EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

(DOKTORA TEZİ)

MİKRODİZİ GEN İFADE VERİTABANLARINDA İÇERİK-TABANLI ARAMA

Esma ERGÜNER ÖZKOÇ

Tez Danışmanı: Prof. Dr. Mehmet Emin DALKILIÇ

İkinci Danışman: Prof.Dr. Hasan OĞUL

Uluslararası Bilgisayar Anabilim Dalı

Sunuş Tarihi: 16.06.2016

Bornova-İZMİR
2016

KABUL VE ONAY SAYFASI

Esma ERGÜNER ÖZKOÇ tarafından doktora tezi olarak sunulan "Mikrodizi Gen İfade Veritabanlarında İçerik-Tabanlı Arama" başlıklı bu çalışma EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliği ile EÜ Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve ..... tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri Üyeleri:		**İmza**
Jüri Başkanı	**:.....**	**.....**
Raportör Üye	**: .....**	**.....**
Üye	**: .....**	**.....**
Üye	**: .....**	**.....**
Üye	**: .....**	**.....**

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Doktora Tezi olarak sunduğum "Mikrodizi Gen İfade Veritabanlarında İçerik-Tabanlı Arama" başlıklı bu tezin kendi çalışmam olduğunu, sunduğum tüm sonuç, doküman, bilgi ve belgeleri bizzat ve bu tez çalışması kapsamında elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara atıf yaptığımı ve bunları kaynaklar listesinde usulüne uygun olarak verdiğimi, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, bu tezin herhangi bir bölümünü bu üniversite veya diğer bir üniversitede başka bir tez çalışması içinde sunmadığımı, bu tezin planlanmasından yazımına kadar bütün safhalarda bilimsel etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim.

.... / .... / 2016

İmzası

Adı-Soyadı

Esma ERGÜNER ÖZKOÇ

ÖZET

MİKRODİZİ GEN İFADE VERİTABANLARINDA İÇERİK-TABANLI ARAMA

Ergüner Özkoç, Esma

Doktora Tezi, Uluslararası Bilgisayar Anabilim Dalı
Tez Danışmanı: Prof. Dr. Mehmet Emin DALKILIÇ

İkinci Danışman: Prof.Dr. Hasan OĞUL

16.06.2016, 106 sayfa

Gen ifade veritabanlarının çok hızlı büyümesi anahtar veya metadata (üstveri) kullanan yapılandırılmamış veritabanı sorgulama işlemlerine alternatif olarak içerik tabanlı arama ihtiyacını getirmiştir. İçerik-tabanlı arama, benzer gen ifade desenine sahip bütün deneylerin, veritabanlarında bulunan biyolojik metinsel açıklamalarına bakılmaksızın getirilmesidir.

Bu tez çalışmasında, içerik-tabanlı arama için gen ifade veritabanlarının özel bir alt kümesi olan zaman serisi deneyleri üzerinde odaklanılmış ve zaman serisi deneyleri sorgulanmıştır. Zaman serisi deneyinin tamamı sorgu deney olarak alınarak, ifade profillerine kümeleme algoritmaları ile boyut indirgeme yapılmış ve halka açık veritabanından oluşturulan test veri kümesinde içerik tabanlı arama uygulaması yapılmıştır. En uygun kümeleme algoritmasının bulunması için FunFem, Mclust ve Kmeans kümeleme algoritmaları denenmiş ve farklı ifade olmuş gen tabanlı yöntem ile karşılaştırılmaları yapılmıştır.

Herhangi iki deneyin benzerliğinin bulunmasında deneylerde bir kümeye atanan ortak gen sayısı ile hastalık ilişkileri açıklamasına (annotation) göre deneylerin benzerliği hesaplanmıştır.

Sonuçlar kümeleme kullanan yöntemin, geleneksel farklı ifade olmuş gen tabanlı yöntemden daha iyi geri getirim yaptığını göstermiştir.

Anahtar sözcükler: Gen ifade veritabanı, zaman serisi verisi, zaman serisi profili içerik-tabanlı arama, bilgi geri getirimi, model-tabanlı kümeleme.

ABSTRACT

CONTENT-BASED SEARCH ON MICROARRAY GENE EXPRESSION DATABASES

ERGÜNER ÖZKOÇ, Esma

PhD in International Computer Department

Supervisor: Prof. Dr. Mehmet Emin DALKILIÇ

Co-Supervisor: Prof.Dr. Hasan OĞUL

06.16.2016, 106 pages

The rapid growth of gene expression databases has created a need for content-based searches as an alternative to unstructured database queries using keyword- or metadata-based searches. Content-based searching is the ability to retrieve all experiments with similar gene expression patterns in a database regardless of the biological annotations provided for these experiments.

While this concept is still in its infancy in a general context, in this thesis we focus on applying it to a specific subset of gene expression datasets, by only querying experiments involving time-series expression profiles. To this end, we propose a novel experiment fingerprinting scheme obtained by clustering expression profiles, for content-based searching of time-series microarray experiments.

To determine the retrieval ability of the proposed scheme, we performed a simulated information retrieval task on a large set of microarray experiments gathered from a public repository. The relevance between any two experiments was then defined using their commonalities based on annotated disease associations.

The results showed that relevant experiments can be more successfully retrieved using this new method compared with traditional differential expression-based methods.

Keywords: Gene expression database, time-course data, time-series profile, content-based search, information retrieval, model-based clustering.

TEŞEKKÜR

Bu çalışma süresince bilgi, deneyim ve önerilerini esirgemeyen, beni araştırma ve geliştirmeye yönlendiren değerli hocalarım Prof. Dr. Mehmet Emin Dalkılıç ve Prof. Dr. Hasan Oğul'a,

Bu süreçte bana katlanan ve desteklerini her fırsatta gösteren canım eşim Önder Özkoç ve kardeşim Duygu Ölmez'e,

Varlığıyla yüzümü güldüren biricik oğlum Ömer Özkoç'a

Yürüttüğümüz ortak çalışmalarda verdiği destekler için değerli arkadaşım Ahmet Hayran'a

Ve destekleriyle beni yüreklendiren arkadaşlarıma en içten teşekkürlerimi sunarım.

Annem Fatma Ergüner, Babam Temel Ergüner, Ablam Saadet Ergüner ve Kardeşim Seda Ergüner'in aziz hatırasına...

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (Devam)

Sayfa

İÇİNDEKİLER (Devam)

Sayfa

ŞEKİLLER DİZİNİ

Şekil Sayfa

ŞEKİLLER DİZİNİ (Devam)

ŞEKİLLER DİZİNİ (Devam)

ÇİZELGELER DİZİNİ

ÇİZELGELER DİZİNİ (Devam)

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

AP	Averaj duyarlılık (Average Precision)
AUC	Eğrinin altında kalan alan (Area Under Curve)
cDNA	Bütünleyici DNA
DNA	Deoksiribonükleik asit
EM	Beklenti Yükseltme (Expectation Maximization)
FİO	Farklı İfade Olmuş
GEO	Gene Expression Omnibus
GSE	GEO veri serisi
GSEA	Zenginleştirilmiş gen kümesi analizi (Gene Set Enrichment Analysis)
ICA	Bağımsız Bileşen Analizi (Independent Component Analysis)
MAP	Ortalama Averaj Duyarlılık (Mean Average Precision)
mRNA	Mesajcı RNA
ROC	İşletim Karakteristik Eğrisi
RNA	Ribonükleik asit
tp	Zaman noktası

1. GİRİŞ

Bilgisayar teknolojisi moleküler mikrobiyolojik çalışmalarla paralel olarak dikkate değer bir değişim göstermektedir. Önceki çalışmalar tek bir genin analizi konusunda yoğunlaşmışken son yıllarda bir çalışmanın bütün genomun analizini yaptığı teknolojiler geliştirilmiştir.

Moleküler biyolojideki geleneksel yöntemlerde genel olarak bir gen analizi bir deneyde yapılmaktadır ve gen fonksiyonlarını bütün olarak görmek oldukça zordur. Geleneksel yöntemlerden farklı olarak DNA mikrodizi teknolojisinin popülaritesinin sebebi, araştırmacıların tek seferde binlerce genin birbirleriyle olan etkileşimlerini görmesini ve bütün genomun basit bir çip üzerinde görüntülenmesini sağlamasıdır. DNA mikrodizi deneyleri gen ifade analizi, antimikrobiyal gen analizi, çevresel araştırmalar, genetik ve mutasyon analizleri gibi birçok teorik ve deneysel çalışmaların temelini oluşturmakta ve hastalıklara tanı koyulmasında bir araç olarak kullanılmaktadır. Araştırmacılar yapılan mikrodizi deneylerini, güncel çalışmaları takip etmek, yeni hipotezler geliştirmek gibi amaçlarla veritabanlarına kaydetmektedir. Çoğunluğu halka açık olan bu veritabanları araştırmacılar için oldukça değerlidir ancak veritabanlarında hali hazırda kullanılan arama fonksiyonları araştırmacıların benzer deneyleri aramasını yapabilecek yeterlilikte değildir. Manuel olarak veritabanında arama yapmak oldukça zahmetli ve zaman harcayıcıdır. Mikrodizi veritabanlarında arama için ortak yaklaşım: metinsel açıklama, dizi ve genlerin tanımlamaları gibi metaveri (üstveri) tabanlıdır. Metaveri tabanlı aramalar, tamamlanmamış kategorizasyon ve örneklerin eksik/yanlış etiketlenmesi gibi nedenlerle sorgu ile ilgili tüm veri setlerini getirememektedir. Bu aramalar, aranan anahtar kelime ile ilgili deneyleri bulmak için verimlidir fakat benzer sonuçlar vermiş çalışmaları bulamaz. Bunun yerine bir deney sonucunda alınan ifade ölçümlerinin tamamı sorgu olarak kullanılarak veritabanındaki benzer sonuçları almış diğer deneyler aranabilir. Ancak gen ifade veri tabanlarında henüz içerik tabanlı arama yapan bir fonksiyonu bulunmamaktadır. Mikrodizi deneyleri statik ve zaman serileri (temporal) olarak iki ana bölümde incelenebilir. Statik deneylerde gen ifadesi ölçümü örneklerin her birinden sadece bir kez alınarak yapılmaktadır. Statik mikrodizi deneyleri genellikle, hasta ve sağlam dokudan alınan örnekler üzerinde genlerin ifade seviyeleri ölçülerek, hastalık mekanizması üzerine yapılmaktadır. Zaman serisi deneylerinde ise tek örneğin farklı zamanlardaki ifade seviyeleri ölçülmektedir.

Bu tez çalışmasında, gen ifade veritabanında benzer sonuçları almış deneyleri getirmek için kümeleme tabanlı içerik-tabanlı arama yöntemi geliştirilmiştir. Yöntemde zaman serisi deneyinin tamamı sorgu olarak alınmış ve oluşturulan test veriseti üzerinde içerik-tabanlı arama yapılmıştır.

Bu tez raporu dört bölümden oluşmaktadır. İlk bölümde çalışma alanıyla ilgili temel bilgiler, bu alanda yapılmış önceki çalışmalar ve tezin katkısı yer almaktadır. İkinci bölüm içerik tabanlı arama yöntemi için uygulanan yöntemler ve veriler ile ilgili detaylı bilgi sunulmaktadır. Üçüncü bölümde test verileri üzerinde yapılan uygulama sonuçları anlatılmakta ve son bölümde tez çalışmasının sonucu verilmektedir.

1.1 Alan Bilgisi

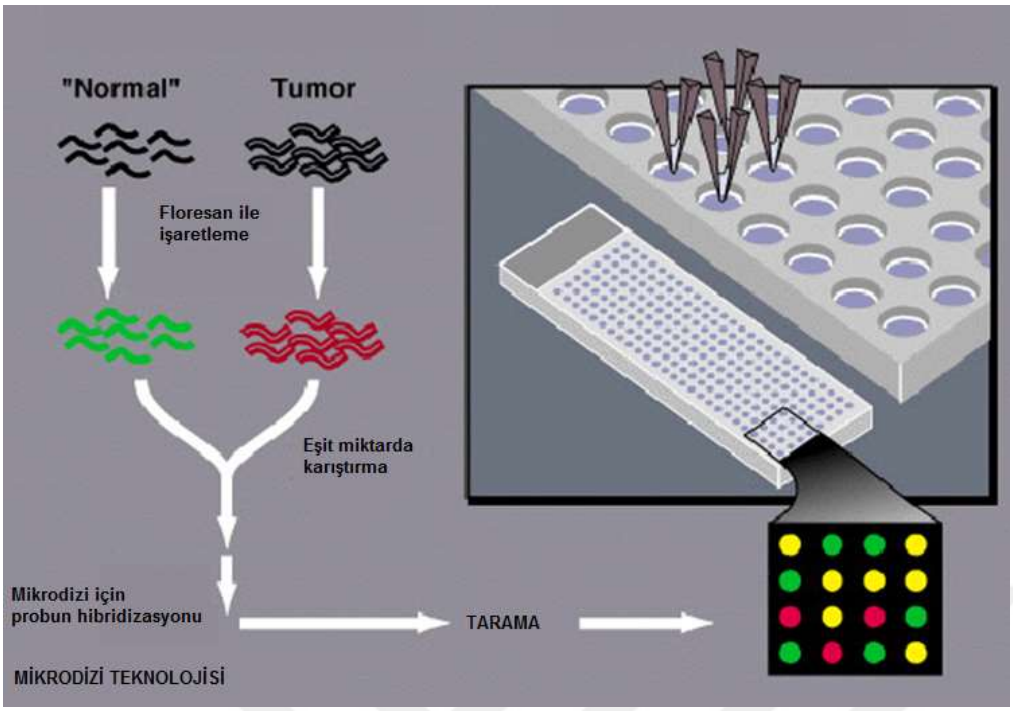
1.1.1 DNA Mikrodizi

Bütün organizmalar ve çeşitli virüslerin yaşam fonksiyonları ve biyolojik gelişmeleri için gerekli olan genetik yönergelri taşıyan nükleik asite Deoksiribonükleik asit (DNA) denir. DNA'nın temel işlevi, RNA(Ribonükleik asit) ve Protein gibi hücrenin diğer tamamlayıcı parçalarının oluşumu için gerekli olan bilgileri uzun süreli saklamaktır. Genetik bilgileri içeren DNA parçalarına gen denir. Bu sebeple DNA bir kalıp veya şablona benzetilebilir. DNA direkt protein üretmek yerine protein sentezini kodlaması için RNA teli şeklinde bir şablon üretir. Bilgi, RNA molekülleri sayesinde DNA moleküllerinden proteinlere gönderilir. DNA molekülünden RNA molekülü sentezlenmesine yazım (transkripsiyon) denilmektedir (Stormo, 2000). mRNA (Mesajcı RNA), DNA'da ki gizli genetik bilginin, protein yapısına transfer edilmesinde şablon görevi yapan aracı moleküldür. mRNA yapılarındaki değişikliklerle hücrelerdeki hemen hemen tüm değişimler açıklanabilmektedir. mRNA ve kodlanmış proteinler gibi gen ürünlerinin üretilmesine Gen İfadesi (Gene Expression) denir. Örneğin mayalarda en az 1000 farklı gen işleviyle spor oluşumu gerçekleşmektedir. Ayrıca çevresel faktörlerin değişimleri de hücrelerin gen ifadelerinde farklılaşmalara sebep olmaktadır. Genlerin ifade desenlerini bilmek, gen fonksiyonlarını anlamlandırabilmek için önemlidir. Genomdaki binlerce-on binlerce genin ifade değişimlerini analiz ederek gen ve gen ürünleri arasındaki etkileşim hakkında fikir sahibi olunabilir. DNA mikrodizi, hücre ve dokulardaki gen ifade ölçümlerindeki bütünsel değişikliklerin araştırılmasında kullanılan bir teknolojidir. DNA mikrodizi teknolojisinde, yazımın genom ölçeğinde değerlendirilmesi, düzenli bir şekilde tüm genomu, içeren cam slaytlar ve ince katman camlar ile sağlanmıştır.

DNA mikrodizi (DNA çip veya biyoçip) cam, silikon çip veya plastik gibi katı bir yüzeye iliştirilerek sıralı bir şekilde birleştirilmiş DNA noktalarıdır. Her DNA noktası, prob adı verilen, belirli DNA diziliminden pikomol (10-12 mol) içerir. Bu genin kısa bir bölümü veya hibridize için kullanılan DNA elemanı olabilir.

Verilen hücre içerisinde genlerin ifade seviyelerini ölçmek için, araştırmacılar ilk olarak bu hücre içerisinde bulunan mRNA'ları toplarlar. Toplanan mRNA'lar ters transkriptaz enzimler (RT) aracılığıyla etiketlenir. Bu enzimler mRNA için bütünleyici cDNA üretirler. Genellikle nükleotidler Cy5 (kırmızı) ve Cy3 (yeşil) gibi floresan işaretlerle ya da kimyasal luminescent saptama ile fark edilebilien digoksijenin (DIG) ile etiketlenmiştir. Bu etiketleme sürecinde floresan nükleotidler (fluorescent nucleodites) cDNA'ya bağlanırlar. Etiketlenmiş olan cDNA'lar DNA mikrodizi üzerine yerleştirirler. Hücre içerisinden mRNA'ları temsil eden etiketlenmiş cDNA'lar mikrodizi üzerinde bulunan yapay cDNA'ları ile hibritleşir/bağlanır. Araştırmacılar özel bir tarayıcı ile mikrodizi üzerindeki noktaları tarayarak floresan yoğunluğunu ölçerler. Eğer çok parlak floresan noktaları var ise bu ilgili genin çok fazla mRNA molekülü ürettiği ve aktif olduğu şeklinde, eğer kısık floresan noktaları var ise bu daha az işaretlenmiş cDNA olduğu ve ilgili genin pasif olduğu şeklinde yorumlanır. Eğer floresan yok ise bu hiçbir mRNA'nın DNA ile hibritleşmediği anlamına gelmektedir ve genin pasif olduğunu göstermektedir (Yoltas ve Karaboz, 2010; Schena et al., 1995).

Şekil 1.1de tümör, kırmızı floresan ile normal hücre ise yeşil floresan ile etiketlenmiştir. Örnekler birlikte hibritleme safhasında mikrodizi üzerindeki yapay cDNA'lar için yarışırlar. Sonuç olarak, eğer nokta kırmızı 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 1.1 Mikrodizi Teknolojisi (NHGRI, 2016)

Mikrodizi teknolojisi için kullanılan çipler/platformlar günümüzde çok çeşitli firmalar tarafından üretilmektedir. Bu firmalar araştırmacıların istedikleri çipleri de tasarlayabilmektedir. Affymetrix, Illumina, Agilent ve Exiqon bu firmalardan en bilinenler arasında sayılabilir.

DNA mikrodizi teknolojisi bir genomdaki gen kopyalarının belirlenmesinde, nükleotid polimorfizmi ve mutasyonun belirlenmesinde, yeni ilaçların keşfedilmesinde ve geliştirilmesinde geleneksel yaklaşımlara üstünlük sağlamaktadır. Hastalıklar için tanı koyucu bir arac olarak kullanılmaktadır. Mikrodizi teknolojisinin (MA), genellikle gen ifade analizi, çevresel araştırmalar, genetik ve mutasyon analizleri, antimikrobiyal gen analizi gibi kullanım alanları mevcuttur. (Yoltas ve Karaboz, 2010).

Genel olarak kullanılan alanlar şu şekilde sıralanabilir. (Mergen, 2015):

1. Gen ifade profillerinin çıkarılması
2. Polimorfizm analizleri
3. Mutasyon analizleri
4. DNA dizi analizi
5. Erimsel çalışmalar

6. Potansiyel terapotik ajanların bulunması, geliştirilmesi, optimizasyonu ve klinik değerlendirmesi

DNA mikrodizi kanser hücrelerindeki gen ifade değişimlerinin sınıflandırılmasında yaygın olarak kullanılmaktadır. Örnek verilirse, ifadesi ölçülen 5500 civarında insan akciğer epitelyum hücrelerindeki genler ile akciğer kanserli doku genleri ile kıyaslanabilmektedir. Bu şekilde kanserleşme sürecinde rol oynayan genler hakkında bilgi toplamak mümkün hale gelir. Mikrodizi teknolojisinin kanser tedavisindeki diğer bir önemli etkisi, gen ifade ölçüm değerlerine göre kanserleşen hücrelerin kategorizasyonudur.

1.1.2 Gen ifadesi veritabanları

Mikrodizi deney sonuç verilerinin, deneylerde kullanılan örneklerin klinik ve/veya patolojik özelliklerinin bulunduğu veritabanları bulunmaktadır. Son yıllarda mikrodizi, biyolojik çalışmaların ortak bölümü haline gelmiştir. 2004 ten itibaren bilimsel yayıncılar tarafından yapılan mikrodizi deney sonuçlarının halka açık bir veritabanına kaydedilmesini zorunlu hale getirilmiştir (Ball et al.,2004). Bu veritabanları, çalışmaların verilerinin ve yapılış şekillerinin belli bir standartta ulaşılabilir olması için koyulan MIAME (Minimum information about a microarray experiment) kriterleri doğrultusunda veri ve bilgi paylaşımı yapmaktadır (Brazma et al., 2001). MIAME kriterlerine göre ham veri, işlenmiş veri, deney ile ilgili doz vb. temel bilgiler, örnek-veri ilişkileri de dâhil olmak üzere deneysel tasarım, platform ile ilgili gen tanımlayıcılar ve genomik koordinatlar gibi anotasyon bilgileri, gerekli laboratuvar ve veri işleme protokolleri (ör: normalizasyon yöntemi) bu veritabanlarında ulaşılabilir olmalıdır.

