoto Uzaklığı	0,602
Pearson Bağıntı Katsayısı	0,598



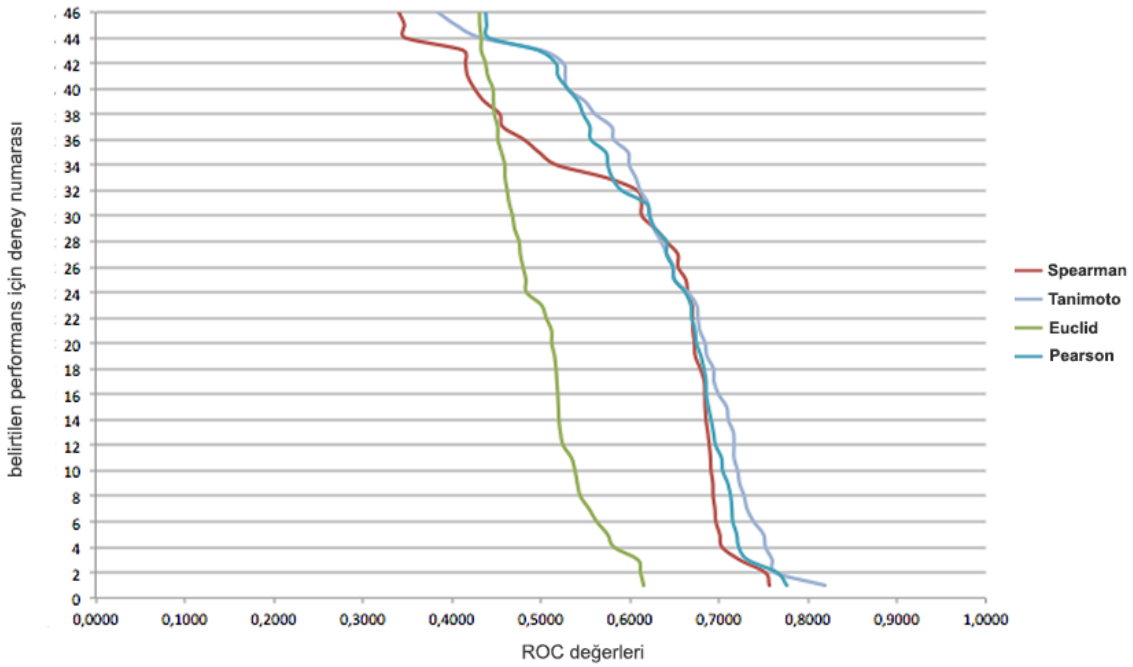
Şekil 4.6 Farklı benzerlik metriklerinin kesişim gen listesi kullanılarak parmak izi verilerine uygulanması sonucu elde edilen en iyi ROC sonuçları

Çizelge 4.8 Farklı benzerlik metriklerinin kesişim gen listesi kullanılarak parmak izi verilerine uygulanması sonucu elde edilen en iyi ortalama ROC değerleri

Metot	**AUC Değeri**
Öklid Uzalığı	0,504
Spearman'ın Derece Bağıntı Katsayısı	0,591
Tanimoto Uzaklığı	0,621
Pearson Bağıntı Katsayısı	0,625

Bir önceki deneyde dört benzerlik metriği kesişim gen listesi için uygulanmıştır. Bu deneyde ise aynı yöntemler birleşim gen listesi için uygulanmaktadır. Sonuçlar sırasıyla Şekil 4.7, Şekil 4.8 ve Şekil 4.9'da gösterilmektedir. Yine her şeklin altında ROC sonuçlarının AUC değerleri yer almaktadır (Çizelge 4.9, Çizelge 4.10 ve Çizelge 4.11). Birleşim gen listesi ile çalışılan bu deneyi bir önceki kesişim gen listesi ile çalışılan deney ile karşılaştırdığımızda büyük değişikliklerin olmadığı görülmektedir. ROC skorları diğerlerinden yüksek ve kendi aralarında birbirlerine