Mikrodizi veritabanları iki ayrı sınıfa ayrılır :

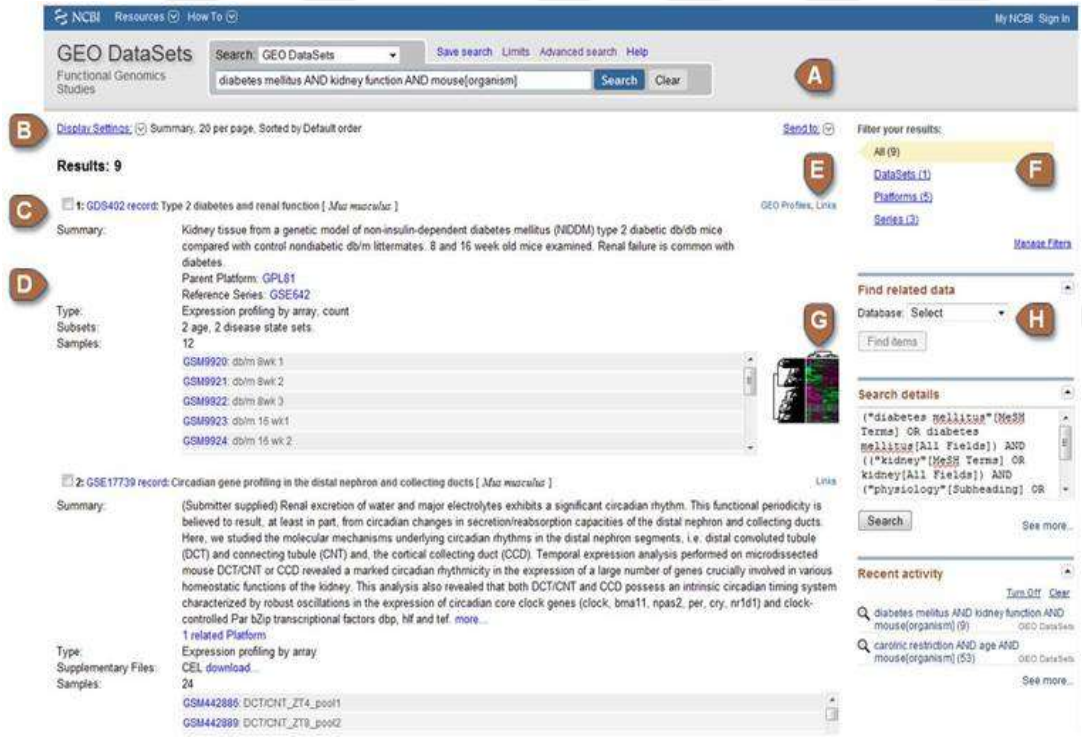
- Hakemli, halka açık, akademik veya endüstri standartlarına uygun ve birçok analiz uygulamaları ve gruplar tarafından kullanılmak üzere tasarlanmış veritabanları: ArrayExpress (Brazma et al.,2003), Gene Expression Omnibus: GEO (Barrett et al.,2009).
- Belirli bir grup ile ilişkili özel veritabanları (laboratuvar, şirket, üniversite, konsorsiyum): Standford mikrodizi veritabanı (Sherlock et al., 2001).

Bu veritabanlarından en popülerleri ArrayExpress ve GEO dur.

ArrayExpress, avrupa kökenli büyük ve kapsamlı bir biyolojik veritabanı olan EBI'nin (The European Bioinformatics Institute) sağladığı mikrodizi veritabanıdır. Günümüz itibariyle bu veritabanında 65.100 deney (EBI, 2016) bulunmaktadır.

Gene Expression Omnibus (GEO) ise Amerika Birleşik Devletleri kökenli ve dünyanın en kapsamlı biyolojik veritabanı olan Ulusal Biyoteknoloji Bilgi Merkezi (National Center for Biotechnology Information-NCBI) tarafından sağlanan mikrodizi veritabanıdır. 2001 yılında kurulan GEO'da 2016 Ocak ayı itibariyle 63.629 çalışma, bu çalışmaların içinde toplam 1641025 örnek ve 15.244 çeşit platform bulunmaktadır (Şekil1.2) .

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. Bu çalışmada kullanılan gen ifade verileri GEO veritabanından sağlanmıştır.



Şekil 1.2 GEO veritabanı arayüzü

Şekil 1.2 de GEO veritabanının arayüz görüntüsü verilmiştir (NCBI, 2016). Burada harfler ile işaretlenen kısımlar aşağıda açıklanmıştır.

A. Arama Kutusu, ilgilenilen GEO veriseti anahtar kelime veya arama terimi bu alana girilerek ulaşılmaktadır. Organizma,veriseti tipi, yazar gibi arama için bir çok terim kullanılabilir.

B. Görüntü ayarları

C. Başlık Satırı Verisetini (GDS), serileri (GSE) veya platform GPL erişim numarası, başlık ve organizma bilgileri ile listelenir.

D. Veriseti, seri veya platformun özet tanımlamaları, tipi, alt küme, ek dosyalar ve örnekler

E. Diğer veritabanlarındaki (PubMed, Epigenomis, SRA) ilgili kayıtların bağlantıları

F. Sonuçların Filtrelenmesi.

G. Verisetleri için kümeler sağlanmıştır, küçük resime tıklanınca birçok veri analizi aracının bulunduğu Veriseti kaydına yönlendirilir.

H. İlgili veri bulma

GEO veritabanında anahtar kelime ile arama yerine karmaşık sorgular oluşturulmasına da imkan verilmiştir. Karmaşık sorgu oluştururken aranacak terim, alan ve boolean operatörler (“VE”, “VEYA”, “DEĞİL”) kullanılmalıdır. Sorgunun sözdizimi: terim[alan] OPERATOR terim [alan] şeklindedir. Örneğin “human[organism] AND (smok* OR diet)” sorgusu içerisinde “smok” veya “diet” kelimelerinin geçtiği insan organiması üzerinde yapılmış çalışmaları listelemektedir. Çizelge 1.1 de GEO da tanımlı bazı alanlar ve bu alanlara yönelik örnek sorgular verilmiştir.

Çizelge1.1 GEO da tanımlı bazı sorgu alanları ve örnekleri

Alan Adı	Örnek
Tüm Alanlar (All Fields)	İçerisinde kanser kelimesi geçen tüm kayıtlar “cancer[All fields]”
Yazar (Author)	A Smith tarafından yazılmış tüm kayıtlar “smith a[Author]”
Tanım (Description)	Tanımlamasında Sigara içmek ile ilgili terimler içeren tüm çalışmalar “smok*[DESC]”
Girdi Tipi (Entry Type)	Sadece Veriseti kayıtları “gds[Entry Type]”
Filtre (Filter)	PubMed bağlantısı olan kayıtlar “gds pubmed[Filter]”
GEO Erişimi (GEO Accession)	GPL570 platformunda yapılan tüm çalışmalar

	"GPL570[GEO Accession]"
Örnek Sayısı (Number of Samples)	100 ile 500 örnek arasında yapılan çalışmalar "100:500[Number of Samples]"
Organizma (Organism)	Fare üzerinde yapılan çalışmalar "Mus musculus[Organism]"
Proje (Project)	NIH Roadmap Epigenomics projesindeki çalışmalar "roadmap epigenomics[Project]"
İlgili Platform (Related Platform)	GSE22474 ile ilgili platformlar "GSE22474[Related Platform]"
Örnek Kaynağı (Sample Source)	Beyinden örneklerle yapılmış çalışmalar "brain[Sample Source]"
Örnek Tipi (Sample Type)	Protein örnekleri kullanılan çalışmalar "protein[Sample Type]"
Teslim eden Enstitü (Submitter Institute)	Broad enstitüsü tarafından yapılmış çalışmalar "Broad Institute[Institute]"
Altküme Tanımı (Subset Description)	Altküme tanımında 'male' terimi geçen verisetleri "male[Subset Description]"
Altküme Değişken Tipi (Subset Variable Type)	Deneysel değişken olarak 'age' kullanan verisetleri "age[Subset Variable Type]"
Başlık (Title)	Başlığında Affymetrix geçen kayıtlar "Affymetrix[Title]"
Güncelleme Tarihi (Update Date)	Haziran 2010 dan itibaren güncellenmiş çalışmalar "2010/06[Update Date]"

1.1.3 İçerik-tabanlı arama

İçerik tabanlı arama, arama işleminin temelinde anahtar kelime, açıklayıcı metin, etiket gibi tanımlayıcılar olmaksızın aranacak veri içeriği analiz edilerek yapılmaktadır. Sadece metin arama değil görüntü-hareketli görüntü erişim

sistemlerinde, müzik verilerinde uygulama alanları mevcuttur (Soydal vd., 2005). Bu çalışmada büyük ölçekli genomik veritabalarında arama yapılacağı için kaydedilen mikrodizi deneylerinin içeriğinin doğru analiz edilmesi gerekmektedir. Bir mikrodizi deneyinin içeriği; gen ifade profillerinin ifade seviyelerini gösteren değerlerdir.

Hali hazırda mikrodizi veritabanlarında arama için ortak yaklaşım: metinsel açıklama (annotation), dizi ve genlerin tanımlamaları gibi üstveri tabanlıdır. GEO, konu ile aramaya imkân vermektedir. Ancak bu tür aramalarda tamamlanmamış kategorizasyon ve örneklerin etiketlenmesi gibi sebeplerden tüm ilgili veri setleri bulunamamaktadır. GEO'nun arama kapasitesini arttırmak için GEOmetadb geliştirilmiştir. GEOmetadb, GEO da arama seçeneği sunan bir SQL veritabanı versiyonudur. Bu aramalar konu bulmak için verimlidir fakat benzer sonuçlar vermiş çalışmaları bulamaz. Benzer sonuçlar vermiş çalışmaları bulmaya "İçerik-tabanlı arama (Content-based search)" denir (Bell and Sacan, 2012).

Benzer içerikli deneyleri bulmak için kullanılan yöntemlerden biri farklı ifade olmuş (Differentially Expressed, FİO) genlerin belirlenmesi tabanlıdır. FİO geni tanımlamanın en basit yolu farklı örneklerdeki veya farklı koşullar altındaki aynı örneğin ölçülen ifade oranına bakmaktır. Eğer oran belirlenen eşik değerini geçerse gen veya protein FİO gendir denir. Eşik değeri genellikle 2 olarak kullanılmaktadır. Eğer eşik değeri yeteri kadar yüksek olursa mesela 12, bulunan sonuçların kesinlikle FİO gen olduğundan emin olunabilir fakat bu durumda birçok FİO gen de gözden kaçabilir. Eğer düşük bir eşik değeri atanırsa gerçekte FİO gen olmayanlar da FİO gen olarak tanımlanabilir. Bu sebeple ifade oranlarındaki değişimin (fold change) belirlenmesi sadece eşik değerine göre değil aynı zamanda istatistiksel hesaplamalar ile de yapılması gerekmektedir.

FİO genler bulunduktan sonra sorgu deney ile içerik olarak benzer deneyler benzerlik ölçütleri kullanılarak listelenebilir. Mikrodizi gen ifade veritabanlarında içerik-tabanlı arama için literatürde farklı yaklaşımlar mevcuttur. Ancak GEO'nun henüz içerik tabanlı arama yapan bir fonksiyonu bulunmamaktadır.

1.2 Önceki Çalışmalar

Mikrodizi gen ifade veritabaları için önerilen yöntemler, ifade profili sunum şekli ve profilleri karşılaştırmak için kullanılan benzerlik ölçütlerine göre gruplandırılabilir. İfade profili sunumlarına göre iki gruptan bahsedebiliriz: farklı ifade olmuş gen profilleri sunumu (Bell and Sacan, 2012; Engreitz et al., 2010a; Engreitz et al., 2011; Fujıbuchi et al., 2007) ve bilinen gen kümesi sunumu (Caldas et al., 2009). Gen profillerini karşılaştırırken kullanılan benzerlik ölçütleri oldukça önemlidir bu sebeple genellikle araştırmacılar birden fazla ölçüt

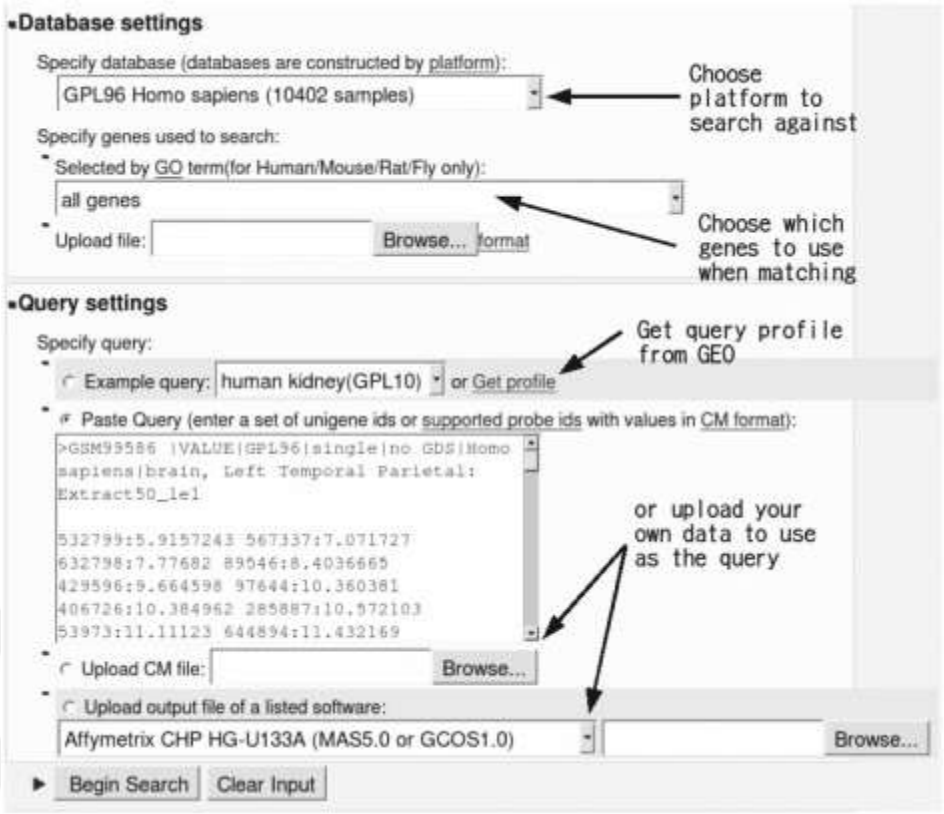
kullanarak en iyi sonuç aldıkları ölçüte karar vermektedirler. En yaygın kullanılan benzerlik ölçütleri olarak Spearman korelasyon katsayısı, Tanimoto katsayısı, Pearson katsayısı ve Öklit uzaklığı sayılabilir.

Bu bölümde bazı önemli yöntemlerin detayları verilecektir. Geliştirilen yöntemler web aracı olarak (Fujıbuchi et al., 2007; Hunter et al. 2001; Chen et al.,2008; Feng et al., 2009) ve algoritma olarak (Caldas et al., 2009; Horton et al. 2006; Georgi et al., 2012; Hayran vd. 2014) ayrı ayrı incelenecektir.

1.2.1 Web araçları

Hunter vd. nin 2001 yılında önerdiği GEST(Gene Expression Search Tool) içerik tabanlı arama için geliştirilen ilk web aracıdır. Bu çalışmada gen ifade verileri için arama yaparken kullanılan uzaklık ölçütleri 92 adet maya ifade deneyi üzerinde denenmiştir. Bunlardan Öklit uzaklığı ve korelasyon benzerliği hali hazırda olan yöntemlerken Bayes ölçütü ilk kez GEST için önerilmiştir. Veritabanı aramalarında kullanılan uzaklık ölçütleri (korelasyon katsayısı, Öklit uzaklığı), gen ifade verisinin karmaşık dağılımlı formda ve çok boyutlu normal olmayan yapısından dolayı kullanışlı olmamaktadır. Bu sebeple Bayes oranı temelli yeni bir benzerlik ölçütü önerilmiştir. Bayes benzerlik ölçütünü kullanarak iki dizi profilini karşılaştırır. Fakat önerilen yeni metrik de önemli ölçüde bir iyileştirme sağlamamıştır. Yöntem çığır açmasına karşın büyük veritabanlarında arama yapmak için yeterli değildir.

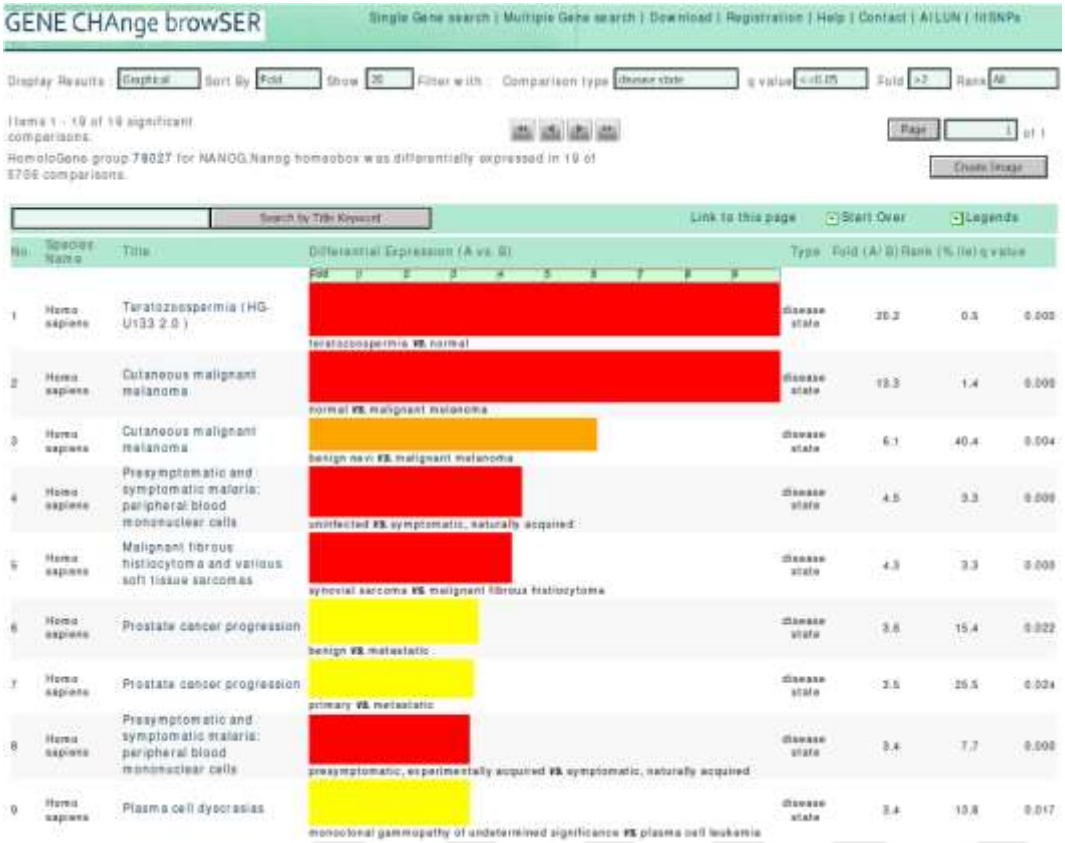
Fujibuchi vd. 2007 yılında RaPiDS algoritmasını (Horton et al.,2006) CellMontage'i yaratmak için kullanmıştır. Horton vd. 2006 yılında önerdiği RaPiDS algoritması gen ifade profili veritabanları için ilk genişletilebilir profil karşılaştırma algoritmasıdır. Fikrin önerilmesi ve ilk örnek sistemin gerçeklemesi her ne kadar GEST ile olsa da büyük veritabanlarında ifade profil benzerliği araması yapan ilk sunucu CellMontage'dir (Şekil1.3). CellMontage hızlı bir şekilde örneklerin sorgu ile aynı tür hücre veya doku olup olmadığını tanımlar. Benzerlik ölçütü olarak Spearman sıralama katsayısını tabanlıdır ve farklı şekillerde normalize edilmiş verileri karşılaştırmaktadır. Yöntem yüksek sıralama korelasyonuna sahip deneyleri getirir, getirilen deneyler sorgu deney ile yeteri kadar ortak gene sahiptir. CellMontage, GEO veritabanından alınmış 65.000 profil kullanır ve gen id lerinin UniGene isimlerine dönüştürerek platformlar arası sorgulamaya imkan verir. Fakat bu, konu ile aramanın üzerine sadece küçük bir gelişme olarak kalmıştır.



Şekil 1.3 CellMontage'in sorgu girdi ekranı (Horton et al.,2006)

2008 yılında Chen vd., Gen Avcısı (GeneChaser-GENE CHAnge browSER) isimli başka bir web sunucusu geliştirmiştir. Gen avcısı herhangi bir biyolojik altyapı gerektirmeden, gen veya gen kümelerinin farklı ifade olma değerlerine göre kolayca biyolojik durumları tanımlayabilmektedir. Yöntem, sonuçları grafiklerle gösterirken istatistiksel karşılaştırma, sıralama/filtreleme ve orijinal çalışmaya erişim gibi imkânları sunmaktadır. Uygulamada 213 farklı mikrodizi platformundan 42 farklı tür içeren 1.515 GEO veri kümesi kullanılmıştır. Gen avcısı her bir karşılaştırma için q-değerine göre tüm farklı ifade olmuş genleri kaydetmektedir.