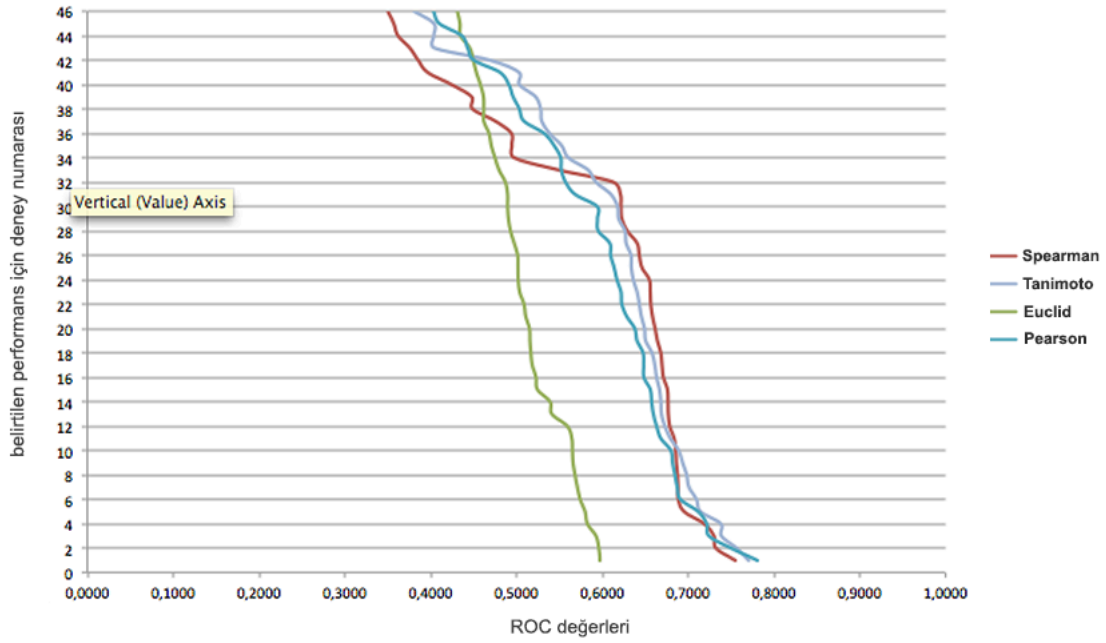
yakın olan Pearson Korelasyon Katsayısı ve Tanimoto Uzaklığı'nın yine diğer metriklere göre daha iyi sonuçlar verdiği ortaya çıkmıştır. Fakat bu sefer Tanimoto Uzaklığı ve Pearson Korelasyonu'na göre az da olsa daha iyi sonuçlar vermiştir.



Şekil 4.7 Farklı benzerlik metriklerinin birleşim gen listesi kullanılarak LAST_DE parmak izi verilerine uygulanması sonucu el edilen ROC sonuçları

Çizelge 4.9 Farklı benzerlik metriklerinin birleşim gen listesi kullanılarak LAST_DE parmak izi verilerine uygulanması sonucu el edilen ortalama ROC değerleri

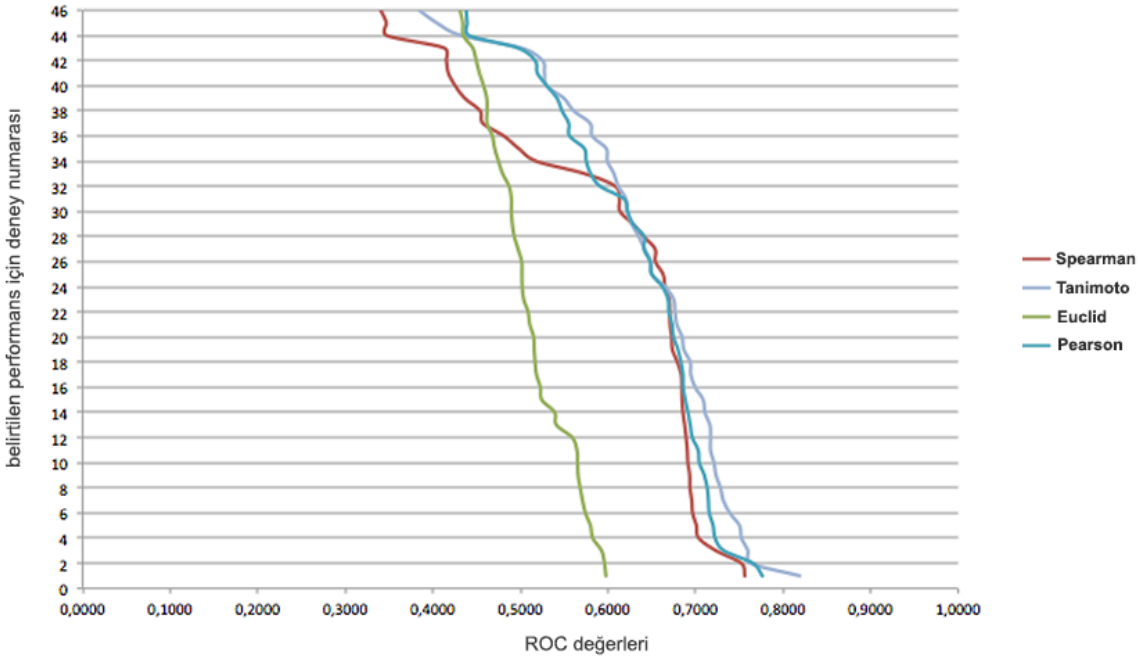
Metot	**AUC Değeri**
Öklid Uzaklığı	0,498
Spearman'ın Derece Bağıntı Katsayısı	0,604
Tanimoto Uzaklığı	0,644
Pearson Bağıntı Katsayısı	0,634



Şekil 4.8 Farklı benzerlik metriklerinin birleşim gen listesi kullanılarak MAX_DE parmak izi verilerine uygulanması sonucu el edilen ROC sonuçları

Çizelge 4.10 Farklı benzerlik metriklerinin birleşim gen listesi kullanılarak MAX_DE parmak izi verilerine uygulanması sonucu el edilen ortalama ROC değerleri

Metot	**AUC Değeri**
Öklid Uzaklığı	0,510
Spearman'ın Derece Bağıntı Katsayısı	0,598
Tanimoto Uzaklığı	0,614
Pearson Bağıntı Katsayısı	0,601



Şekil 4.9 Farklı benzerlik metriklerinin birleşim gen listesi kullanılarak parmak izi verilerine uygulanması sonucu el edilen en iyi ROC sonuçları

Çizelge 4.11 Farklı benzerlik metriklerinin birleşim gen listesi kullanılarak parmak izi verilerine uygulanması sonucu el edilen en iyi ortalama ROC değerleri

Metot	**AUC Değeri**
Öklid Uzalığı	0,510
Spearman'ın Derece Bağıntı Katsayısı	0,604
Tanimoto Uzaklığı	0,644
Pearson Bağıntı Katsayısı	0,634

Ayrıca en iyi sonuçları veren Pearson Korelasyon Katsayısı ve Tanimoto Uzaklığı metriklerinin benzerlik matrisindeki en iyi ROC skorunu vermiş 10 nokta aşağıdaki Çizelge 4.12 ve Çizelge 4.13'te listelenmektedir.