Tek gen arama ve çoklu gen arama olmak üzere iki farklı arama fonksiyonu vardır. Tek gen arama girdi olarak bir genin tanımlayıcısını alır ve o genin farklı ifade olmuş hali ile ilgili karşılaştırmaları listeler. Örneğin Şekil 1.4 (Chen vd.,2008) de Nanog geni için arama yapılmış ve toplam 5.706 deney durumundan 19 hasta durumunda farklı ifade olduğu GeneChaser tarafından bulunmuştur. Ayrıca top 9 hastalık durumu karşılaştırması listelenmiştir.



Şekil 1.4 GeneChaser Nanog için tek gen arama sonuçları (Chen vd.,2008)

Çoklu gen arama ise girdi olarak iki tane gen tanımlayıcı listesi alır (fazla-ifade olmuş, az-ifade olmuş) ve ilgili karşılaştırmaları listeler. Örneğin Şekil 1.5 de Nanog, Oct4 ve Sox2 genleri için yapılan arama sonuçları verilmiştir. Bu uygulamanın dezavantajı aynı anda az sayıda genin aranmasına olanak vermesidir.

GENE CHAnge browSER

Single Gene search | Multiple Gene search | Download | Registration | Help | Contact | AILUN | fitSNPs

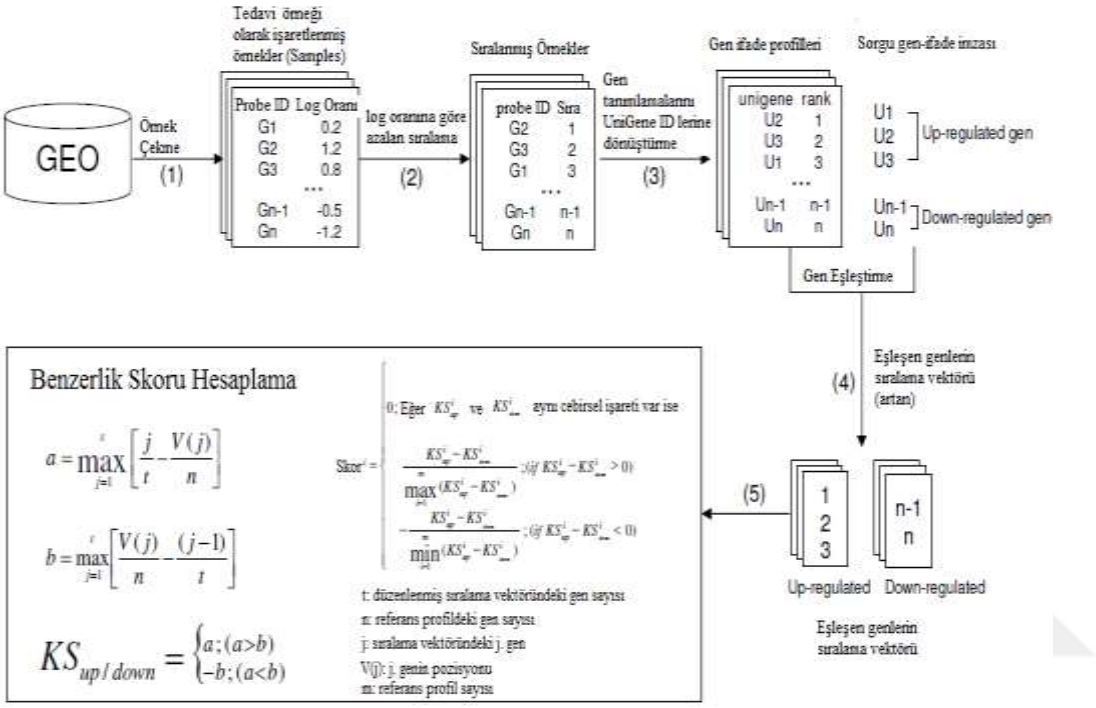
Link to this page

No.	Title	Subset A *vs.* Subset B	Type	Up Fold(A/ B)	Down Fold(A/ B)	q value(Avg.)
1	Expression data from different research centers	universal RNA ***vs.*** liver	tissue	28.7	0.0	0.0060
2	Expression data from different research centers	universal RNA ***vs.*** spleen	tissue	15.9	0.0	0.0083
3	Expression data from different research centers	universal RNA ***vs.*** kidney	tissue	9.7	0.0	0.0020
4	Teratozoospermia (HG-U133 2.0)	teratozoospermia ***vs.*** normal	disease state	8.2 Symbol / Fold NANOG / 20.236 POU5F1 / 2.037 SOX2 / 2.275	0.0	0.0000 Symbol / q value NANOG / 0 POU5F1 / 0 SOX2 / 0
5	Immortalized endothelial cell line response to atorvastatin	primary ***vs.*** immortalized	cell line	6.7	0.0	0.0063
6	Cutaneous malignant melanoma	normal ***vs.*** malignant melanoma	disease state	6.2 Symbol / Fold NANOG / 13.3333 POU5F1 / 3.81679 SOX2 / 1.5083	0.0	0.0060 Symbol / q value NANOG / 0 POU5F1 / 0 SOX2 / 0.018
7	Hematopoietic stem cell engraftment in goat	Capra hircus ***vs.*** Homo sapiens	species	6.1	0.0	0.0127
8	Whole blood RNA isolation and leukocyte RNA purification: comparison of methods	PAXgene ***vs.*** buffy coat	protocol	4.2	0.0	0.0143
9	Whole blood RNA isolation and leukocyte RNA purification: comparison of methods	whole blood ***vs.*** leukocyte	tissue	4.2	0.0	0.0143
10	Presymptomatic and symptomatic malaria: peripheral blood mononuclear cells	uninfected ***vs.*** symptomatic, naturally acquired	disease state	3.3 Symbol / Fold NANOG / 4.4843 POU5F1 / 2.34742 SOX2 / 3.0303	0.0	0.0000 Symbol / q value NANOG / 0 POU5F1 / 0 SOX2 / 0
11	Presymptomatic and symptomatic malaria: peripheral blood mononuclear cells	uninfected baseline ***vs.*** acute infection	infection	3.3	0.0	0.0000
12	Plasma cell dyscrasias	monoclonal gammopathy of undetermined significance ***vs.*** plasma cell leukemia	disease state	2.7 Symbol / Fold NANOG / 3.41297 POU5F1 / 2.25734 SOX2 / 2.39808	0.0	0.0273 Symbol / q value NANOG / 0.017 POU5F1 / 0.045 SOX2 / 0.02
13	Various normal pure cell cultures	epithelial ***vs.*** endothelial	cell type	2.6	0.0	0.0047

Şekil 1.5 GeneChaserda Nanog,Oct4 ve Sox2 genleri için çoklu gen arama sonuçları (Chen vd.,2008)

2009 yılında Feng vd. GEO veritabanından alınan sorgu verilerini gen ifadesi imzasına veya gen ifadesi oranına dönüştürüp arama yapan Gem-Trend (Gene Expression data Mining Toward Relevant Network Discovery) isimli web aracını geliştirmişlerdir. İki gen ifade imzasının benzerliğinin karşılaştırılması parametrik olmayan, sıralama tabanlı desen eşleştirme yaklaşımı (Lamb vd. 2006) ile yapılmaktadır. Yöntem, ilk olarak Kolmogorov-Smirnov (KS) skorlarını sorgunun hem up-regulated hem de down-regulated gen kümeleri için hesaplamaktadır. Sonra hesaplanan skorlar tek bir skora birleştirilir. GEO örnekleri bu birleştirilmiş skora göre azalan olarak sıralanmaktadır (Şekil1.6).

Yüksek skora sahip örnekler sorguya daha yakın örneklerdir. GemTrend gen ifade imzası oluşturma kapasitesi 500 gen ile sınırlıdır. Eğer verilen veri 500 genden fazla gen içeriyor ise tüm genler mutlak oran değerlerine göre sıralanır ve 2 eşik değerini geçen ilk 500 gen seçilir. Son olarak seçilen genler işaretlerine göre iki kümeye ayrılır: up-regulated, down regulated.



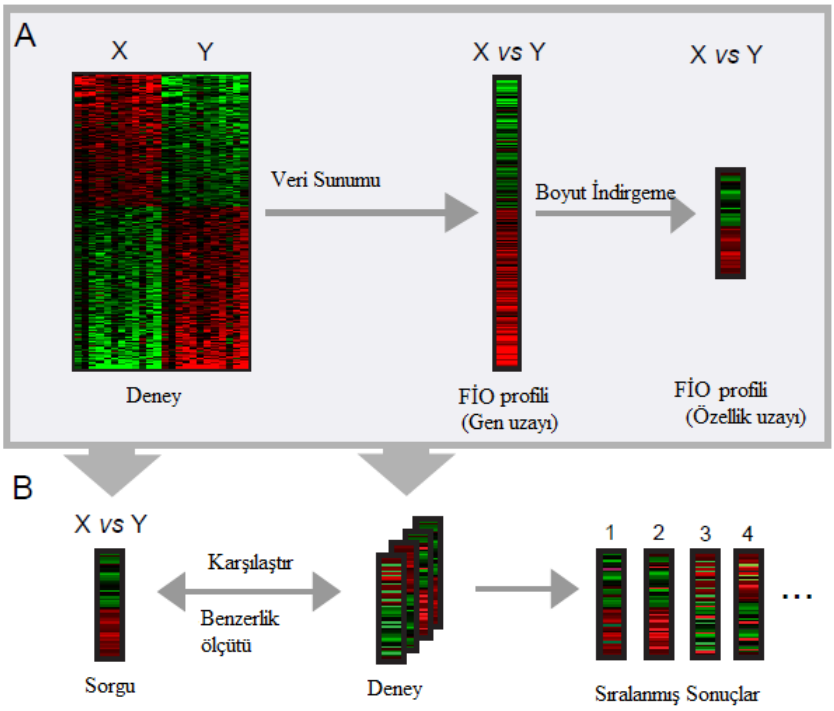
Şekil 1.6 Gem-trend'de referans gen ifadesi profili oluşturma ve benzerlik skorunun hesaplanması. (Feng et al., 2009)

Şekil1.6 da 1) hastalık olarak işaretlenmiş (annotated) Gen ifadesi verilerinin GEO veritabanından çekilmesi, 2) Hastalık ve kontrol ölçümlerinin log-oranlarına göre her bir örnekteki genlerin azalan şekilde sıralanması. 3) platform özelliklerine göre verilen gen tanımlamalarının UniGene ID formatına dönüştürülmesi 4) Sorgu ve referans ile eşleşen up-down regulated genlerin sıralanmış vektörün sıralanması 5)Benzelik skorunun hesaplanması adımları gösterilmektedir.

ProfileChaser isimli son gen arama web sunucusu Engreitz vd. tarafından önerilmiştir (Engreitz et al., 2010a, 2011). Eşik değerini görmezden gelerek deneyde ölçülen her gen için "skor" değer eklenerek yeni FİO gen profilleri oluşturan bir yöntemdir. Çeşitli boyut indirgeme yöntemleri ile korelasyon ölçütleri, ilgili deneyleri sıralamak için kullanılmıştır. Bunlardan matris dekompozisyonu ve p-weighted Pearson korelasyon kombinasyonunun, FİO gen profili karşılaştırmak için en iyi ikili olduğu görülmüştür.

Çalışma, 32 deneyden hasta ve normal dokuları karşılaştıran 1278 dizi üzerinde denenmiştir. Dokular, Duchenne Müsküler Distrofisi, Meme kanseri ve Huntington Koresi hastalıklarını taşımaktadır. Deneyler çeşitli dokular, türler ve platformlar üzerinde yapıldığı için ifade verisinin kümelenmesini etkilemektedir. Farklı veri setlerindeki bu bilgileri standartlaştırmak için prob kümeleri Entrez Gen tanımlamalarına dönüştürülmüştür. Bioconductor yazılımı (R-paketi) kullanılarak her bir karşılaştırma için FİO gen profilleri oluşturulmuştur. FİO gen profilleri her biri bir skora sahip olan özellik listesinden oluşmaktadır. Yöntemi, veri sunumu, boyut indirgeme ve arama algoritması olarak 3 ana bölümde özetleyebiliriz (Şekil1.7).

Tipik mikrodizi analiz yöntemleri FİO genleri ifade değerlerindeki değişim oranı (fold change) olarak tanımlarken, bu çalışmada hem parametrik hem de non-parametrik veri sunumu – FİO gen profillerini karşılaştırırken en iyi yöntemi bulmak için- kullanılmıştır. Öncelikle FİO gen profilleri oluşturulup ilgili skor, ifade değerlerindeki değişim oranının logaritması (log fold change), p-değer profili ve sıra profili yöntemi ile atanmıştır. Tüm mikrodizi verilerinin normalize edilmiş prob düzeyindeki maksimum ve minimum değerlerine bakılarak logaritma değerleri hesaplanmış gerekli yerlerde log2 transformasyonu uygulanmıştır. Her prob'un varyansı ile notlandırılmış genlerin ortalamaları hesaplanarak prob ile gen birleştirilmiştir. Normal ve hasta doku arasındaki fold-change farkı herbir gruptaki örneklerin ortalamaları ile hesaplanmıştır.



Şekil 1.7 ProfileChaser yaklaşımının şematik diyagramı (Engreitz et al., 2010a)

Şekil 1.7 A) FİO gen profillerinin oluşturulması X ve Y koşullarını karşılaştıran bir deney tek FİO gen profilinde birleştirilmiştir. B) FİO gen profillerinin kütüphanede aranması işlemeni göstermektedir. Sorgu profili diğer oluşturulan tüm FİO gen profilleri ile benzerlik ölçütleri kullanılarak karşılaştırılır.

Her bir deneydeki genlerin FİO olup olmadığı t-istatistik (Emprical Bayes moderated) ile belirlenmiştir. İndirgenmiş özellikler ile oluşturulan FİO gen profillerinin FİO değerini ölçmek için limma (R paketi) yazılımı kullanılmış ve p-değer profiller oluşturulmuştur. Sıralama profilini oluştururken her deneydeki örnekler için prob işlenmemiş ifade skorlarına göre sıralanmıştır. Sonra bir prob için tüm skorların averajı alınarak normal ve hasta örnek grupları için tek bir skor hesaplanmıştır. Bu şekilde sıralama profili oluşturulmuştur.

Yukarıdaki üç FİO gen profili sunumunun NCBI homologene kullanılarak insan geni karşılıkları bulunmuştur. Birebir insan karşılığı olmayan genler silinmiştir. Daha sonra ICA (Independent Component Analysis)'nın izdüşüm yöntemi (Engreitz et al., 2010b) ve fixed effect meta estimate yöntemi kullanılarak boyut indirgeme yapılmıştır. Sonraki testlerde ICA yönteminin daha başarılı olduğu belirlenmiştir.

Çalışmada denenen benzerlik ölçütleri, Pearson korelasyon katsayısı ve Spearman sıralama korelasyonu tabanlıdır. Çalışmada kısaca çeşitli boyut indirgeme yöntemleri (ICA-projection, Gene, MsigDB, HsGxModule) ve çeşitli benzerlik ölçütleri (Pearson, Spearman,var-weight Pearson, P-weight Perason, P-weight Spearman, RankDiff Pearson) farklı kombinasyonlarda kullanılarak test edilmiştir. En iyi sonuç veren kombinasyon: ICA-izdüşüm ve P-weight Pearson olmuştur. P-weight Pearson korelasyonu ile en iyi benzer FİO gen profilleri bulunmuş, ICA tabanlı yöntem ile de özellikler 50 katı kadar azaltılmıştır.

1.2.2 Algoritmalar

Bu bölümde mikrodizi verileri için içerik-tabanlı arama algoritmalarına değinilecektir.

Caldas vd. (2009) sorgu deney ile aynı biyolojik süreçten geçmiş deneyleri bulmayı amaçlayan bir algoritma önermişlerdir. Bunu başarmak için var olan tekniklere ek olarak yeni olasılıksal makine öğrenme tekniği kullanmışlardır. Önerilen yöntem, her deneyde farklı şekilde aktive edilmiş biyolojik süreçten bilgi çıkarmakta, aynı süreçlerin aktive ettiği önceki deneyleri getirmekte ve getirilen sonuçları yorumlamaktadır. Yöntem hesaplanabilir verilerdeki gizli bileşenleri bulmak için kullanılan konu modelini adapte etmiştir (Griffiths et al., 2004; Blei

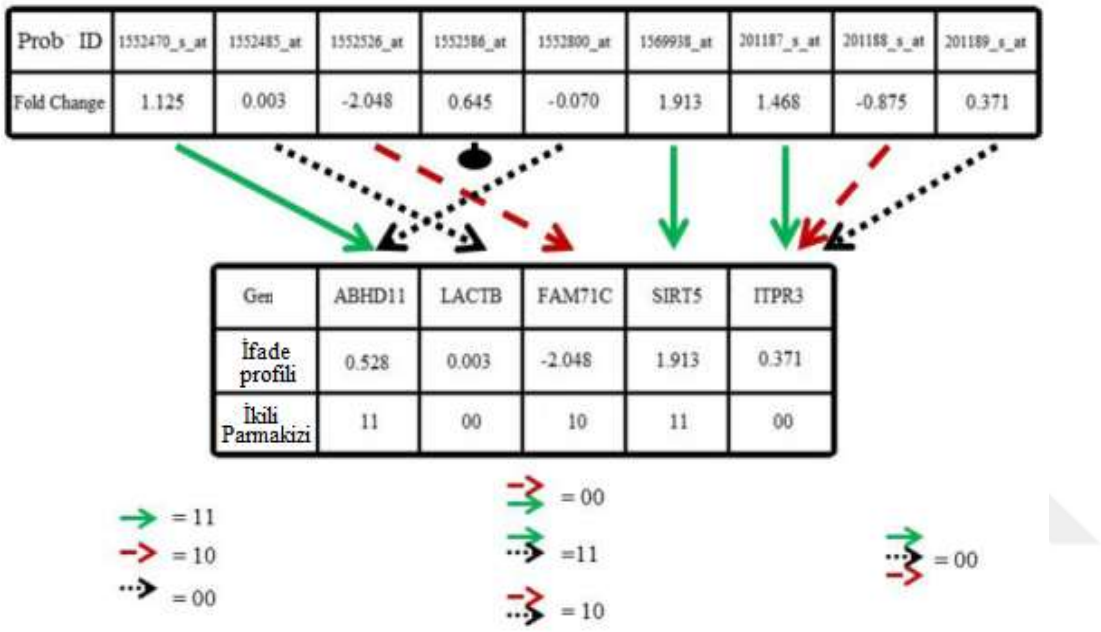
et al.,2003; Buntine and Jakulin, 2004). Adapte ederken kelime sayısını, gen kümesindeki farklı ifade olmuş gen sayısına, kelime çeşidini gen kümesine dönüştürmektedir.

ArrayExpress veritabanından 288 işlenmiş insan gen ifade mikrodizi deneyleri yöntemi test etmek için kullanılmıştır. Konu modeli zenginleştirilmiş gen kümesi analizi (GSEA-Gene Set Enrichment analysis) (Subramanian et al.,2005) ile işlenen deneyleri modellemek için kullanılmıştır. GSEA ilk olarak fold-change değerine göre genleri sıralar ve liste genelinde Kolmogorov–Smirnov istatistiği kullanarak zenginleştirme skoru hesaplar. Sonrasında hesap olarak her kümenin farklı ifade olma büyüklüğü tanımlanır. Biyolojik resim tek bir gen yerine üst sıradaki gen kümeleri için oluşturulur. Son olarak biyolojik süreçlerine göre ağırlıklandırılan kümeler, konular üzerindeki olasılık dağılımlarına göre karşılaştırılır.

2012 yılında Bell vd. içerik tabanlı aramaya yeni bir bakış açısı ile kemoinformatik alanındaki ikili parmak izi yöntemini mikrodizi deney verilerine uyarlamıştır. Çalışmada farklı ifade olmuş gen profillerini bit vektörü şeklinde sunarak daha hızlı bir karşılaştırma yapılmıştır. Yöntem daha önce Engreitz'in (2010a) derlediği 32 mikrodizi çalışması üzerinde denenmiştir. Bu çalışamalar Duchenne düsküler distrofisi, meme kanseri ve Huntington koresi hastalıkları ile ilgili mikrodizi örneklerini içermektedir. Hastalıklar farklı doku tiplerini etkiledikleri için, farklı ifade profilleri üretmesi beklenmektedir. Gerekli ifade değerleri GEO'dan alınmıştır. Çalışma için toplam 715 alt küme tanımlanmıştır. Yöntem öncelikle verilerin parmak izi için hazırlanması ile başlar. İlk olarak, GEO ya gönderilen örnekler farklı mikrodizi platformundan oldukları için standart gen kümesi ve bu genlerin homologları tanımlanmıştır. Gerekli olduğu durumlarda NCBI homologene uygulaması kullanılarak türler arası haritalama yapılmıştır. Her bir alt kümenin her probunun ortalama ifade değeri hesaplanmış, fold-change olarak bu değerler arasındaki oran kullanılmıştır.