Çizelge 4.12 Pearson Korelasyon Katsayısı metriği ile oluşturulan benzerlik matrisindeki en yüksek 10 Benzerlik skoruna sahip nokta

No	**Deney X**	**Deney Y**	**Benzerlik Skoru**
1	GSE6462_HRG_1_0	GSE6462_HRG_10_0	0.856
2	GSE6751_36	GSE6751_35	0.855
3	GSE6462_HRG_1_0	GSE6462_HRG_0_5	0.823
4	GSE6521_HRG_U0126	GSE6521_HRG_AG1478	0.782
5	GSE6462_EGF_10_0	GSE6521_HRG_AG1478	0.754
6	GSE9936_Ad+EQ	GSE9936_Ad+E2	0.751
7	GSE6521_HRG_U0126	GSE6521_HRG	0.751
8	GSE6462_HRG_10_0	GSE6462_HRG_0_5	0.734
9	GSE6521_HRG_U0126	GSE6462_EGF_0_5	0.718
10	GSE6462_EGF_1_0	GSE6462_EGF_0_5	0.708

Çizelge 4.13 Tanimoto Uzaklığı metriği ile oluşturulan benzerlik matrisindeki en yüksek 10 ROC skora sahip nokta

No	Deney X	Deney Y	ROC Skoru
1	GSE9936_Ad+E2	GSE9936_Ad+EQ	0.662
2	GSE9936_Ad+HG	GSE9936_Ad+E2	0.594
3	GSE9936_Ad+HG	GSE9936_Ad+Q	0.573
4	GSE6461_HRG_1_0	GSE6461_HRG_10_0	0.552
5	GSE6751_36	GSE6751_35	0.522
6	GSE9936_AdERb+EQ	GSE9936_AdERb+E2	0.475
7	GSE9105_S10	GSE9105_S8	0.396
8	GSE6462_HRG_10_0	GSE6462_HRG_0_5	0.372
9	GSE5808_p3	GSE5808_p1	0.364
10	GSE9105_S4	GSE9105_S10	0.364

Çizelge 4.12 ve Çizelge 4.13'e bakıldığında bu çalışmada önerilen zaman serisi deneylerde içerik tabanlı arama yöntemlerinin iyi sonuçlar verdiği anlaşılmaktadır. Deneylerin isimlerine bakıldığında örneğin GSE6462_HRG_1_0 ilk alt çizgiye kadar olan bölüm deney serisi numarası sonraki bölüm deney tasarımında kullanılan etken madde/maddeler veya dozunu ifade etmektedir. Örneğin Çizelge 4.12'de benzerlik skorlarına göre sıralanmış en iyi 10 deneyden 8'inin zaten aynı serinin farklı hormon veya inhibitörler gibi maddeler eşliğinde yapıldığı anlaşılmaktadır. İçerik tabanlı aramada amaç benzer deneyleri bulmak olduğundan bu geri getirilen deneylerin benzerlik olarak çok yüksek skorlara sahip olması deneyimizin önerdiğimiz model ve yöntem açısından başarılı sonuçlar verdiğini göstermektedir. Ayrıca 9 ve 5 nolu deneylere baktığımızda zaman serisi deneyin farklı GEO Serisine ait olduğu fakat her iki deneyde platform, organizma ve deney tasarımının hemen hemen aynı olduğu görülmektedir. Bu durum ise deneylerin yakınlık derecelerinin neden yüksek çıktığını açıklamaktadır.

5. SONUÇLAR VE TARTIŞMA

Bu çalışmada zaman serisi mikrodizilerde içerik tabanlı aramanın uygulanabilirliği deneylerle gösterilerek ve yorumlanarak anlatılmaya çalışılmıştır. Özet olarak, zaman serisi çok boyutlu ifade verileri için farklı parmak izi çıkarma çalışmaları ve benzerlik metriği stratejileri karşılaştırılmıştır. Bizim oluşturduğumuz veri tabanı için, sonuçlar Pearson Bağıntı Katsayısı ve Tanimoto Uzaklığı'nın farkı ifadeye dayalı parmak izlerinin karşılaştırılmasında daha iyi olduğunu göstermektedir. Ayrıca, zaman serisi deneylerde farklı ifade olmuş genlerin tespitinde ilk ve son zaman noktalarının alınması ilk ve farklı ifade olma olasılığı en yüksek olan noktaya göre daha iyi sonuçlar verdiği gözlemlenmiştir.

Sonuçlar, tüm genlerin zaman serileri ifade davranışlarını tanımlamak için bir kerede tüm zaman noktalarının değerlendirilmesine olanak sağlamaktadır. Aynı zamanda, değerlendirme kriterleri tartışmaya açık bir konudur. Bazı durumlarda, diğer hastalıklar için de aynı tedavinin bulunması, aynı hastalıklar ile ilişkili deneylerin bulunmasından daha fazla arzulanabilir. Bu çalışma büyük veri depolarından zaman serisi deneylerin bulunması üzerine bilgi geri getirimi ve biyoenformatik alanlarında çalışan araştırmacılar için önemli bir tartışma açmaktadır.

KAYNAKLAR LİSTESİ

[1] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, and A Soboleva, "NCBI GEO: archive for functional genomics data sets—update," Nucleic Acids Res., 2013, D991-5.LOCKHART, D. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays, Nat.Biotechnol, vol.14, s.1675–1680, 1996.

[2] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma, "ArrayExpress—a public database of microarray experiments and gene expression profiles," Nucleic Acids Res., 2007, 35: D747–D750.

[3] L. Hunter, R.C. Taylar, S.M. Leach, and R. Simon, "GEST: a gene expression search tool based on a novel Bayesian similarity metric," Bioinformatics, vol. 17, 2001, pp. S115-S122.

[4] A. Tanay, I. Steingeld, M. Kupiec, and R. Shamir, "Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium," Mol. Syst. Biol., 2005, 1: 2005.0002.

[5] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J. Lerner, J.P. Brunet, A. Subramanian, K.N. Ross, M. Reich, H. Hieronymus, G. Wei, S.A. Armstrong, S.J. Haggarty, P.A. Clemons, R. Wei, S.A. Carr, E.S. Lander, and T.R. Golub, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," Science, 313(5795):1929-35, 2006.

[6] M.A. Hibbs, D.C. Hess, C.L. Myers, , C. Huttenhower, K. Li, and O.G. Troyanskaya, "Exploring the functional landscape of gene expression: directed search of large microarray compendia," Bioinformatics 23, 2692–2699, 2007.

[7] D.C. Hassane, M.L. Guzman, C. Corbett, X. Li, R. Abboud, F. Young, J.L. Liesveld, M. Carroll, and C.T. Jordan, "Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data," Blood, 111(12):5654-62, 2008.

[8] J.T. Dudley, R. Tibshirani, T. Deshpande, and A.J. Butte, "Disease signatures are robust across tissues and experiments," Mol. Sys. Biol., 2009, 5:307.

[9] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, , J. Lerner, J.P. Brunet, A. Subramanian, K.N. Ross, M. Reich, H. Hieronymus, G. Wei, S.A. Armstrong, S.J. Haggarty, P.A. Clemons, R. Wei, S.A. Carr, E.S. Lander, and T.R. Golub, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," Science, 313(5795):1929-35, 2006.

[10] P.B. Horton, L. Kiseleva, and W. Fujibuchi, “RaPiDS: an algorithm for rapid expression profile database search,” International Conference on Genome Informatics, vol. 17, pp. 67-76, 2006.

[11] W. Fujibuchi, L. Kiseleva, T. Taniguchi, H. Harada, and P. Horton, “CellMontage: similar expression profile search server,” Bioinformatics, vol. 23, pp. 3103-3104, 2007.

[12] R. Chen, R. Mallelwar, A. Thosar, S. Venkatasubrahmanyam, and A.J. Butte, “GeneChaser: Identifying all biological and clinical conditions in which genes of interest are differentially expressed,” BMC Bioinformatics, vol. 9, p. 548, 2008.

[13] C. Feng, M. Araki, R. Kunimoto, A. Tamon, H. Makiguchi, S. Niijima, G. Tsujimoto, Y. Okuno, “GEM-TREND: a web tool for gene expression data mining toward relevant network discovery,” BMC Genomics, 2009, 10:411.