İkili parmak izi hesaplamak için öncelikle ifadedeki anlamlı değişikliği tanımlamak gerekmektedir. Bu sebeple tüm alt kümelere t-test - 5% olasılık anlamlılığı tanımlamak için kullanılmıştır. Sonrasında standart kümedeki tüm genlere 2 bitlik ikili parmak izi atanmıştır (Şekil1.8). Buradaki ilk bit ifadedeki değişikliğin önemli olup olmadığı ile ilgili, ikinci bit ise değişikliğin artan -1- yönde mi azalan-0- yönde mi olduğunu anlamak için kullanılmıştır. Sadece ifade düzeyindeki değişiklik önemli olan problara ikili parmak izi atanarak diğerleri ihmal edilmiştir. Kullanılan benzerlik ölçütleri, Pearson korelasyonu, Spearman Sıralama Korelasyonu, Tanimoto Katsayısı ve Modifiye Edilmiş Tanimoto

Katsayısı'dır. Örnekler hastalık durumuna göre üç gruba ayrılmıştır. Her örnek diğer 715 örnek ile her bir benzerlik ölçütüne göre karşılaştırılmıştır. Son olarak örneklerin sorgu örneğine benzerliğine göre sıralama yapılmıştır. Her bir benzerlik ölçütü için ROC eğrileri çizilmiştir. Doğru-pozitif seçimlerin sorgu ile aynı, yanlış–pozitif seçimlerin ise farklı hastalık durumuna sahip oldukları varsayılmıştır. Eğrinin altında kalan alan (AUC) karşılaştırma için hesaplanmıştır. Spearman Sıralama en fazla AUC üreten, Tanimoto ise en az AUC üreten benzerlik ölçütü olmuştur.



Şekil 1.8 İfade profillerinin ve ikili parmak izlerinin oluşturulması

Şekil 1.8 de yeşil oklar ifade miktarının arttığını gösterir ve parmak izi "11" olarak oluşturulur. Kırmızı kesikli oklar ifade miktarının azaldığını gösterir bu durumda parmak izi "10" olarak oluşturulur. Kesikli siyah oklar ifade miktarında önemli bir değişiklik olmadığını gösterir ve parmak izi "00"dır. Standart kümede olmayan genler siyah daire ile gösterilmiştir ve iptal edilmiştir. Eğer birden fazla prob aynı geni etkiliyorsa fold-change in ortalama değeri alınmıştır.

Giorgi vd. (2012) model-tabanlı bir yaklaşım ile yeni bir algoritma önermişlerdir. Algoritmanın gen ifadesi karşılaştırmaları iki bölümden oluşur. İlki verilen hedef gen listesinin durumsal ifade modelinin oluşturulması, ikincisi oluşturulan modellere göre Fisher skor benzerliği hesaplamaktır. Algoritmanın sezgisel model-tabanlı bilgi erişim modeli iki ifade profilinin benzerliğini karşılaştırırken Fisher-Kernel (Jaakkola and Haussler, 1999) benzerliği kullanmaktadır. Algoritma GEO ve ArrayExpress veritabanlarından sağlanan iki durum çalışması ile (İnsan lösemi ve bitki stresi) test edilmiştir. Fisher-Kernel

benzerliği benzer deneylerin getirilmesinde kullanılmıştır ve daha temel benzerlik ölçütleri (Perason korelasyon, Öklit uzaklığı) ile karşılaştırılmıştır. Ayrıca hibrit bir yaklaşım geliştirilmiştir: belirleyici genler öğrenilen ifade modelinden alınır ve ifade değerleri ile Perason korelasyon, Öklit uzaklığı hesaplanır. Sonuç olarak bu yaklaşım geri getirim bölümünde (retrieval) daha iyi sonuçlar vermiştir.

Son olarak 2014 yılında Hayran ve arkadaşları farklı ifade olmuş gen tabanlı içerik tabanlı arama yöntemi önermişlerdir. Bu çalışmayı diğerlerinden ayıran en belirgin özellik üzerinde çalışılan veri kümesidir (zaman serisi deneyleri). Bu tez çalışmasında geliştirilen yöntem de aynı veri kümesi üzerinde çalışacağı için algoritma FİO gen tabanlı parmak izi oluşturma bölümde detaylı olarak incelenecektir. (Hayran vd., 2014)

1.3 Tezin Katkıları

Mikrodizi gen ifade veritabanlarında içerik tabanlı arama yapan birçok çalışma mevcuttur. Önceki yöntemlerden farklı olarak bu çalışma, zaman serisi deneylerini kümeleyerek tamamını sorgu olarak alan ve gen ifade veri tabanında arayan ilk yöntemdir.

Bu tez çalışmasının katkıları şu şekilde maddelenebilir:

- Önerilen kümeleme-tabanlı geri getirim modelinde kullanılan yöntemlerin zaman serisi mikrodizi verilerinde içerik tabanlı arama için etkili bir yöntem olup olmayacağının araştırılmasına ve üzerinde tartışılmasına zemin hazırlayacaktır.
- "Kümeleme-tabanlı imza çıkarımı" ile mikrodizi deneyleri ile ölçülen her genin ifade değerleri değişimi betimlenmiştir.
- Kıyaslama veri kümesi oluşturulmuş olması, ileride bu alanda yapılacak diğer çalışmalar için de faydalı olacaktır.
- Önerilen yöntemin gen ifade veritabanlarında uygulanması ile dev veri depolarında sorguya dayalı arama yapmak zorunda olan araştırmacıların işleri kolaylaşacaktır. Eksik/yanlış kategorizasyon veya etiketlemeden kaynaklanan arama sonuçlarının yetersizliğinin önüne geçilecektir.
- Binlerce deney içinde keşfedilmeyi bekleyen gerçek bilgi daha kolay bulunabilecektir.

Bu amaçla, kümelenmiş gen profillerine dayanan bir temel kurulmuş, deneyleri betimlemek ve karşılaştırmak için birkaç kümeleme algoritması değerlendirilmiştir. Deneysel çalışma GEO'dan (Gene Expression Omnibus) alınan örnek veri kümeleri üzerinde yapılmıştır.

2. GEREÇ VE YÖNTEMLER

2.1 Geri Getirim Modeli

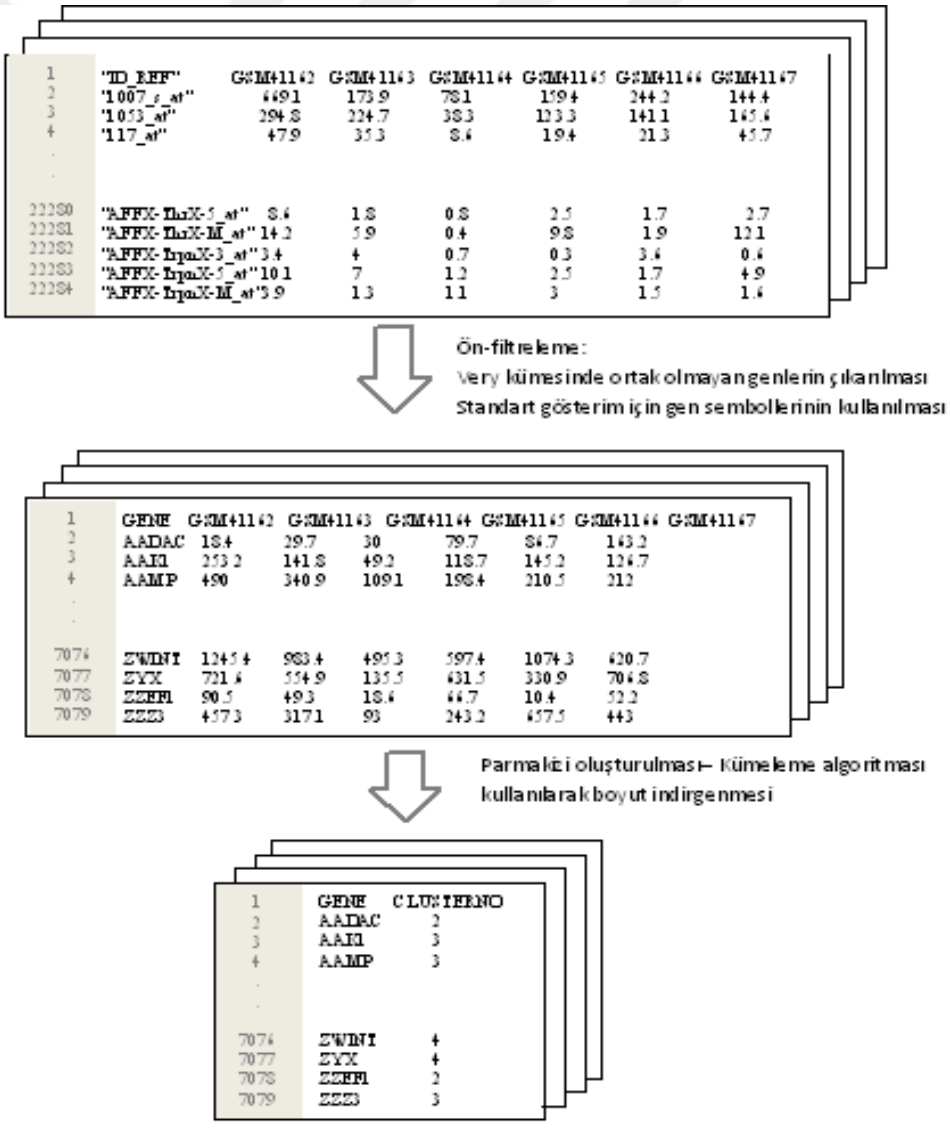
Geri getirim, büyük miktardaki verilerden arama yapma ve bu verilerden istenen bilgiyi üretme/görüntüleme tekniği ve sürecidir. Bilgi geri getirim sistemlerinin temel amacı dermedeki (koleksiyon) ilgili belgelere erişmek, ilgili olmayanları ise ayıklamaktır. İdeal bir geri getirim sistemi salt-ilgili belgelerin tümüne erişmelidir. Fakat genellikle geri getirim sistemleri ilgili tüm belgelere erişemez ve ideal bir geri getirim sisteminin pratikte uygulanabilirliği yoktur. Bu nedenle ideal geri getirim yerine kabul edilebilir geri getirim tanımlanmıştır. Kabul edilebilir geri getirim sistemleri ile oluşturulan benzer belge sıralamasında, ilgili belgeler ilgisizlere göre daha ön sırada yeralması gerekmektedir (Raghavan ve Sever, 1995). Bunun yanında, sistemin kullanıcıları için hızlı bir şekilde birkaç ilgili belgeye erişimi yeterli olmaktadır.

Dermeye eklenen belgelerin karakteristik özellikleri dizinleme işlemi ile belirlenmekte ve belgeler bu dizinler ile kategorize edilmektedir. Kullanıcılar belge ararken uygun arama terimleri için dizinlerden faydalanabilirler. Ancak birçok kullanıcı bu dizinleri kullanmak yerine kendi belirlediği terimler ile arama yapmaktadır. Kullanıcının seçtiği arama terimi ile sistem terimlerinin uyuşmaması durumunda geri getirimlerin başarısız olması kaçınılmazdır. Bilgi geri getirim sistemlerinin başarılı olabilmesi için; belgelere uygun dizin terimlerinin atanması ve kullanıcıların bu dizin terimlerini doğru şekilde sorgularında kullanmaları gerekmektedir. Bu sebeple aranacak belgenin dizin terimi yerine içeriği ile yapılan sorgular geri getirim sistemlerinin daha başarılı sonuçlar almasını sağlamaktadır.

Geri getirim, mikrodizi gen ifade verilerine uygulandığında belgeler, yapılan deneylerde ölçülen gen ifade ölçümlerinin yer aldığı matrislerdir. Mikrodizi gen ifade veritabanlarındaki deneylerin tamamı veya belirli bir bölümü ile de dermeler oluşturulabilir. Mikrodizi deney geri-getirimi, sorgu olarak verilen bir mikrodizi deneyine içerik olarak benzer deneylerin veri tabanında aranarak benzerlik sırasına göre listelenmesi şeklinde tanımlanabilir. Verilen bir ifade matrisi M'de m_{ij}, j. durumdaki i. genin ifade ölçüm değerini gösterir. Deney geri getiriminde amaç, mikrodizi veritabanındaki zaman serisi matrisleri $\{M_1, M_2, \ldots, M_n\}$ içinden, sorgu matrisi M ile en yüksek benzerliği gösteren M_x $(1 \leq x \leq n)$ matrisinin bulunmasıdır.

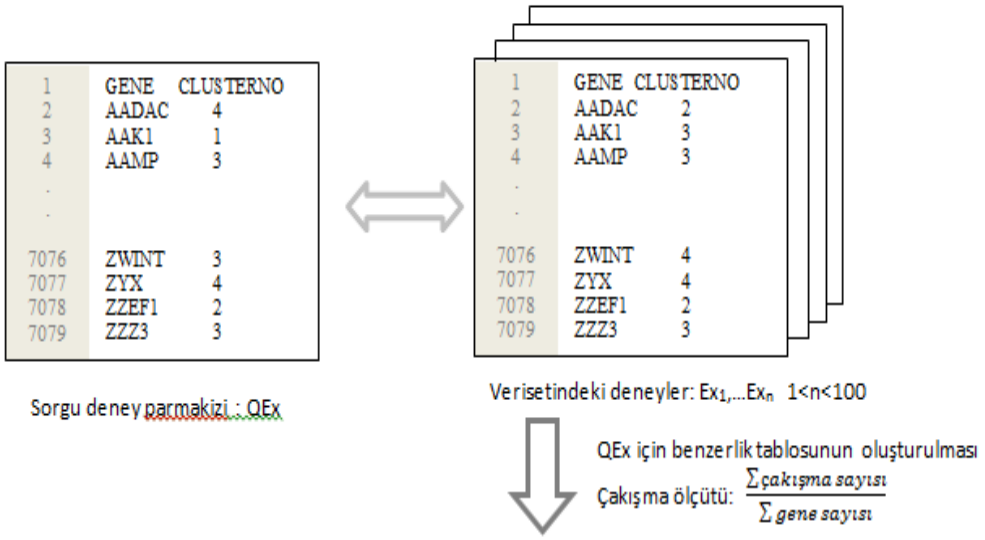
Tez çalışmasında önerilen geri getirim modeli, parmak izi dosyalarının oluşturulması (Şekil 2.1) ve parmak izlerinin karşılaştırılması (Şekil 2.2) adımlarından oluşmaktadır. Parmak izi dosyalarının oluşturulmasındaki amaç, deneylerde farklı sayıdaki zaman noktaları için ölçülen ifade seviyelerindeki

değişimlerin doğru ve etkin biçimde tek bir sütuna indirgemektir. Burada bunun için parmak izi dosyalarının oluşturulması adımında model-tabanlı kümeleme algoritmaları kullanılmıştır. İfade seviyelerindeki değişim modellenerek genler kümelenmiştir. Kümeleme öncesinde deneyler bir ön-filtreleme sürecine tabii tutulur. Mikrodizi gen ifade veritabanından (GEO) toplanan test verisi üzerinde ortak bir gen kümesi belirlenmiştir. Test veri kümesindeki her bir deney için, ortak gen kümesindeki genlerin ifade seviyeleri ve karşılık gelen gen sembolleri kaydedilmiş diğer genler hesaplamaya katılmamıştır. Bu ön filtreleme ile tüm parmak izi dosyalarının aynı uzunlukta olması da sağlamıştır. Ön- filtreleme sonucu oluşan dosyalardaki genler kümeleme algoritması kullanılarak kümelenmiş ve hangi kümeye ait olduğu bilgisi kaydedilmiştir. Gen sembolü ve karşılık gelen küme numarasından oluşan dosyalara parmak izi dosyası denilmiştir. (Şekil 2.1)



Şekil 2.1 Parmak izi oluşturma akış çizgesi

Bilgi getirim modelinin ikinci adımında oluşturulan parmak izi dosyalarının karşılaştırılması yer almaktadır. Burada iki deney arasındaki benzerliği ölçmek için aynı genin farklı iki deneyde aynı kümeye düşme sayısı hesaplanmıştır. Başka bir deyişle farklı iki deneydeki aynı genin, zaman noktaları arasındaki ifade değeri değişimi aynı şekilde olduysa ve iki deneydeki bu şekildeki ortak gen sayısı ne kadar fazla ise deneyler birbirine o kadar benzerdir denilir. Geri getirim modelinde, sorgu deney (parmak izi dosyası) test verilerindeki tüm deneylerle karşılaştırılarak benzerliğe göre sıralanarak benzerlik tablosu oluşturulur (Şekil 2.2). Benzer çıkan deneylerin gerçekten ilgili olup olmadığı hesaplanan ROC ve AP değerleri ile ölçülmüştür.



	QEx	Ex_1	Ex_2	Ex_3	.....	Ex_{n-1}	Ex_n
QEx	1.00	0.7980	0.5010	0.3439		0.1143	0.1099

Şekil 2.2 Parmak izi dosyalarının karşılaştırılarak içerik tabanlı arama

2.2 Parmak İzi Çıkarım Modeli

2.2.1 FİO parmak izi yöntemi

Hayran vd. (2014) tarafından önerilen yöntem, zaman serisi deneyleri üzerinde yapılan ilk içerik tabanlı arama yöntemidir. Yöntemde FİO genlerin bulunması için NUDGE yöntemi (Normal Uniform Differential Gene Expression Detection) (Dean and Raftery, 2005) kullanılmıştır. Nudge yöntemi, temelde normalize edilmiş 2 tabanına göre logaritma oranlarını kullanmaktadır. Model normal-uniform karışım modelidir. Modelde genler FİO olan ve olmayan olarak

iki gruba ayrılır. Her grup kendi yoğunluğuna göre modellenir böylece veri bütün olarak bu karışık yoğunluklar ile ağırlıklandırılır. Ağırlıklar her iki grup için de öncül olasılığına (prior-probability) karşılık gelir. İlk grup FİO olmayan genler grubudur ve logaritma oranları sıfırdır. Gözlenen log oranları bu genler için uygun dönüşümlerden sonra Gaussian yoğunluğu ile modellenir. İkinci grup FİO genlerdir, log oranları diğer gruptan oldukça uzaktır ve uygun aralıkta uniform dağılımlı olarak modellenir.

Model şöyledir ; $x_i \sim \pi N(x_i|\mu,\sigma^2) + (1-\pi)\bigcup_{[a,b]} x_i \ , i = 1,\dots,N$

Model parametreleri ;

$x_i \rightarrow i$ geni için gözlenen log değeri

$\pi \rightarrow$ öncül olasılık

$N(x_i \mid \mu,\sigma 2) \rightarrow \mu$ mean ve σ^2 varyanslı Gaussian dağılımı

$\bigcup_{[a,b]} x \rightarrow [a,b]$ aralığı için uniform dağılım

$N \rightarrow$ gen sayısı

Model, EM (Expectation-Maximization) algoritması kullanarak en çok olabilirlik (maximum likelihood) ile tahmin edilir.

Öncelikle tanımlanmamış etiketler belirlenir:

$z_i, i = 1,\dots,N \rightarrow z_i$ eğer gen FİO gen ise 1 değilse 0 olarak etiketlenir.

EM algoritması iki adımdan oluşur,

1. E adımı (Expectation): verilen parametre tahminleri ile etiketler tahmin edilir.

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

   k. iterasyon için;

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

2. M adımı (Maximization): verilen etiket tahminleri ile π, μ, σ^2 model parametreleri tahmin edilir.
   - $\hat{\pi}^{(k)} = \frac{\sum_{i=1}^{N}(1-\hat{z}_i^{(k)})}{N}$
   - $\hat{\mu}^{(k)} = \frac{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right).x_i}{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)}$

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

Verilen parametre tahminleri ile k. iterasyonda olabilirlik:

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

Yukarıdaki adımlar yakınsayana (convergence) kadar tekrar eder. Yakınsama, parametre tahminlerini, etiketleri ve olabilirlik logaritmalarını her adımda kontrol ederek bulunabilir. Miktarlardaki değişim adımlar arasında yeterince küçüldüğü zaman algoritma yakınsadı denilebilir. Burada z_i'nin başlangıç değeri: i. genin gözlenen log değerinden tüm genlerin ortalama değeri çıkarılıp standart sapmaya bölündüğünde 2 den büyükse 1 değilse 0 olarak kabul edilmiştir.

i. genin final etiket tahmini $\hat{z}_i$, genin FİO olup olmadığı ile ilgili son olasılık (posterior probability) değerini vermektedir.

Nudge yöntemi ile FİO genler bulunurken deneylerdeki zaman noktaları her bir deney için farklılık gösterdiği için temelde iki farklı yöntem uygulanmıştır. İlk yöntemde (son_FİO) FİO genler bulunurken ilk zaman noktası ile son zaman noktasının ifade değerleri esas alınmıştır. Son zaman noktası n ise Nudge (t_1, t_n) değeri hesaplanmıştır. Hesaplanan değer büyükten küçüğe sıralanarak, karşılık gelen gen sembolü ile FİO parmak izi dosyaları oluşturulmuştur.

İkinci yöntemde (max_FİO) ilk ve son zaman noktaları arasındaki olasılık yerine, ilk zaman noktası ile diğer tüm zaman noktaları arasındaki FİO gen olma olasılığı Nudge ile hesaplanmıştır. Hesaplanan olasılıklardan en yüksek olan FİO gen olma olasılığı olarak alınmıştır. Son zaman noktası n ise $i = 2$'den n'ye kadar max $|nudge(\,t_1, t_i\,)$ değeri hesaplanmıştır. Bu yöntem ile amaç, ifade değeri ilk ve son zaman noktasında aynı kalmış, fakat ara zaman noktalarında bir değişim var ise bu değişimi de hesaba katmaktır.