[14] A.C. Gower, A. Spira, and M.E. Lenburg, “Discovering biological connections between experimental conditions based on common patterns of differential gene expression,” BMC Bioinformatics, 2011, 12: 381.

[15] G. Williams, “SPIEDw: a searchable platform-independent expression database web tool,” BMC Genomics, 2013, 14:765.

[16] J.M. Engreitz, A.A. Morgan, J.T. Dudley, R. Chen, R. Thathoo, R.B. Altman, and A.J. Butte, “Content-based microarray search using differential expression profiles,” BMC Bioinformatics, 2010, 11:603.

[17] J.M. Engreitz, R. Chen, A.A. Morgan, J.T. Dudley, R. Mallelwar, and A.J. Butte, “ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression,” Bioinformatics, vol. 27, pp. 3317-3318, 2011.

[18] F. Bell, and A. Sacan, “Content based searching of gene expression databases using binary fingerprints of differential expression profiles,” Health Informatics and Bioinformatics (HIBIT) 7th International Symposium, pp. 107-113, 2012.

[19] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski, “Probabilistic retrieval and visualization of biologically relevant microarray experiments,” Bioinformatics, 2009, 25:i145-153.

[20] S. Suthram, J.T. Dudley, A.P. Chiang, R. Chen, T.J. Hastie, and A.J. Butte “Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets,” PLoS Comput Biol, 2010, 6(2):e1000662.

[21] E. Georgii, J. Salojärvi, M. Brosché, J. Kangasjärvi, and S. Kaski, “Targeted retrieval of gene expression measurements using regulatory models,” Bioinformatics, 2012, 28:2349-2356.

[22] http://www.genetiklab.com/, 08.06.14.

[23] Tian, Tianhai, "Stochastic Models for Studying the Degradation of mRNA Molecules," Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on , vol., no., pp.167,172, 12-15 Nov. 2011.

[24] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux,Dmitry Rudnev, et al. NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucleic Acids Research, vol. 35, pp.760-765, 2007.

[25] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Research, Vol. 37, pp.885-890, 2009.

[26] Ron Edgar and Tanya Barrett.NCBI GEO standards and services for microarray data.Nat Biotechnol, vol. 24(12), pp.1471-1472, 2006.

[27] Tanya Barrett and Ron Edgar.Gene Expression Omnibus (GEO): Microarray data storage, submission, retrieval, and analysis. Methods Enzymol, vol.411,pp. 352-369, 2006.

[28] J.M. Engreitz , A. A. Morgan, J.T. Dudley, R.Chen, R.Thathoo, R.B. Altman, A. J. Butte “Content-based microarray search using differential expression profiles”. BMC Bioinformatics 2010 11:603.

[29] F.Bell, A.Sacan “Content based searching of gene expression databases using binary fingerprints of differential expression profiles” Health Informatics and Bioinformatics (HIBIT) 2012 7th International Symposium. Pages 107-113.

[30] Hofmann, T., (2001), Unsupervised Learning by Probabilistic Latent Semantic Analysis, Machine Learning, 42, 177-196.

[31] Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. (2006), ‘Microarray data anal- ysis: from disarray to consolidation to consensus’, Nature Reviews: Genetics 7, 55–65.

[32] Jeffery, I. B., Higgins, D. G. & Culhane, A. C. (2006), ‘Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data’, BMC Bioinformatics 7, 359.

[33] Witten, M. D. and Tibshirani R. (2007, Kasım). A comparison of fold-change and the t-statistic for microarray data analysis. Retrieved July 1, 2014 from Standford University, http://statweb.stanford.edu/~tibs/ftp/FCTComparison.pdf.

[34] Z. Bar-Joseph. Analyzing time series gene expression data. Bioinformatics, 20(16):2493–2503, 2004.

[35] R. B. Stoughton. Applications of DNA microarrays in biology. Annu Rev Biochem, 74:53–82, 2005.