GEO veritabanından 46 tanesi meme kanseri olmak üzere toplam 111 zaman serisi deneyi üzerinde içerik-tabanlı arama çalışmaları yapılmıştır. Farklı ifade olmuş genleri tespit ederken eşik değerinin belirlenmesi oldukça önemlidir. Örneğin eşik küçük bir değer olarak seçilirse gerçekte FİO olmayan genler de FİO gen olarak değerlendirilebilir. Eşik büyük bir değer olursa gerçekte FİO olan bir

gen eşiği geçemediği için değerlendirme dışı kalabilir. Her iki durumda da gerçekçi olmayan sonuçlar alınabilir.

Bu sebeple uygulanan yöntemler birden fazla eşik değeri için denenmiştir. Ayrıca eşik değerinin yanı sıra alınan sonuçların büyükten küçüğe sıralanarak dosyanın belirli bir kısmının FİO olduğu kabul edilip sonuçlar yüzde olarak karşılaştırılmıştır.

2.2.2 Kümeleme

Verileri benzerliklerine göre gruplara ayırma işlemine "Kümeleme" denir. İki temel ilkesi vardır: küme içi benzerliği en fazla; kümeler arası benzerliği ise en az yapmak. Verilerin sahip oldukları karakteristik özellikler temel alınarak ve çok değişkenli istatistiksel yöntemler kullanılarak kümeleme yapılır. Verileri kümelere ayırırken, verilerin birbirlerine olan benzerlikleri/ uzaklıkları ölçülür. Verinin özelliklerine göre (kesikli/sürekli, nominal/ ordinal, vb.) hangi ölçütün kullanılacağı belirlenir. En sık kullanılan ölçütler: Öklit uzaklığı, Pearson uzaklığı, Spearsman sıralama korelasyonu ve Manhattan uzaklığıdır.

Kümeleme tekniklerinin gen fonksiyonlarının, gen düzenlemesinin ve hücresel süreçlerin anlaşılması konusunda yardımcı olduğu kanıtlanmıştır. (Jiang et al., 2004). Gen ifadesi verilerinde iki gen profilinin benzerlikleri temel alınarak genler kümelenir. Bir gen profilinde aynı genin farklı zamanlarda/durumlarda ölçülmüş ifade seviyeleri yer almaktadır. İki gen profilinin benzerliği ölçülen gen ifade değerlerinin arasındaki mesafe ile belirlenir. Herhangi iki gen (g_i, g_j) arasındaki mesafe (d_{ij})'nin taşıması gereken beş özellik vardır (Das et al., 2009)

1) Herhangi iki profil arasındaki mesafe negatif olamaz.

2) Gen profilinin kendisi ile arasındaki mesafe sıfırdır.

3) İki gen arsındaki mesafe sıfır ise profiller aynıdır.

4) g_i ve $g_j$profilleri arsındaki mesafe, $g_j$ve g_i profilleri arsındaki mesafeye eşittir.

$$d(g_i, g_j) = \mathrm{d}(g_j, g_i)$$

5) Mesafe ölçütü üçgenin eşitsizliği kuralını sağlamalıdır. g_k, g_i, g_j gen profilleri için kural :

$$d(g_i, g_k) <= \mathrm{d}(g_i, g_j) + \mathrm{d}(g_j, g_k)$$

Gen ifade profillerini kümeleyebilmek için verinin özelliklerini bilmek gerekmektedir. Gen ifadesi zaman serisi (gen profili), tek örneğin farklı durumlarda/zamanlardaki ifade seviyelerinin ölçülmesi ile oluşturulmuş verilerdir. Gen ifadesi zaman serilerinin, kısa ve düzensiz (unevenly) örneklenmiş olmak üzere iki ana özelliği vardır. Standford Mikrodizi veritabanındaki zaman serisi deneylerinin % 80 inden fazlası 9 dan daha az zaman noktası içermektedir (Kuenzel, 2010) 50'nin altındaki gözlem istatistiksel analiz için oldukça kısa kabul edilir. Gen ifadesi zaman serisi verileri bu özellikler ile diğer zaman serisi verilerinden (işletme,vb.) ayrılırlar.

Gen ifadesi zaman serileri için 3 temel benzerlik gereksinimi tanımlanabilir. Ölçekleme ve kaydırma, düzensiz dağılmış örnekleme noktaları ve şekil (dahili yapı) (Moller-Levet et al., 2003).

1.Ölçekleme ve Kaydırma:

Ortak dizilime sahip genlerin ifade olma şekilleri de benzerdir fakat bu durumda genler aynı zamanda ifade olurken aynı düzeyde ifade olmaları gerekmez. Bu, ölçekleme ve kaydırma problemine neden olan durumlardan biridir. İfade seviyesindeki ölçekleme ve kaydırma faktörü benzer ifadeleri saklayabilir ve iki ifade profili arasındaki benzerlik ölçülürken dikkate alınmamalı elenmelidir. Bu probleme neden olan diğer bir durum ise mikrodizi teknolojisidir ki çoğu zaman normalizasyon ile düzeltilir.

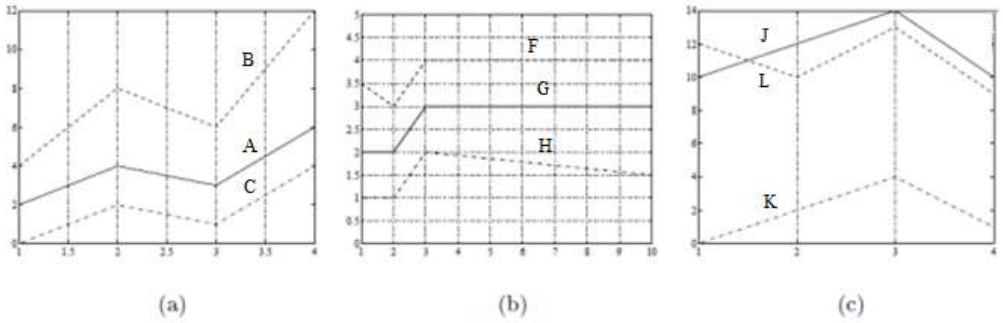
Matematiksel olarak kaydırma ve ölçeklendirme, doğrusal transformasyonun olduğuna işarettir: x ve y iki zaman serisi olsun y eğer $y = mx + b$ olarak ifade edilebilirse x'in doğrusal transformasyonudur. Burada m ölçekleme faktörü, b dikey kaydırma miktarını gösterir. Örneğin, Şekil 2.3 (a) da A, B, C üç zaman serisi, B, A'nın iki ile ölçeklenmiş, C, A'nın 2 kaydırılmış halidir.

2. Düzensiz dağılmış örnekleme noktaları: Örnekleme aralığı uzunluğu bilgi vericidir ve benzerlik karşılaştırmalarında dikkate alınmalıdır. $F[3.5\ 3\ 4\ 4]$, $G[2\ 2\ 3\ 3]$ ve $H[1\ 1\ 2\ 1.5]$ üç zaman serisi olsun. $F\ ve\ G, G\ ve\ H$ serileri arasındaki mutlak hata sırasıyla $[1.5\ 1\ 1\ 1]$ ve $[1\ 1\ 1\ 1.5]$ dir. Görüldüğü gibi seriler aynı hata değerine sahiptir. Örnekleme aralığı zamanla değişim oranı - komşu zaman noktalarındaki ifade seviyesinin değişiminin örnekleme aralığı uzunluğuna oranı- olarak alındığında G, F den daha çok H ye benzer Şekil 2.3(b).

3. Şekil (dahili yapı): Farklı modellerle tanımlanabilir ve genellikle serinin "şekli" olarak yansıtılır. Mikrodizi deneylerinde gen ifadesinin yoğunluğu yerine yoğunluğun değişikliği ifade profilinin şeklini karakterize eder. Dahili yapı,

deterministik fonksiyon, seriyi tanımlayan semboller veya istatistiksel modellerle gösterilebilir.

Şekil 2.3 (c)'de J, K, L zaman serileri olsun. Yoğunluk düşünüldüğünde J, K ya daha benzer ancak yoğunluktaki değişim düşünüldüğünde J, L ye daha çok benzer.



Şekil 2.3 Gen ifadesi zaman serileri benzerliği

Gen ifadesi verileri için birçok popüler kümeleme tekniği mevcuttur. Hepsinin ortak amacı önemli biyolojik süreçlerin oyuncusu olan genlerin farklı fonksiyonel rolünü açıklamaktır. Benzer şekilde ifade olmuş genlerin süreçte benzer fonksiyonel rolü vardır denilebilir (Beal and Krishnamurthy, 2012).

Eğer iki gen aynı kümede ise benzer ifade deseni vardır. Gen ifade verilerini kümelerken karşılaşılan iki büyük soru vardır: Birincisi benzer ifade desenli genler nasıl gruplanır? Diğeri gürültülü veri kümesinden nasıl işe yarar kullanışlı desenler çıkarılır? Gen ifadesi için kümeleme tekniklerini: Paylaştırma, Hiyerarşik, Yoğunluk tabanlı, Model tabanlı ve Çizge tabanlı olarak kategorize edebiliriz.

1. Parçalı Yaklaşım

K-means algoritması tipik bir parçalı tabanlı kümeleme algoritmasıdır, öyle ki veriyi, önceden tanımlanmış küme sayısına, yine önceden tanımlanmış kriterleri optimize edecek şekilde bölmektedir. En önemli avantajı basitliği ve hızıdır. Böylece büyük veri kümelerine de uygulanabilir. Fakat algoritma her çalışmada aynı sonucu üretmeyebilir ve aykırı değerlerin üstesinden gelemez. Self- Organising Map (Tamayo et al.,1999), K-means'ten gürültülü veriyi kümeleme açısından daha sağlamdır. Kullanıcıdan küme sayısı ve ızgara kümeleri girmesi istenir. Gen ifade verileri için küme sayısını belirlemek oldukça zordur. Ayrıca paylaştırma yaklaşımı az boyutlu, özünde iyi ayrılmış veriler için uygundur. Fakat gen ifadesi veri setleri çok boyutludur ve genellikle kesişim, gömülü kümeler içerir.

Özünde bu algoritmalar n zaman noktalı ifade profillerine n-boyutlu vektör olarak davranır. İki profil arasındaki benzerlik miktarını belirlemek için uzaklık veya korelasyon fonksiyonlarını uygular. Öklit uzaklığı, Manhattan uzaklığı, Pearson korelasyon katsayısı en fazla kullanılan fonksiyonlardır.

2. Hiyerarşik Yaklaşım

Veriyi hemen kesişmeyen parçalara ayırmayı amaçlayan parçalı yaklaşımın tersine, Hiyerarşik yaklaşım, grafiksel olarak gösterilebilen (dendongram -ağaca benzer diyagram-) içiçe kümelerin hiyerarşik serisini üretir. Dendogramın dalları, kümelerin biçimlendirme bilgisini tutmasının yanısıra kümeler arası benzerliği de gösterir. Dendogramın belirli bir düzeyde kesilmesi ile belirlenmiş sayıda küme elde edilebilir.

Hiyerarşik kümeleme yöntemleri, kümelerin bir bütün olarak incelenerek kademeli olarak içerdiği alt kümelere bölümlendirilmesi veya ayrı ayrı incelenen kümelerin kademeli olarak bir küme biçiminde birleştirilmesi mantığına dayanır. (Beal and Krishnamurthy, 2006). Hiyerarşik kümeleme yöntemleri dendogramın oluşturulma şekline göre birleştirici hiyerarşik kümeleme yöntemleri (agglomerative) ve ayrıştırıcı hiyerarşik kümeleme yöntemleri (divisive) olmak üzere ikiye ayrılır.

Birleştirici hiyerarşik kümeleme yöntemlerinde başlangıçta her bir gözlem, bağımsız bir küme olarak ele alınır ve daha sonra tekrarlı bir biçimde, bütün gözlemleri içeren tek bir küme elde edilene kadar, her bir gözlemin, kendisine en yakın olan gözlem ile bir küme oluşturması sağlanır.

Ayrıştırıcı hiyerarşik kümeleme yöntemlerinde ise, başlangıçta bütün gözlemler tek bir küme olarak değerlendirilir ve daha sonra tekrarlı bir biçimde, bütün gözlemler birbirlerinden bağımsız tek bir küme oluncaya kadar, her bir gözlem, kendisine en uzak olan gözlemden ayrılıp, yeni bir küme oluşturacak şekilde ayrıştırılır.

Hiyerarşik kümeleme sadece benzer ifade desenli genleri bir grupta toplamaz aynı zamanda veri setini grafiksel olarak da modeller. Grafiksel sunum kullanıcının bütün veriyi gözden geçirmesine ve veri dağılımı hakkında bir fikre sahip olmasını sağlar. Fakat veri setindeki ufak bir değişiklik hiyerarşik dendogramda büyük değişikliklere yol açar. Diğer bir dezavantaj ise yüksek hesaplama karmaşıklığıdır.

3. Yoğunluk-tabanlı yaklaşım

Yoğunluk-tabanlı kümeleme yaklaşımı, nesnelerin yoğunluk (density) ve çekim (attraction) kavramları üzerine kurulmuştur. Ana fikir, nesnelerin birbirini çektiği çok boyutlu yoğun alanların kümeleri oluşturduğudur. Yoğun alanların çekirdeklerinde nesneler birbirine çok yakın ve kalabalıktır yani yoğundurlar. Kümelerin çeperlerindeki nesneler, çekirdeğe göre daha seyrek dağılmışlardır.

Başka bir deyişle yoğunluk tabanlı kümeleme, nesne uzayındaki yoğun alanları belirler. Kümeler, seyrek yoğun alanlar ile ayrılmış yüksek yoğun alanlardır. Benzer ifade desenli gen gruplarını tanımlamak için DHC (density-based hierarchical clustering) yöntemi geliştirilmiştir (Jiang et al., 2003). Nesnelerin, çekimi ve yoğunlukları belirlendikten sonra DHC, verilerin küme yapısını iki düzeyli hiyerarşik yapıda organize eder. İlk düzeyde, yoğun alandaki verilerin ilişkisini gösteren çekim ağacı yapılandırılır. Çekim ağacındaki her bir düğüm bir nesneye karşılık gelir ve düğümün üst düğümü o düğümün çekicisini gösterir. En yüksek yoğunluk derecesine sahip düğüm, tek istisnai veri nesnesidir. Bu nesne çekim ağacının kökü haline gelir. Ağacın yapısı, veri kümesi kalabalıklaştıkça ve veri yapısı karmaşıklaştıkca yorumlanabilirlik açısından zorlayıcı olmaktadır. Bu probleme çözüm olması açısından DHC'nin ikinci düzeyinde çekim ağacını yoğunluk ağacı olarak özetler. Yoğunluk ağacındaki her düğüm yoğun alanları simgeler. Başlangıçta tüm veri tek bir yoğun alanmış gibi değerlendirilir ve yoğunluk ağacının kökünü oluşturur. Sonrasında bu yoğun alan belirli bir kritere göre bölünerek bir çok alt-yoğun alan elde edilir. Her alt- yoğun alan kök düğümün çocuk düğümüdür. Bölünme işlemi her bir alt-yoğun alanın bir kümeyi ifade etmesine kadar sürdürülür.

Yoğunluk tabanlı yaklaşım gürültüden, eş-ifade olmuş genlerin ayrışması açısından başarılı yani gürültülü çevreler için dayanıklıdır. Yöntem gömülü kümeleri tanımlarken aykırı değerlerle de başa çıkmaktadır. Ayrıca DHC, gen ifade verilerinin yüksek bağlantılı yapısı için de uygundur. Fakat yoğunluk tabanlı kümeleme teknikleri boyutsal indeks yapısı kullanılmadığında yüksek hesaplama karmaşıklığı yüzünden ve girdi parametresi bağımlılığından sıkıntı yaratmaktadır.

4. Model-tabanlı Yaklaşım

Model-tabanlı yaklaşım, gen ifade verilerindeki küme yapısını modellemek için istatistiksel bir alt yapı kullanır. Verilerin altta yatan olasılık dağılımlarının sonlu karışımdan geldiği varsayılır. Model-tabanlı algoritmalar modelin parametrelerine genellikle beklenti yükseltme (EM) gibi bazı istatistik teknikleri uygulayarak ençok olabilirliği tahmin etmeye çalışmaktadır. Başka bir ifadeyle,

amaç: $L_{mix}(\theta,\tau) = \sum_{i=1}^{k} \gamma_r^i f_i(x_r|\theta_i)$ olabilirliğini ençoklayan model parametresi (θ) ve gizli parametreyi (τ) tahmin etmektir.

$\theta = \{\theta_i | 1 \le i \le k\}$ k -> küme sayısı

$\tau = \{\gamma_r^i | 1 \le i \le k, 1 \le r \le n\}$ n -> veri nesnesi sayısı

γ_r^i -> x_r nin i. kümeye ait olma olasığı

x_r -> veri nesnesi (örn. Bir gen ifadesi deseni)

$f_i(x_r|\theta_i)$ -> x_r 'nin yoğunluk fonksiyonu

(θ) ve (τ) EM algoritması ile tahmin edilmektedir. EM algoritması E (expectation) ve M (maximization) adımları arasında tekrarlardan oluşmaktadır. E adımında τ gizli parametresi veriden o anki tahmin edilen θ değeri ile koşullu olarak tahmin edilir. M adımında ise θ model parametresi tahmin edilir. Verinin tamamı için olabilirliği ençoklayacak şekilde EM algoritması yakınsayınca her bir veri nesnesi bir kümeye en yüksek olasılıkla atanır.

Model-tabanlı yaklaşımın en önemli avantajı i. nesnenin k. kümeye ait olma olasılığının γ_r^i tahmin edilmesidir. Gen ifade verilerinin önemli karakteristik özelliklerinden birisi yüksek bağlılık (highly-connected) sergilemeleridir. Başka bir ifadeyle gen ifade verilerindeki genlerin birden fazla kümeye ait olma durumu vardır. Bu durum, gen ifade verileri için model tabanlı yaklaşımların olasılıksal sonuç verme özelliği, tercih edilme sebebi olmaktadır. Bu yaklaşımda veri kümesinin belirli bir dağılımı olduğu varsayılır fakat bu varsayım her zaman doğru değildir.

5. Çizge-tabanlı yaklaşım

Verilen Z veri kümesi için, yakınlık matrisi Y, $Y[i,j] = yakınlık\ (N_i, N_j)$

ve ağırlık çizgesi $G(V,E)$ den oluşan yakınlık çizgesi oluşturulabilir. Yakınlık çizgesinde her bir veri noktası bir köşeyi göstermektedir. Bazı kümeleme yöntemleri her nesne çiftini, nesneler arası yakınlık değerine göre atanan ağırlıklı bir kenar ile bağlar. Diğer yöntemlerde ise yakınlığı belirli bir eşik değerine göre 0 veya 1 olarak belirler, i. nesne ile j. nesne eğer $Y[i,j]$ 1 değerine sahipse iki düğüm arasında bir kenar çizilmektedir. Çizge teorik kümeleme teknikleri, verileri kümeleme problemini, yakınlık çizgesinde en küçük kesme (minimum cut) veya maksimum klik bulan çizge problemine dönüştürmektedir.

En bilinen iki çizge tabanlı algoritma: CLICK (Shamir and Sharan, 2000) ve CAST (Ben-Dor, 1999) dir. CLICK (CLuster Identification via Connectivity Kernels) yakınlık çizgesinde birbirine sıkı bağlanan bileşenleri bulmayı hedeflemektedir. Kümeleme işlemi, tekrarlı olarak yakınlık çizgesinde minimum kesik bulmaya çalışır, özyinemeli olarak veriyi en küçük kesmedeki bağlı

bileşenler kümelerine böler. Gen ifade verilerinde, eş-ifade olmuş genler birbiri ile büyük oranda kesişen iki kümeye atanabilmektedir. Böyle bir durumda CLICK, kümeleri ayırmayıp, bir tane bağlı bileşen olarak sunabilir. CAST (Cluster Affinity Search Technique), veri noktalarını, her bir kliğin bir kümeyi temsil ettiği alt-çizgelerin kesişmeyen birleşimi olan klik çizgesi (clique graph) olarak sunar. CAST, kullanıcı tanımlı küme sayısına bağlı kalmamasına rağmen iyi global parametre tanımlama da sıkıntı çekmektedir.

Bütün bu yaklaşımların yanı sıra gen ifade verilerinin kümelenmesini gen-tabanlı kümeleme, örnek-tabanlı kümeleme ve altuzay kümeleme olarak üç ayrı sınıfta da incelemek mümkündür (Jiang et al., 2004). Gen-tabanlı kümeleme de genler birer nesne, örnekler (zaman noktası/ hasta-sağlam) de birer özellik (feature) gibi ele alınır. Örnek-tabanlı kümeleme de ise tam tersidir: örnekler nesne, genler özellik olarak ele alınır. Bu iki kümeleme yaklaşımının ayrımı, gen ifade verisi için kullanılan kümeleme işleminin temel karakteristiği tabanlıdır. K-means ve hiyerarşik yaklaşım gibi bazı kümeleme algoritmaları hem genleri kümelemek hem de örnekleri parçalara ayırmak için kullanılabilir. Moleküler biyolojide, "hücre içindeki herhangi bir işlev genlerin küçük bir alt kümesinin katılımı ile gerçekleştirilir ve hücresel işlev sadece küçük bir örnek alt küme üzerinde gerçekleşir" düşüncesi vardır. Bu düşünce ile altuzay kümeleme de genler ve örnekler simetrik olarak ele alınır; gen veya örnek nesne veya özellik olarak kabul edilebilir.

Gen-tabanlı kümeleme de amaç eş-ifade (co-expressed) olmuş genleri birlikte gruplamaktır. Fakat mikrodizinin deneylerinin karmaşık yapısı dolayısıyla gen ifade verisinin sıklıkla yüksek miktarda gürültü içermesi, gen ifade verisinin genellikle birbiri ile bağlı olması (kümelerin genellikle yüksek kesişim oranına sahip olması) gibi karakteristik özelliklerinden ve biyolojik alandan gelen kısıtlamalardan kaynaklanan bazı problemler vardır. Ayrıca mikrodizi verilerinin kullanacak biyologlar arasında genellikle genlerin kümeleri değil, kümenin içindeki ilişki halinde olan genlerin veya kümeler arasındaki ilişki daha favori bir konudur. Yani algoritmanın sadece kümelemesi değil grafiksel sunum yapması da önemlidir. Kmeans, Self-organizing haritalar (SOM), Hiyerarşik kümeleme, çizge-teorik yaklaşım, Model-tabanlı kümeleme, yoğunluk- tabanlı yaklaşım (DHC) gen-tabanlı kümeleme algoritmalarına örnek olarak verilebilir.

Örnek-tabanlı yaklaşımın amacı, fenotip yapısını veya örneğin alt-yapısını bulmaktır. Yapılan çalışmalar (Golub et al., 1999) örneklerin fenotipleri sadece ifade düzeyleri küme ayrımı ile yüksek korelasyona sahip olan küçük gen alt kümeleri ile ayrıştırılabilir. Bu genlere tanıtıcı gen (informative gene) denir. İfade

matrisindeki diğer genlerin örneklerin ayrışmasında bir rolü yoktur ve verisetinde gürültü olarak değerlendirilirler. Kmeans, Self-organizing haritalar (SOM) (Kohonen 1984), Hiyerarşik kümeleme gibi geleneksel kümeleme algoritmaları tüm genleri özellik olarak alıp örnek kümelemeye doğrudan uygulanabilir. Tanıtıcı genlerin, ilgili olmayan genlere oranı (gürültü oranı) genellikle 1:10 'dir. Bu durum da kümeleme algoritmasının güvenilirliğini zedeler. Bu yöntemler, tanıtıcı genlerin belirlenmesinde kullanılır. Tanıtıcı genlerin seçilmesi denetimli ve denetimsiz (supervised -unsupervised) olarak iki farklı kategoride incelenir.

Denetimli yaklaşım, örneğin "hasta" ve "sağlam" gibi fenotip bilgilerinin eklendiğini durumda kullanılır. Örneklerde bu bilgi kullanılarak sadece tanıtıcı genleri içeren sınıflandırıcı yapılandırılır. Denetimli yaklaşım biyologlar arasında sıkça tanıtıcı genlerin belirlenmesinde kullanılmaktadır. Sınıflandırıcı üç adımda yapılandırılır;

1. Eğitim örneğinin seçimi: Bu adımda eğitim verisinden, örnek sayısının kısıtlı olduğu ($n < 100$) bir alt küme seçilir.
2. Tanıtıcı gen seçimi: bu adımın amacı, ifade deseni ile örneklerin fenotipinde farklılığa sebep olan genlerin seçilmesidir. Komşu analizi, destek vektör makinaları (SVM) ve birçok sıralama (ranking) tabanlı yöntem bu amaçla kullanılmaktadır.
3. Örnek kümeleme ve sınıflandırma: bir önceki adımda belirlenen tanıtıcı genler kullanılarak tüm örnek kümesi kümelenir.

Denetimsiz yaklaşımda ise örneklerin fenotipini belirleyen herhangi bir etiket konulmaz. Etiketin olmaması dolayısıyla da tanıtıcı genlerin kümelere ayrılmaya rehberlik etmemesi, denetimsiz yaklaşımı daha karmaşık hale getirmektedir. Denetimsiz yaklaşımın çözmesi gereken iki problem vardır: gen hacminin oldukça yüksek olmasına karşı örnek sayısının sınırlı olması ve toplanan genlerin büyük çoğunluğunun ilgili olamaması. Denetimsiz yaklaşımdaki bu problemler için iki stratejiden bahsedilebilir: Denetimsiz gen seçimi, ve ilişkili kümeleme. Denetimsiz gen seçiminde, gen seçimi ve örnek kümeleme iki ayrı işlem olarak ele alınır. Öncelikle gen boyutu indirgenir sonrasında klasik kümeleme algoritmaları uygulanır. Eğitim kümesi olmadığı için gen seçimi sadece gen ifade verilerinin varyansını analiz eden istatistiksel modele dayanır. İlişkili kümeleme, dinamik olarak, gen ve örnekler arasındaki ilişkinin kullanılmasını ve tekrarlı olarak kümeleme ile gen seçimi işlemlerini birleştirmeyi destekler. Sezgisel olarak gerçek örnek parçaları bilinmediğinden, her tekrarda hedefe bir adım daha yakın parçaların oluşması beklenir. Yaklaşık parçalar kısmen iyi gen alt kümelerinin seçilmesine olanak sağlar. Birçok tekrardan sonra örnek parçalar

gerçek örnek yapısına yakınsar ve seçilen genler, tanıtıcı gen kümesinin muhtemel adaylarıdır.

Altuzay kümeleme, gen ifadesi verilerine uygulandığında, kümeler genlerin ve deneysel koşulların altkümelerinden oluşan birer "blok" olarak değerlendirilir. Aynı blok içindeki genlerin, o bloktaki koşul altında gösterdiği ifade deseni tutarlıdır. Optimal çözüme yaklaşmak için farklı açgözlü sezgisel yaklaşımlar uyarlanmıştır. Altuzay kümeleme ilk olarak Agrawal ve diğ. tarafından 1998 yılında genel veri madenciliği konusunda önerilmiştir. Altuzay kümeleme de iki altuzay kümesi aynı nesneleri ve özellikleri paylaşabilirken bazı nesneler hiçbir altuzay kümesine ait olmayabilir. Optimal blok seçme problemi NP-harddır ;n gen ve m örnekten oluşan bir ifade matrisi için genlerin ve örneklerin kombinasyonlarının hesaplama maliyeti 2^{n+m} dir. Altuzay kümeleme yöntemleri genellikle hedef bloğu belirlemek için model tanımlar ve sonrasında gen-örnek uzayında arama yapar. Gen ifade verileri için önerilen alt uzay kümele yöntemlerine Biclustering (Cheng and Church 2000), CTWC (Coupled two way clustering) (Getz et. al.,2000) ve Plaid model (Lazzeroni et. al., 2002) örnek olarak verilebilir.

Farklı kümeleme kriterlerine göre veriler kümelenebilir, gen ifade verilerinde bu kriterler eş-ifade olan gen grupları, aynı fenotipe ait olan örnekler, aynı biyolojik süreçten geçmiş örnekler veya genler olabilir. Fakat farklı kümeleme algoritmalarında aynı kriterler kullanılsa dahi veriler farklı şekillerde kümelenebilir. Bu sebeple birden fazla algoritma denenerek veri dağılımına en uygun olanın seçilmesi gerekmektedir.

Kümeleme doğrulama, farklı kümeleme işlemlerinden elde edilen kümelerin kalitesi ve güvenilirliğini ölçme işlemi olarak tanımlanabilir. Kümeleme doğrulama, kümelerin güvenilirliğine veya küme yapılarının şans eseri oluşmadığını gösteren benzeşime odaklanır.

Genel olarak kümelerin kalitesi, küme tanımlamasına göre homojenlik ve ayrışma terimleri ile ölçülebilir. Başka bir deyişle bir kümedeki tüm nesneler birbiri ile benzemeli diğer kümedeki nesneler ile ayrışmalıdır. Kümedeki nesnelerin benzerliğini ölçen küme homojeniliği için bir çok tanım mevcuttur.

$$H_1(C) = \frac{\sum_{N_i N_j \in C, N_i \neq N_j} Benzerlik\ (N_i, N_j)}{||C||.(||C|| - 1)}$$

Yukarıdaki tanıma göre C kümesinin homojenliği, C kümesi içindeki ortalama ikili nesne benzerliği ile gösterilmektedir. Bazı tanımlamalar ise homojenliği kümenin ağırlık merkezine göre değerlendirmektedir:

$$H_2(C) = \frac{1}{||C||} \sum_{N_i \in C} Benzerlik(N_i, \overline{N})$$

Burada $\overline{N}$, C kümesinin ağırlık merkezini (centroid) ifade etmektedir.

Kümelerin ayrışması (Separation) ise iki kümenin (C_1, C_2) benzer olmaması anlamını taşımaktadır.

$$S_1(C_1, C_2) = \frac{\sum_{N_i \in C_1, N_J \in C_2} Benzerlik\ (N_i, N_j)}{\left|\left|C_1\right|\right|.||C_2||}$$

$$S_2(C_1, C_2) = Benzerlik(\overline{N_1}, \overline{N_2})$$

Homojenliğin ve ayrışmanın tanımları nesnelerin benzerliğine dayandığı için, C kümesinin kalitesi, küme içinde yüksek homojenlik değeri ve kümeler arası düşük ayrışma değeri ile artmaktadır.

Verilen $C = \{C_1, C_2, .., C_K\}$ kümeleri için ortalama homojenlik (Shamir et al. 2000);

$$H_{ort} = \frac{1}{K} \sum_{C_i \in C} \left|\left|C_i\right|\right|. \frac{1}{||C_i||} \sum_{N_{ij} \in C_i} Benzerlik(N_j, \overline{N}_\iota)$$

C_i, i. küme,

N_{ij}, i. kümedeki j. nesne,

$\overline{N}_\iota$, i. kümenin ağırlık merkezi

Ortalama ayrışma (Sharan et al. 2000);

$$S_{ort} = \frac{1}{\sum_{C_i \neq C_j} \left|\left|C_i\right|\right|. ||C_j||} \sum_{C_i \neq C_j} \left(\left|\left|C_i\right|\right|. \left|\left|C_j\right|\right|\right). Benzerlik\ (\overline{N}_\iota, \overline{N}_J)$$

Diğer bir bakış açısı ile kümeleme verilen "zemin gerçeği" (ground truth) üzerine kurulmalıdır. Zemin gerçeği, bilgi alanından gelmelidir, genlerin bilinen fonksiyon ailesi veya dokuların sağlam-kanserli klinik teşhisi gibi. Eğer varolan verinin küme yapısının zemin gerçeği var ise kümeleme işleminin performansı kümeleme sonuçları ile zemin gerçeği karşılaştırılarak ölçülebilir. Örneğin kümeleme sonuçları $C = \{C_1, C_2, \dots, C_p\}$ olsun, n veri nesnesi olmak üzere $n * n$

ikili sistemde bir C matrisi oluşturulur. Bu matriste hücreler, eğer iki nesne aynı sınıfa atanmış ise $C_{ij} = 1$ aksi durumda $C_{ij} = 0$ olarak belirlenir. Benzer şekilde bir de zemin gerçeği matrisi oluşturulur: $Z = \{Z_1, Z_2, \dots, Z_s\}$, iki matris arasındaki benzerlik aşağıda tanımlanan şekilde açığa çıkar:

(N_i, N_j) iki nesne çifti için $C_{ij} = 1,\ Z_{ij} = 1$ olduğu durumda n_{11}

(N_i, N_j) iki nesne çifti için $C_{ij} = 1,\ Z_{ij} = 0$ olduğu durumda n_{10}

(N_i, N_j) iki nesne çifti için $C_{ij} = 0,\ Z_{ij} = 1$ olduğu durumda n_{01}

(N_i, N_j) iki nesne çifti için $C_{ij} = 0,\ Z_{ij} = 0$ olduğu durumda n_{00}

C ve Z nin benzerlik derecesi (Halkidi et al., 2001);

$$Sıralama\ indeksi = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$$

$$Jaccard\ katsayısı = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

$$Minkowski\ ölçütü = \sqrt{\frac{n_{10} + n_{01}}{n_{11} + n_{01}}}$$

Sıralama indeksi ve Jaccard katsayısı C ve Z nin uzlaşma derecesini, Minkowski ölçütü Z de aynı kümedeki toplam nesne çiftlerinin (N_i, N_j) sayısının, uzlaşmazlık oranını ifade etmektedir. Hem Jakard katsayısı hem de Minkowski ölçeğinde n_{00} direk hesaplamaya katılmaz. Bu iki indeks gen-tabanlı kümeleme için daha etkili olabilir. Çünkü nesne çiftlerinin büyük bir kısmı aynı kümeye ait olmayacaklardır ve n_{00} diğer tüm durumlara baskın gelecektir.

Doğrulama indeksleri iki farklı kümeleme sonuçlarını karşılaştırır fakat bu karşılaştırma kümeleme sonuçlarının güvenilirliğini ölçmez. Kümeleme güvenilirliği, kümelerin şans eseri oluşturulmama olasılığıdır. Türetilen kümelerin istatistiksel olarak anlamlı olup olmadığı kümelerin p-değerleri ve tahmin gücü ile ölçülebilir.

2.2.2.1 Mclust (Model-tabanlı Kümeleme)

Mclust, sınıflandırma olabilirliği için hiyerarşik birleştirici yaklaşımı (Murtagh et al., 2000; Banfield and Raftery 1993) ve çok değişkenli mixture modelin ençok olabilirlik tahmini için beklenti yükseltme (EM) algoritması (Celeux and Govaert, 1995; McLachlan and Basford, 1988) kullanır.

EM algoritması, z matrisini i gözlemin k kümesine ait olma durumsal olasılığı "z_{ik} " parametre tahminin yapıldığı E-adımı ile z verildiğinde en çok olabilirlik parametre tahminin yapıldığı M –adımı arasında yinelenir.

Her küme Gauss modeli ile gösterilir.

$$\varphi_k(x|\mu_k, \Sigma_k) = (2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\,(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right\}$$

x, k, Σ_k sırasıyla veri, küme ve kovaryansı gösterir. Kümeler, oval ve μ_k ortalaması merkezlidir. Σ_k, geometrik özellikleri belirler. Her kovaryans matrisi $\Sigma_k = \lambda_k D_k A_k D_k^T$ formunda parametreli eigen değer dekompozisyonudur.

D_k -> eigen vektörünün ortogonal matrisi,

A_k -> elemanları Σ_k 'nın eigen değerleri ile orantılı olan diyagonal matris

λ_k -> sayısal (scalar) bir değerdir.

A_k yoğunluğun şeklini tanımlarken, D_k, Σ_k nın temel bileşenlerini tanımlar. λ_k ilgili elipsoidin hacmini belirler. Dağılımın karakteristiği (yönelimi, hacmi ve şekli) genellikle veriden tahmin edilir. Dağılımın karakteristiğinin kümeler arasında değişimine izin verilir veya tüm kümeler için aynı olması zorlanır.

Model tabanlı hiyerarşik birleştirici yaklaşım, kümeleme için hiçbir bilgi olmadan başlasa bile oldukça iyi bölümlendirme yapma eğilimindedir. Oysaki EM algoritması için başlangıç değerleri kritiktir. Bayesci Bilgi Kriteri (Bayesian Information Criterion- BIC) veriyi temsil eden küme sayısına karar verir (Kass and Raftery, 1995)

Kümelemede karşılaşılan iki temel sorun: kümeleme yönteminin seçimi ve küme sayısının belirlenmesidir. Mixture modelde bu sorun teke indirilir o da modelin seçimidir. Kümelemede model seçimi için Mclust yaklaşımı Bayes faktörü ve ardıl model olasılığı kullanan Bayesci model seçimi tabanlıdır.

Temel fikir eğer birçok model $M_1, M_2, \ldots M_K$ var ise öncül olasılığı $P(M_k)$, $k=1 \ldots K$ (genellikle birbirine eşit alınır) olduğu durumda Bayes teoremine göre: veri D için M_k modelinin ardıl olasılığı, verilen model M_k'nın veri olasılığı ile modelin öncül olasılığı çarpımı ile orantılıdır:

$$P(M_k|D) \propto P(D|M_k).\, P(M_k)$$

Bilinmeyen parametreler olduğunda toplam olasılık kuralına göre $P(D|M_k)$, maksimizasyon yerine integralle hesaplanır. Model seçimi için Bayesci yaklaşım, ardıl olasığına göre karar verir. Eğer öncül model olasılıkları $P(M_k)$ aynı ise en

yüksek integralli benzerliğe sahip model seçilir. İki modeli M_1, M_2 karşılaştırıken Bayes faktörü: iki integrali alınmış benzerliğin oranı olarak hesaplanır. $B_{12}= P(D|M_1)/ P(D|M_2)$, eğer $B_{12}>1$ ise M_1 modeli seçilir. Çizelge 2.1 de Mclust yönteminde bulunan modeller listesi verilmiştir.

Çizelge 2.1 MClust yönteminde bulunan modeller listesi

Tek değişkenli mixture	
"E"	eşit varyans (tek- boyutlu)
"V"	Değişken varyans (tek- boyutlu)
Çok değişkenli mixture	
"EII"	küresel, eşit hacim
"VII"	küresel, eşit olmayan hacim
"EEI"	köşeli, eşit hacim ve şekil
"VEI"	köşeli, değişen hacim, eşit şekil
"EVI"	köşeli, eşit hacim, değişen şekil
"VVI"	köşeli, değişen hacim ve şekil
"EEE"	oval, eşit hacim, şekil ve yönelim
"EVE"	oval, eşit hacim ve yönelim
"VEE"	oval, eşit şekil ve yönelim
"VVE"	oval, eşit yönelim
"EEV"	oval, eşit hacim ve eşit şekil
"VEV"	oval, eşit şekil
"EVV"	oval, eşit hacim
"VVV"	oval, değişen hacim, şekil ve yönelim
Tek bileşenli	
"X"	tek değişkenli normal
"XII"	küresel, çok değişkenli, normal
"XXI"	köşeli çok değişkenli normal
"XXX"	oval çok değişkenli normal

Mclust ile gen ifade verilerinin kümelenmesi

Test verimizdeki tüm genleri aynı anda kümeleyebilmemiz mümkün olmadığı için öncelikle her veri setinden 6 zaman noktası ve 108 er tane gen alınarak bir eğitim verisi oluşturulmuştur.

Veri setindeki deneylerden 6 zaman noktası seçerken verinin yapısının bozulmaması amacıyla tüm deneylere aynı yöntem uygulanmıştır. Örneğin 4 zaman noktası (*tp*) ve 3 tekrar (*rep*) içeren deneyler için 12 sütunlü matrisin sütunları $\{t_1, t_1, t_1, t_2, t_2, t_2, t_3, t_3, t_3, t_4, t_4, t_4\}$ şeklinde olacaktır. Burada deneyi 6 sütuna indirirken $\{t_1, t_1, t_2, t_2, t_3, t_4\}$ yöntemi uygulanmıştır. Veri setinde bulunan diğer deneyler için uygulanan yöntemler aşağıda verilmiştir. Hali hazırda 6 zaman noktası olan deneyler olduğu gibi alınmıştır.

3 tp, 3 rep $\{t_1, t_1, t_1, t_2, t_2, t_2, t_3, t_3, t_3\} \rightarrow \{t_1, t_1, t_2, t_2, t_3, t_3\}$

8 tp, 1 rep $\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\} \rightarrow \{t_1, t_2, t_3, t_4, t_5, t_6\}$

3 tp, 1 rep $\{t_1, t_2, t_3\} \rightarrow \{t_1, t_1, t_2, t_2, t_3, t_3\}$

4 tp, 1 rep $\{t_1, t_2, t_3, t_4\} \rightarrow \{t_1, t_1, t_2, t_2, t_3, t_4\}$

2 tp, 2 rep $\{t_1, t_1, t_2, t_2\} \rightarrow \{t_1, t_1, t_1, t_2, t_2, t_2\}$

6 tp, 2 rep $\{t_1, t_1, t_2, t_2, t_3, t_3, t_4, t_4, t_5, t_5, t_6, t_6\} \rightarrow \{t_1, t_2, t_3, t_4, t_5, t_6\}$

6 tp, 3 rep $\{t_1, t_1, t_1, t_2, t_2, t_2, t_3, t_3, t_3, t_4, t_4, t_4, t_5, t_5, t_5, t_6, t_6, t_6\} \rightarrow \{t_1, t_2, t_3, t_4, t_5, t_6\}$

3 tp, 4 rep $\{t_1, t_1, t_1, t_1, t_2, t_2, t_2, t_2, t_3, t_3, t_3, t_3\} \rightarrow \{t_1, t_1, t_2, t_2, t_3, t_3\}$

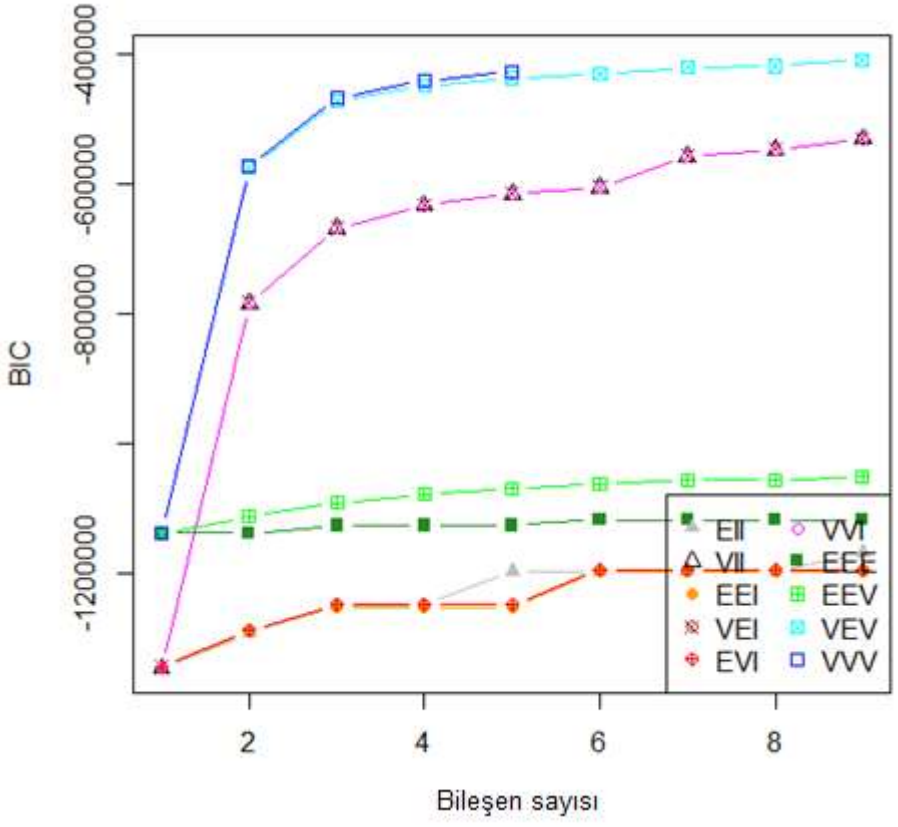
9 tp, 2 rep $\{t_1, t_1, t_2, t_2, t_3, t_3, t_4, t_4, t_5, t_5, t_6, t_6, t_7, t_7, t_8, t_8, t_9, t_9\} \rightarrow \{t_1, t_2, t_3, t_4, t_5, t_6\}$

Eğitim verisi Mclust fonksiyonu ile kümelenmiştir. Kümeleme sonucu fonksiyon aşağıdaki bilgileri (Şekil 2.4) vermiştir. Buna göre elimizdeki veri, eşit şekilli elips (VEV) modeline uygun 9 kümeye ayrılmıştır.