[36] I. P. Androulakis, E. Yang, and R. R. Almon. Analysis of time-series gene expression data: Methods, challenges, and opportunities. Annu Rev Biomed Eng, 9:205–228, 2007.

[37] S. D. Ginsberg, S. E. Hemby, S. E. Lee, V. M. Lee, and J. H. Eberwine. Expres- sion profiling of transcripts in Alzheimer's disease tangle-bearing CA1 neurons. Ann Neurol, 48:77–87, 2000.

[38] J. M. Ross, C. Fan, M. D. Ross, T.-H. Chu, Y. Shi, L. Kaufman, W. Zhang, M. E. Klotman, and P. E. Klotman. HIV- 1 infection initiatites an inflammatory cascade in human renal tubular epithe- lial cells. J Acquir Immune Defic Syndr, 42(1):1–11, 2006.

[39] M. L. Whitfield, G. Sherlock, A. J. Sal- danha, J. I. Murray, C. A. Ball, K. E. Alexander, J. C. Matese, C. M. Perou, M. M. Hurt, P. O. Brown, and D. Bosteon. Identification of genes periodically ex- pressed in the human cell cycle and their expression in tumors. Mol Biol Cell, 13(6):1977–2000, 2002.

[40] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell, 11(12):4241–4257, 2000.

[41] Androulakis IP, Yang E, Almon RR: Analysis of time-series gene expression data: Methods, challenges, and opportunities. Annual Review of Biomedical Engineering 2007, 9:205-228.

[42] Bar-Joseph Z: Analyzing time series gene expression data. Bio- informatics (Oxford, England) 2004, 20(16):2493-2503.

[43] Opgen-Rhein R, Strimmer K: Learning causal networks from sys- tems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC bioin- formatics 2007, 8 Suppl 2:S3.

[44] Opgen-Rhein R, Strimmer K: From correlation to causation net- works: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. Bmc Syst Biol 2007, 1:37.

[45] Ernst J, Bar-Joseph Z: STEM: a tool for the analysis of short time series gene expression data. BMC bioinformatics 2006, 7:191.

[46] Ding M, Cui SY, Li CJ, Jothy S, Haase V, Steer BM, Marsden PA, Pippin J, Shankland S, Rastaldi MP, Cohen CD, Kretzler M, Quaggin SE: Loss of the tumor suppressor Vhlh leads to upregulation of Cxcr4 and rapidly progressive glomerulonephritis in mice. Nat Med 2006, 12(9):1081-1087.

[47] Karpuj MV, Becher MW, Springer JE, Chabas D, Youssef S, Pedotti R, Mitchell D, Steinman L: Prolonged survival and decreased abnormal movements in transgenic model of Huntington disease, with administration of the transglutaminase inhibi- tor cystamine. Nat Med 2002, 8(2):143-149.

[48] Braga-Neto U: Fads and fallacies in the name of small-sample microarray classification. Ieee Signal Proc Mag 2007, 24(1):91-99.

[49] Ernst J, Nau GJ, Bar-Joseph Z: Clustering short time series gene expression data. Bioinformatics (Oxford, England) 2005, 21:I159-I168.

[50] N. Dean and A.E. Raftery, “Normal uniform mixture differential gene expression detection for cDNA microarrays,” BMC Bioinformatics, 2005, 6:173.

[51] Sahiner, B.; Heang-Ping Chan; Hadjiiski, L.M., "Performance Analysis of Three-Class Classifiers: Properties of a 3-D ROC Surface and the Normalized Volume Under the Surface for the Ideal Observer," Medical Imaging, IEEE Transactions on , vol.27, no.2, pp.215,227, Feb. 2008.

[52] Landgrebe, T.C.W.; Duin, R. P W, "Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.30, no.5, pp.810,822, May 2008.

[53] http://en.wikipedia.org/wiki/Receiver_operating_characteristic, 12.06.2014.

[54] Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.