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 9 components:

 log.likelihood     n  df       BIC       ICL
      -202948.4 11652 211 -407872.4 -409416.7

Clustering table:
   1    2    3    4    5    6    7    8    9
1017  804 2420  511 1308  977   74 1501 3040
```

Şekil 2.4 Eğitim verisi için Mclust sonuçları

Mclust içerisinde tanımlı diğer modellere ve küme sayısına bakıldığında verimiz için en uygun modelin (maksimum BIC değeri) VEV olduğu görülmektedir (Şekil 2.5)



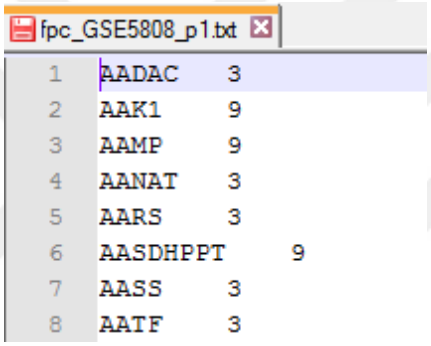
Şekil 2.5 Eğitim verisi için Mclust fonksiyonunun Model seçimi

Mclust fonksiyonu her bir gen için hangi kümeye ait olabileceği ile ilgili olasılık hesaplamaktadır. Örneğin Şekil 2.6 da GSE5808 verisinin ilk 6 geninin 9 küme için olasılık tablosu verilmiştir.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
GSE5808_p1AADAC	5.459128e-09	3.571337e-10	9.990741e-01	2.414914e-24	2.602580e-14	1.015830e-20	2.351968e-30	2.163908e-17	5.925851e-03
GSE5808_p1AAK1	1.809749e-101	3.428114e-07	3.637350e-05	2.287405e-21	2.476085e-11	9.626675e-18	2.227844e-27	2.052085e-14	9.999633e-01
GSE5808_p1AAMP	8.856204e-10	2.414540e-10	9.957005e-01	1.631102e-24	1.758381e-14	6.862450e-21	1.588587e-30	1.461778e-17	4.299496e-03
GSE5808_p1AANAT	1.764557e-27	1.148200e-09	9.838877e-01	7.734774e-24	8.346273e-14	3.253940e-20	7.532777e-30	6.932901e-17	1.611226e-02
GSE5808_p1AATF	2.349611e-139	1.715493e-07	7.755691e-02	1.142674e-21	1.239562e-11	4.809718e-18	1.112700e-27	1.026340e-14	9.224429e-01
GSE5808_p1ABCA3	0.000000e+00	9.999028e-01	2.253766e-164	9.195714e-15	9.716052e-05	3.866976e-11	8.964662e-21	8.235669e-08	4.373759e-16

Şekil 2.6 Küme olasılıkları

Eğitim verilerinden alınan sonuçlara göre veri setinde bulunan tüm deneyler (her bir deneyde 7078 adet gen ve 6 zaman noktası mevcuttur), Mclust fonksiyonu kullanılarak kümelenmiştir. Sonuçlar her bir deney için gen sembolü ve en yüksek olasılıklı küme numarası olacak şekilde " fpc_" ön isimli dosyalara kaydedilmiştir (Şekil 2.7).



Şekil2.7 fpc dosya içeriği

Tüm fpc dosyaları kullanılarak benzerlik matrisi oluşturulmuştur. Benzerlik matrisi oluşturulurken benzerlik ölçütü olarak iki kategorize veriyi karşılaştırırken çakışma (overlap) yöntemi uygulanmıştır. Öyle ki, iki fpc de aynı genlere karşılık gelen kümeler aynı ise bu bir çakışma olarak belirlenmiştir. Çakışma sayısına göre benzerlik Benzerlik = toplam çakışma sayısı / toplam gen sayısı formülüne göre hesaplanmıştır. Benzerlik matrisi oluşturulduktan sonra her bir deneye ait ROC skoru ve AP değerleri hesaplanmıştır.

2.2.2.2 Funfem

Funfem (Model-tabanlı Kümeleme) (Bouveyron et al., 2014) Ayrıştırıcı fonksiyonel mixture model tabanlı bir kümeleme yöntemidir. Funfem, eğrileri ortak ve ayrıştırıcı fonksiyonel alt uzayda modelleyerek, zaman-serileri verilerinin, daha genel bir ifadeyle fonksiyonel verinin kümelenmesini sağlar.

Funfem algoritması, ortak ve ayırt edici fonksiyonel alt uzayda eğrileri modelleyerek fonksiyonel verileri kümeler. Gözlemlenen eğriler $\{x_1, x_2 \dots x_n\}$ ise Funfem, bunları K homojen grupta kümeler. Gözlemlenmeyen rasgele değişkenler Z=$\{z_1, z_2 \dots z_n\} \in \{0,1\}^k$ olduğunu varsayarsak, x eğer $k.$ gruba ait ise Z_k=1 değilse 0 olarak belirlenir. Kümeleme adımının amacı her gözlemlenen x_i eğrisi için z_i=(z_{i1},… z_{ik}) tahmin etmektir. Tanımlanan 12 model için benzerliği maksimum yapan modelin belirlenmesi oldukça zordur. Bunun için Fisher-EM algoritması kullanır (Bouveyron et al., 2012) ve hangi modelin en uygun olduğunu belirler.

Fisher-EM algoritması, ayrıştırıcı uzay ve mixture model parametrelerini tahmin eder ve EM algoritması tabanlıdır. E ve M adımı arasına bir adım daha ekler: F- adımı, amacı projeksiyon matrisi (projection matrix) U'yu hesaplamaktır.

E-adımı: Gözlemin k. gruba ait olma ardıl olasılığını hesaplar.

F-adımı: Fisher kriterini maksimize ederek ayrıştırıcı saklı uzayın yönelim matrisi $U^{(q)}$'nun durumsal ardıl olasılığını tahmin eder.

M-adımı: Saklı alt uzayda, durumsal beklentinin log-olabilirliği (log-likelihood) maksimize ederek mixture model parametreleri tahmin edilir.

Fisher–EM algoritması Aitken kriteri (Bouveyron and Brunet, 2012) sağlanana kadar tekrarlı olarak parametreleri günceller. Saklı alt uzayda az boyutlu ve tüm gruplar için ortak olduğundan kümelenen veri kolayca görüntülenebilir.

Funfem, küme sayısına karar verirken, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) ve ICL (Integrated Completed Likelihood) değerlerinden birini seçmeye imkan verir. BIC kriterinde hata terimi: $\frac{\gamma(M)}{2}\log(n)$, AIC kriterinde $\gamma(M)$, ICL kriterinde $\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}\log(t_{ik})$ dir. Burada $\gamma(M)$, M modelindeki parametre sayısını, n gözlem sayısını, K küme sayısını gösterir. t_{ik} ise $i.$ gözlemin $k.$ kümeye ait olma ardıl olasılığını ifade eder.

Funfem yönteminde 12 model (Discriminative Functional Mixture (DFM) models) vardır (Bouveyron and Brunet, 2012) (Çizelge 2.2).

Çizelge 2.2 DFM modelleri ve *d=K-1* olduğunda kovaryans matrisindeki serbest parametre sayısı

DFM-Model	Σ_k	β_k	Varyans parametre sayısı
$\Sigma_k\beta_k$	Serbest	Serbest	*(K-1)(p-K/2) +K²(K-1)/2+K*
$\Sigma_k\beta$	Serbest	Ortak	*(K-1)(p-K/2) +K²(K-1)/2+1*
$\Sigma\beta_k$	Ortak	Serbest	*(K-1)(p-K/2) +K(K-1)/2+K*
$\Sigma\beta$	Ortak	Ortak	*(K-1)(p-K/2) +K²(K-1)/2+1*
$\propto_{kj}\beta_k$	Diyagonal	Serbest	*(K-1)(p-K/2) +K²*
$\propto_{kj}\beta$	Diyagonal	Ortak	*(K-1)(p-K/2) +K(K-1)+1*
$\propto_k\beta_k$	Küresel	Serbest	*(K-1)(K-1)(p-K/2) +2K*
$\propto_k\beta$	Küresel	Ortak	*(K-1)(p-K/2) +K+1*
$\propto_j\beta_k$	Diyagonal& ortak	Serbest	*(K-1)(p-K/2) +(K-1)+K*
$\propto_j\beta$	Diyagonal& ortak	Ortak	*(K-1)(p-K/2) +(K-1)+1*
$\propto\beta_k$	Küresel & ortak	Serbest	*(K-1)(p-K/2) +K+1*
$\propto\beta$	Küresel & ortak	Ortak	*(K-1)(p-K/2) +K+2*

Modeli belirleyen parametreler :

Σ_k : *k.* grubun kovaryans matrisi

β_k: *k.* grubun gürültü varyans matrisi

Σ_k = diag($\propto_{k1}$,…., $\propto_{kd}$) her grup için kovaryans matrisinin diyagonal olduğu durum

Funfem, 12 model, 3 farklı model seçme kriteri (BIC, AIC, ICL) ve 3 farklı (hclust, random, kmeans) başlangıç ile verimize en uygun olanlarına karar verir.

Funfem ile Genlerin Kümelenmesi

Test verimizdeki tüm genleri aynı anda kümeleyebilmemiz mümkün olmadığı için öncelikle her veri setinden 6 zaman noktası ve 70 er tane gen alınarak bir eğitim verisi oluşturulmuştur. Bu eğitim verisi Funfem fonksiyonunun tüm seçenekleri ile kümelenmiştir. Kümeleme sonucu fonksiyon aşağıdaki bilgileri (Şekil 2.8) vermiştir. Buna göre en iyi sonuç, hclust tabanlı,

$\Sigma_k\beta_k$ modelinde ve 4 küme olduğunda alınmıştır (The best model is DkBk with $K = 4$ (bic = -152654.5)). Hclust, R kütüphanesinde tanımlı, n sayıda nesnenin farklılıklarını kullanarak hiyerarşik kümeleme yapan bir fonksiyondur. Hclust ile başlangıçta her nesne kendi kümesine atanır, sonra algoritma tekrarlı olarak her adımda tek benzer bir küme olana kadar, en çok benzeyen iki kümeyi birleştirir.

```
>> K = 4
DkBk    :       bic = -152654.5
DkB     :       bic = -224074.5
DBk     :       bic = -257110.7
DB      :       bic = -321082.5
AkjBk   :       bic = -158019.3
AkjB    :       bic = -234338.3
AkBk    :       bic = -168207
AkB     :       bic = -247110.8
AjBk    :       bic = -266210.5
AjB     :       bic = -339530.6
ABk     :       bic = -273281.1
AB      :       bic = -340090.4

The best model is DkBk with K = 4 ( bic = -152654.5 )
```

Şekil 2.8 Eğitim verisi için Funfem sonuçları

Eğitim verilerinden alınan sonuçlara göre veri setinde bulunan tüm deneylerdeki genler, Funfem fonksiyonu kullanılarak kümelenmiştir. Burada Mclust yönteminden farklı olarak deneylerdeki zaman noktaları sabitlenmemiştir, her deney kendi özgün hali ile kümelenmiştir. Sonuçlar her bir deney için gen sembolü ve küme numarası olacak şekilde "fpcFUN_" ön isimli parmak izi dosyalarına kaydedilmiştir.

Tüm fpcFUN dosyaları kullanılarak benzerlik matrisi oluşturulmuştur. Benzerlik matrisi oluşturulurken benzerlik ölçütü olarak Mclust'ta olduğu gibi çakışma (overlap) yöntemi uygulanmıştır. Benzerlik matrisi oluşturulduktan sonra her bir deneye ait ROC skoru ve ortalama duyarlılık (Average Precision -AP) hesaplanmıştır.

2.2.2.3 K-means Kümeleme Algoritması

Kümeleme problemini çözen en basit gözetimsiz öğrenme algoritmalarından biri Mac Queen tarafından 1967 yılında önerilen k-means algoritmasıdır. K-means, bilimsel ve endüstriyel uygulamalarda, veri madenciliği alanında en

yaygın olarak kullanılan parçalı kümeleme algoritmaları arasındadır (Syed, 2004), (Berkhin, 2006). K-means algoritmasında amaç, n tane gözlemden oluşan bir veri setini, parametre olarak verilen k adet farklı kümeye parçalamaktadır. Burada, parçalama işlemi sonunda elde edilen kümelerin iki önemli niteliği vardır: (i) küme içi benzerlikleri maksimum (ii) kümeler arası benzerlikleri minimum olmalıdır. Küme benzerliği, kümenin merkezi olarak kabul edilen bir nesne ile o kümedeki diğer nesneler arasındaki mesafelerin ortalama değeridir (Miller and Han, 2001).

K-means algoritması sırasıyla şu adımları gerçekleştirir:

1. Adım: Küme merkezleri iki farklı yöntem ile belirlenir. Birinci yöntemde *k* (küme sayısı) tane rasgele nokta gözlemler arasından seçilir. İkinci yöntemde ise merkez nokta, tüm gözlemlerin ortalaması alınarak belirlenir.

2. Adım: Her bir gözlem için belirlenen merkez noktalara olan mesafesi hesaplanır. Hesaplanan sonuçlara göre tüm gözlemler *k* tane kümeden kendilerine en yakın olan kümeye atanır. Bunun için dört farklı mesafe ölçüm yöntemi vardır.

- Öklit Uzaklığı-(Euclidean Distance) ve Öklit Uzaklığının Karesi (Squared Euclidean Distance): Öklit uzaklığı ve Öklit uzaklığının karesi formülleri ile işlenmemiş verilerle hesaplanır. Öklit uzaklıkları kümeleme analizine sıra dışı olabilecek yeni nesnelerin eklenmesinden etkilenmezken boyutlar arasındaki ölçek farklılıkları Öklit uzaklıklarını önemli ölçüde etkilemektedir. Öklit uzaklık en sık kullanılan uzaklık hesaplama yöntemidir.

  Öklit uzaklığı : $Mesafe\ (x, y) = \sqrt{\sum_i (x_i - y_i)^2}$

  Öklit uzaklığının karesi formülü : $Mesafe\ (x, y) = \sum_i (x_i - y_i)^2$

- Manhattan Uzaklık Formülü (City-block (Manhattan) Distance):

  Boyutlar arasındaki ortalama farktır. Bu ölçüt ile farkın karesi hesaplanmadığı için sıra dışılıkların etkisi azdır.

  Manhattan uzaklığı: = $Mesafe\ (x, y) = \sum_i |x_i - y_i|$

- Chebychev Uzaklığı (Chebychev Distance): İki gözlem arasındaki mutlak maksimum uzaklıktır.

  Chebychev uzaklığı: $Mesafe\ (x, y) = max|x_i - y_i|$

3. Adım: Kümelerin belirlenen merkez noktaları o kümedeki tüm gözlemlerin ortalama değeri ile güncellenir.

4. Adım: Merkez noktalar güncellenmeyene kadar 2. ve 3. adımlar tekrar edilir.

Kısaca, Kmeans algoritması verilen k sayısı için aşağıdaki amaç fonksiyonunu optimize ederek verileri k adet ayrık kümeye ayırır.

$$E = \sum_{i=1}^{k} \sum_{O \in C_i} |O - \mu_i|^2$$

Burada O, C_i kümesindeki veri nesnesi ve μ_i, C_i kümesindeki nesnelerin ortalamasını ifade etmektedir. Amaç fonksiyonu E, nesnelerin küme merkezlerine olan uzaklıklarının karelerinin toplamını en aza indirmeye çalışmaktadır.
Kmeans algoritması basit ve hızlıdır, zaman karmaşıklığı l, iterasyon sayısı, k küme sayısı olduğu durumda $O(l*k*n)$ dir. Fakat gen –tabanlı kümeleme için birçok dezavantajı vardır. Öncelikle, gen küme sayısı gen ifade verisi için belirsizdir. Kullanıcılar optimal küme sayısına ulaşmak için algoritmayı farklı küme sayısı değerleri için çalıştırarak kümeleme sonuçlarını karşılaştırmaktadırlar. Binlerce gen ölçümü içeren büyük gen ifade verileri için bu yöntem pratik değildir. Diğer bir problem de gen ifade verisinin büyük miktarda gürültü içermesidir. Fakat Kmeans algoritması her geni bir kümeye atamaya çalışır ve bu algoritmanın gürültüye karşı duyarlı olmasına sebep olur.

Kmeans algoritması ile Genlerin Kümelenmesi

Gen ifadesi verilerini Kmeans algoritması ile kümeleyerek her bir deney için bir parmak izi dosyası oluşturulmuştur. Verileri kümelerken daha önce ROC ve MAP değerlerine göre en iyi sonuçları veren Funfem algoritması tarafından hesaplanan küme sayısı k=4 olarak alınmıştır.

2.3 Öznitelik Seçimi

Makine öğrenmesinın en temel problemlerinden biri girdi $X = \{x_1, x_2, .., x_N\}$ ve çıktı Y arasındaki fonksiyonel ilişkinin $f()$ tahmin edilmesidir. $\{X_i, Y_i\}$, $i = 1, .., M$ genellikle X_i gerçek sayılar vektörü, Y_i de gerçek sayıdır. Bazı durumlarda Y çıktısı, girdi öznitelikleri $\{x_1, x_2, .., x_N\}$ yerine sadece öznitelik alt kümesi $\{x_{(1)}, x_{(2)}, .., x_{(n)}\}$, $n<N$ tarafından tahmin edilebilir. Girdi ve çıktı arasındaki fonksiyonel ilişki tahmin edilirken yeterli veri ve zaman var ise ilişkili olmayan öznitelikler dahil tüm girdi özniteliklerinin kullanılması uygundur. Fakat pratikte ilişkisi olmayan özniteliklerin makina öğrenmesi

sürecine dahil edilmesi ile iki problem açığa çıkmaktadır: (i) İlişkili olmayan öznitelikler, hesaplama maliyetini arttırmaktadır. (ii) İlişkili olmayan girdi özniteliklerinin hesaba katılması, sonuçların birden fazla modele uygun olmasına ve yanlış değerlendirmesine sebebiyet verebilir. Örneğin, hastalığın belirtileri ve teşhisi arasındaki ilişkinin konulacağı tıbbi bir araştırmaya, yanlışlıkla bir girdi özniteliği olarak hastanın tanıtıcı bilgisi (ID) eklenir ise, ID ile hastalığın tanımlanması gibi yanlış bir sonuç açığa çıkabilir. Bunların yanı sıra girdi ile çıktı arasında bir bağlantı kurmayı hedeflerken, çıktı üzerinde az etkisi olan özniteliklerin ayıklanması sonuca daha verimli ulaşılmasına yardımcı olacaktır. Öznitelik seçimi üzerinde makine öğrenme ve istatistikte uzun yıllardır çalışma yapılmaktadır. Fakat veri boyutunun, artması ile veri madenciliğinde arama konularının artması ile daha popüler olmuştur. Öznitelik seçimi; Fitreleme, altküme seçimi olarak da isimlendirilir. Kısaca öznitelik seçiminin amacı, $X = \{x_1, x_2, .., x_N\}$ girdi öznitelik kümesi nin tamamı arasındanY çıktısını tahmin edebilecek X_s altkümesi seçmektir.

Öznitelik seçimi tüm öznitelikler içinden azaltılmış ve muhtemelen daha iyi sınıflandırma performansına sahip bir öznitelik alt kümesinin bulunmasıdır. Sınıflandırıcıya verilecek öznitelik sayısı, tüm değişkenleri denemeye elverişli değilse, sınıflandırıcının işini kolaylaştıracak bir ön işleme gerek duyulur. Bu şekilde sınıflandırma gücü olmayan öznitelikler ayıklanabilir. Öznitelik uzayının boyutu arttırması sınıflandırıcının başarısını azaltıcı etki yapar. Bu nedenlerden öznitelik seçme yöntemleri kullanılır.

Burada öznitelik seçim yöntemleri ile deneyleri diğerlerinden daha iyi tanımlayan genler seçilmeye çalışılmıştır. Bunun için birkaç farklı öznitelik seçimi algoritması kullanılmıştır. Bunlar; Relief algoritması, OneR nitelik algoritması ve bilgi kazanımı (Info Gain) algoritmasıdır. Öznitelik seçimi algoritmalarının genel yapısı aşağıda verilmiştir (Molina et al., 2002).

Girdi:

S- *X* öznitelikli veri örneği, |*X*|=*n*

J- maksimize edilecek değerlendirme ölçeği

GS- ardıl üretim operatorü

Çıktı:

Çözüm- ağırlıklandırılmış öznitelik kümesi

L:= Başlangıç noktası (*X*);

Çözüm:= { *J*'e göre en iyi *L*};

Tekrar:

*L:=*Arama_stratejisi *(L, GS(J),X);*

X' :={ J'e göre en iyi *L };*

EĞER *J(X')>= J(*çözüm*) || J(X') = J(*çözüm*)*

ve |*X'*|<*(*|çözüm|*)*

İSE Çözüm*:=X';*

Kadar Dur (*J,L*)

Öznitelik seçimi yöntemlerinin tümü WEKA (Waikato Environment for Knowledge Analysis) makine öğrenme araç seti üzerinde ve özel bir parametre optimizasyon yöntemi kullanılmadan gerçeklenmiştir.

2.3.1 Relief algoritması

Özniteliklerin değerini, aralarında bulunan bağımlılıkları ortaya çıkarmaya çalışarak bulmayı amaçlar. Bunu yapmak için, özelliğin bulunduğu örneğin ait olduğu ve olmadığı sınıflarda bulunan en yakın örnekleri karşılaştırarak ağırlıklandırır (Molina et al., 2002).

X örneğindeki *A* özelliği için *W*[*A*] ağırlığı:

$$W[A] = W[A] - \frac{diff(A,X,H)}{m} + \sum_{C \neq sınıf(X)} \frac{\left[P(c)x\ diff\left(A,X,M(C)\right)\right]}{m}$$

formülü ile hesaplanır. Burada *H*, ait olduğu sınıftaki en yakın örnek, *M*(*C*), ait olmadığı *C* adet sınıftaki en yakın örnekler, *m*, normalizasyon katsayısı, *diff* ise iki örnek arasındaki fark fonksiyonudur.

2.3.2 Bilgi kazanımı algoritması

Verilen bir öznitelik için verilen sınıflandırma sonuçlarının ne kadar değer ile kazanılabileceğini gösterir (Information Gain Attribute Evaluation-InfoGain). Bilgi kazanımı entropinin tersidir ve 0 ile 1 arasında değer alır. Sınıflandırılan bir verinin, verilen bir özelliğine göre kazandığı değeri göstermektedir. Örneğin bir sınıflandırma sonucunda 5 farklı sınıf oluşturuldu ve her bir sınıf için ayrı bir özellikten bahsedilebiliyorsa bu durumda bilgi kazanımı 1 olmalıdır. Yani sınıf ile özellik arasında birinci derece bir ilişkiden bahsedilebilmektedir. Bilgi

kazanımının sıfıra yakın bir değer olması, elde edilen sınıfların, özelliklerle ilişkili olmadığını göstermektedir.

Bilgi kazanımı (*IG(X)*), *X* özniteliği için aşağıdaki formül kullanılarak hesaplanır. (Novakoviç, 2009)

$$IG(X) = H(S) - \sum_{t \in T} p(t)H(t)$$

H(S) = *S* kümesinin entropisi

t = *S* hariç, *X* özniteliğin altkümesi

p(t)= t deki eleman sayısı/ *S* kümesinin alt kümesi

H(t)=*t* alt kümesinin entropisi

2.3.3 OneR (One-attribute-rule) nitelik algoritması

Holte tarafından geliştirilen basit bir algoritmadır (OneR Atribute Algorithm) (Holte, 1993). Algoritma girdi olarak birçok farklı öznitelikte ve sınıftaki veri örneği kümesini girdi olarak alır. Çıktı olarak, verilen örnek kümesinden bir kural çıkarır. Bu kural, verilen sınıfı en hatasız tahmin eden tek bir öznitelik tabanlıdır. Özünde tek seviyeli karar ağacı oluşturulur. Algoritmanın temel hedefi en az hata oranına sahip özniteliği belirlemektir. Öncelikle özniteliği alır, bu özniteliğin herbir değerini, bu değerin görüldüğü en sık olan sınıfa atar. Bu öznitelik değerinin sahip olduğu hata sayısı farklı bir sınıfa karşılık gelir. Öznitelik için tüm değerlerin toplam hataları hesaplanır. En az toplam hataya sahip olan öznitelik one-r kuralı olarak belirlenir. Tüm numerik değerli özelliklerin sürekli olduğunu varsayar ve birçok ayrık aralıklara böler.

OneR algoritmasının sözde kodu (Mahmood and Lei, 2009);

1- her bir öznitelik için

2 -özniteliğin her bir değeri için, aşağıdaki gibi kurallar oluştur

3 -her sınıfın görülme sıklığını hesapla

4 -en sık görülen sınıfı bul

5 -kuralı oluştur ve bu öznitelik-değerini sınıfa ata

6 -kuralın doğruluğunu hesapla

7 -en yüksek doğruluğa (en düşük hata oranı)sahip kuralı seç.

2.4 Benzerlik Ölçütleri ve Değerlendirilmesi

Bu bölümde iki veri setinin benzerliğini değerlendirmek için kullanılmış yöntemlerden bahsedilecektir.

Benzerlik; andırış (resemblance), benzeyiş (likeness), iki nesnenin eşitliği olarak bilinir. İkiden fazla eleman karşılaştırıldığında veya bir eşik belirlendiğinde daha anlamlı hale gelir. Benzerliği ölçmek için benzeyişin niceliksel olarak belirlenmesi gerekir. Bunun için korelasyon ortak kullanılan yöntemdir (Moller-Levet et al., 2003).

Korelasyon katsayısı, bağımsız değişkenler arasındaki ilişkinin yönü ve büyüklüğünü belirten katsayıdır.

Korelasyon katsayısı -1 ile +1 arasında bir değere sahiptir. Sıfırdan büyük değerler direk yönlü doğrusal ilişkiyi; küçük değerler ise ters yönlü doğrusal ilişkiyi gösterir. Korelasyon katsayısının sıfıra eşit olması belirtilen değişkenler arasında doğrusal bir ilişki olmadığını gösterir. Korelasyon katsayısının 1'e yakınlık derecesi, ilişkinin doğrusallığı ile doğru orantılıdır.

Bu bölümde daha önce iki ifade profilinin benzerliğini karşılaştırmak için kullanılan Pearson korelasyon katsayısı, Spearman sıralama korelasyon katsayısı Tanimoto katsayısı ve Öklid uzaklığı hakkında kısaca bilgi verilecektir.

2.4.1 Pearson korelasyon katsayısı

Bir rasgele örneklem olarak *n* büyüklükte *X* ve *Y* değişkenleri için sayısal veri serileri bulunmaktadır. Bu seriler *n* satırlı ve 2 sütunlu bir veri matrisi olarak tutulur. Bu veriler *i = 1, 2, ..., n* için x_i ve y_i olarak yazılır ve Pearson korelasyon katsayısı olan r_{xy} şu formül ile hesaplanır:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Burada toplama $\Sigma\, i=1$ ile n arasındadır.

Eğer veriler (x, y) normal dağılımdan gelmişlerse, Pearson'un korelasyon katsayısı bu iki değişken arasında bulunan korelasyon için en iyi korelasyon kestirimi olduğu gösterilmiştir. (Aktürk, 2009 ; Pearson, 1896)

2.4.2 Spearman sıralama korelasyon katsayısı:

İki değişken arasındaki ilişkinin gücünü ölçmek için kullanılan parametrik olmayan bir testtir. Verilerin normal dağılmış olması gerekmez. Veriler aralıklı ya da oranlı ölçüm düzeyinde toplanmışsa önce sıralanır, sonra test uygulanır. Pearson korelasyon katsayısının özel bir halidir. İki değişken ($Y\ ve\ X$) içinde

verilerin sıralanmış olmaları Spearman'ın sıralama korelasyon katsayısı ρ (rho) hesaplanması için gereklidir. Genellikle veriler aralıksal ölçekli, oransal ölçekli, veya sırasal ölçeklidir. Öncelikle sıralama düzeni haline getirilmeleri gerekmektedir. ρ hesabı için sıralanmış x_i ve y_i verileri kullanılır.

Sıralandıktan sonra, iki değişkenin (x_i ve y_i) sıra numaraları arasındaki fark d_i $i=1,...n$ hesaplanır. Bu, tüm karşılıklı veriler ($i=1...n$) için yapılır. Eğer herhangi iki değişkenin sıra numaraları aynı değilse ρ değeri aşağıdaki formül ile hesaplanır:

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

Burada; n iki değişkenli örneklemde toplam gözlem sayısı, $d_i = x_i - y_i$ x_i ve y_i sıra numaraları arasındaki farktır.

Eğer sıra numaraları aynı olan iki veya daha fazla değişken var ise, Pearson korelasyon katsayısı sıralama numaraları veri olarak kullanılarak hesaplanır. Sıra numaraları aynı olan verilere, sıra numaralarının ortalamaları atanır (örneğin 1 2,5 2,5 4). Bu durumda formül şu şekilde olur:

$$\rho = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Spearman'ın ρ katsayısı -1 ile +1 aralığında değer alır. Sınır değerler iki değişken sıralaması arasında bağlantının çok iyi olduğunu gösterir. Eğer $\rho<0$ ise, sıralamalar arasında ters orantılı, Eğer $\rho>0$ ise sıralamalar arasında doğru orantılı değişme görülmektedir. Eğer Spearman'ın ρ katsayısı sıfır ise, sıralamalar arasında hiçbir bağlantı bulunmadığı anlamı taşımaktadır (Spearman, 1904).

2.4.3 Tanimoto katsayısı

İlk olarak 1960 yılında iki görüntü arasındaki benzerliğin hesaplanması çalışmasında kullanılmıştır. İkili tabanda tutulan (1 ve 0) görüntülerin benzerliği karşılaştırılırken iki görüntüdeki aynı koordinatta her bit (&, ||) mantıksal operatörleri ile işleme tabi tutulur ve birbirine oranı alınır. Örneğin X_i, X görüntüsünün i. bitini ve Y_i, Y görüntüsünün i. bitini simgelediği durumda Tanimoto benzerliği

$T_s(X,Y) = fracsum_i(X_i \& Y_i) sum_i(X_i || Y_i)$ şeklinde hesaplanır.

Bu benzerlik katsayısı, Jaccard indeksine benzer şekilde iki görüntünün birleşimi ve kesişimi şeklinde ele alınabilir. Tanimoto katsayısı 0 ile 1 arasında

bir değerdir, değer 1'e yaklaştıkça benzerliğin arttığını göstermektedir.Tanimoto uzaklığı, iki görüntünün birbirine olan uzaklığı olarak ifade edilir;

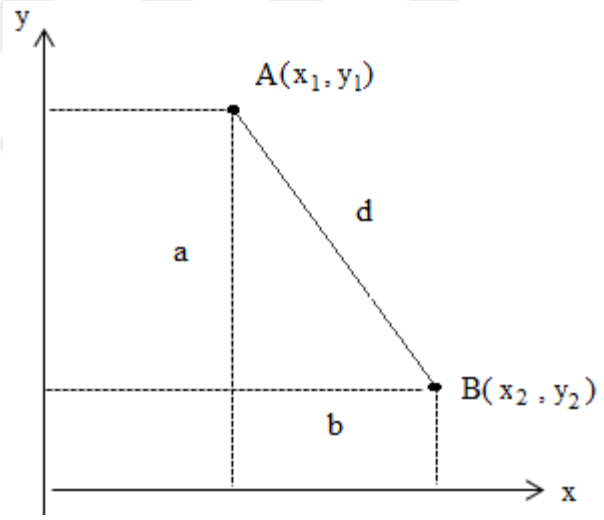
$$T_d(X,Y) = -log_2(T_s(X,Y))$$

Tanimoto benzerliği görüntüler için kullanıldığı kadar metinler içinde kullanılmaktadır. Günümüzde yaygın olarak kimyasal bileşenlerin ikili parmak izi sunumunun analizlerini yaparken kullanılmaktadır. Kısaca Tanimoto katsayısı ikili sunumlar (*binary*) için ortak pozitif bitlerin toplam pozitif bitlere oranı olarak ifade edilebilir (Bell and Sacan, 2012)

$$\tau = \frac{\sum(A^+ \cap B^+)}{\sum(A^+ \cup B^+)}$$

2.4.4 Öklit Uzaklığı

Öklit uzaklığı, iki nokta arasındaki doğrusal uzaklıktır. İki boyutlu uzayda Pisagor teoreminin bir uygulamasıdır (Şekil 2.9). A ve B noktaları arasındaki uzaklık:



Şekil 2.9 Öklit uzaklığı gösterimi

$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Bu durumda n uzunluklu x ve y noktaları arasındaki Öklit uzaklığı d(x,y):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

formülü ile hesaplanır.

2.5 Karmaşıklık Analizi

Önerilen içerik tabanlı arama modeli karmaşıklık açısından parmak izinin oluşturulması ve benzer deneylerin aranması olarak iki bölüme incelenebilir.

Parmak izi dosyasının oluşturulması için öncelikle deneyin platform bağımlı tanımlamalarının gen sembollerine dönüştürülmesi gerekmektedir. Deneydeki genler, test verilerinin alındığı platformların tanımlamaları ve karşılık gelen gen sembolü listesine göre gen sembollerine dönüştürülür.

t-> gen sembolü dizini uzunluğu (136.415)

s ->deneydeki gen sayısı (her deney için farklılık gösterir)

} *O(t.s)*

Gen sembollerine dönüştürülen deney ortak gen kümesine göre tekrar yapılandırılır

n-> ortak gen listesi uzunluğu (7078)→*O(n.s)*

Parmak izi dosyasının oluşturmak için gerekli son adım ise kümeleme; bu çalışmada kümeleme için iki farklı bütünleştirici hiyerarşik model tabanlı kümeleme algoritması kullanılmıştır. Genel olarak kümeleme algoritmasının karmaşıklığı (Zhong and Ghosh, 2003) aşağıda verilmiştir.

Parçalı (partitional) kümeleme yaklaşımı içeren modellerde (Gaussian, Multinomial vb.) model parametrelerinin tahmini kapalı-form çözümlüdür ve tekrarlı süreçlere ihtiyaç duymaz. Her tekrarın zaman karmaşıklığı, atama ve model tahmini adımlarında *N* tane veri nesnesi ve *K* tane küme için doğrusaldır. *M* tekrar sayısı olmak üzere toplam karmaşıklık *O(KNM)* dir. Eğer parametre tahmininin tekrar gerektirdiği modeller (hidden Markov models with Gaussian observation density) kullanılmış ise model tahmini karmaşıklığı, M_1 tekrar sayısı olmak üzere her bir kümeleme tekrarı için $O(KNM_1)$ dir. Bu durumda toplam karmaşıklık $O(KNMM_1)$ dir. Teorik olarak *M* ve M_1 oldukça büyük olabilir. Fakat pratikte EM algoritması oldukça hızlı bir şekilde yakınsar (yaklaşık 30-40 tekrar).

Model tabanlı hiyerarşik birleştirici yaklaşım, i. seviyede i tane model ve en altta *N* küme ile başlar. Küme-içi uzaklık *N*. seviyede $N(N-1)/2$ ve i. seviye için $i-1$ $(i < N)$ olarak hesaplanır. Tüm hiyerarşi için toplam uzaklık sayısı:

$$N(N-1)/2 + (N-2) + (N-3) + \cdots + 1 \rightarrow O(N^2)$$

Aynı mantığı kullanarak gerekli toplam uzaklık karşılaştırma sayısı :

$$N(N-1)/2 + ((N-1)(N-2))/2 + \cdots + 2 * 1/2 \rightarrow O(N^3)$$

Toplam model tahmini karmaşıklığı:

$$N.1M_1 + N.2M_1 + 1.3M_1 + \cdots + 1NM_1 \rightarrow O(N^2M_1)$$

Küme-içi uzaklık karşılaştırmaları karmaşıklığı karşılaştırma sonuçlarının kaydedildiği akıllı veri yapıları kullanılarak $O(N^2\ log\ N)$ 'e düşürülebilir.

Son olarak, sorgu parmak izinin test veri tabanında bulunan parmak izi dosyaları ile karşılaştırılarak arama sonucunun sıralanması:

İki parmak izinin karşılaştırılması (çakışma metriği) p-> test veritabanında karşılaştırılacak parmak izi sayısı (99)→ $p.O(n)$

Sonuçların sıralanması -> $O(p\ log\ p)$

2.6 Değerlendirme Yöntemi

2.6.1 İşletim karakteristik eğrisi (Receiver Operating Characteristic – ROC)

ROC eğrisi; testlerin etkinliklerinin karşılaştırılmasında, ayırt etme yeteneğinin belirlenmesinde, uygun pozitiflik eşiğinin bulunmasına, araştırmanın gelişiminin takip edilmesine imkan verir.

Gerçek hastalara konan tanı açısından; A: Doğru tanıya uygun olarak tanı testinin de hasta dediği gerçek pozitifler (GP) C: tanı testinin yanlış olarak sağlam dediği, gerçekte hasta olan yanlış negatifler (YN)

Gerçek sağlamlara konan tanılar açısından; D: Testin sağlam dediği ve gerçekte de sağlam olan gerçek negatifler (GN) B: Testin hatalı olarak hasta dediği gerçekte sağlam olan yanlış pozitifler (YP)

Duyarlılık (*Sensitivity*): Gerçekte hasta olanlara testin de hasta tanısı koyma yeteneğidir. *DUYARLILIK = A / (A+C) = GP / (GP + YN)*

Özgüllük (Specificity): Gerçekte sağlam olanlara testin de sağlam tanısı koyma yeteneğidir. *ÖZGÜLLÜK = D / (D + B) = GN / (GN + YP)* (Dirican, 2001; Tomak ve Bek,2010)

Bir ROC eğrisi, farklı eşik değerleri için y ekseninde doğru pozitiflik (duyarlılık) ve x ekseninde yanlış pozitiflik (1-özgüllük) oranlarının bulunduğu bir eğridir. Genelde doğru pozitiflik oranı düşük olan eşik değerlerinin, yanlış pozitiflik oranları da düşüktür.

ROC eğrisi altında kalan alan (AUC –area under the curve) ROC puanı olarak tanımlanabilir. ROC puanı 1 ise pozitifler mükemmel bir şekilde negatiflerden ayrılmıştır. ROC puanı sıfır olduğunda ise herhangi bir pozitif bulunamamıştır anlamı çıkarılmaktadır.

Test verilerimiz için kullandığımız ROC algoritması Çizelge 2.3 de verilmiştir (Fawcett, 2006).

Çizelge 2.3 ROC skoru hesaplama algoritması

Girdi: Parmak izi dosyaları
Çıktı: ROC skoru
tp=0 //true pozitiflere başlangıç değerin atanması
fp=0 //false pozitiflere başlangıç değerin atanması
roc=0 // ROC skoruna başlangıç değerin atanması
Her bir sıralı etiket için
{

- **EĞER** Etiket =1 **İSE**
  tp++
  DEĞİLSE
  fp++
  roc=roc+tp

}

- **EĞER** tp=0 **İSE**
  roc=0
  DEĞİLSE EĞER fp=0 **İSE**
  roc=1
  DEĞİLSE
  roc/=tp*fp

Tez çalışmasında ROC skorları hesaplanmadan önce test verilerinin tamamı meme kanseri verisi ise 1 değilse 0 olarak etiketlenmiştir. Etiketlenen değerlere göre doğru pozitif (tp) veya yanlış pozitif (fp) sayıları tespit edilmiştir. Hesaplanan benzerliğe göre büyükten küçüğe sıralanan deneylerin ilk 41 tanesi içinde 0 etiketli deney bulunursa yanlış pozitif olarak sayılmıştır. Sonraki adımda yp ve tp sayılarına göre ROC skorları hesaplanmıştır.

2.6.2 Ortalama Duyarlılık (Mean Average Precision)

Bilgi getirim sistemlerinin (Information Retrieval System) başarımını ölçmek için "Duyarlılık" ve "Anma" olarak iki temel ölçüt kullanılır. Duyarlılık (Precision-P) ölçütü arama sonucunda getirilen belgelerin sorguya uygun olma olasılığını gösterir. Bir başka ifade ile duyarlılık, sorguya benzer belgelerin getirilen toplam belge sayısına oranıdır (Berber ve Alpkoçak, 2009).

$$P = \frac{|\{ilgili\ belge\} \cap \{getirilen\ belge\}|}{|\{getirilen\ belge\}|}$$

Getirilen belgelerden ilk sıralardaki belgelerin sorgunun içeriğine uygun olması bir sistemin değerlendirmesi açısından önem taşımaktadır. Bu nedenle, bir diğer başarım ölçütü de ilk sıralarda getirilen belgelerin duyarlılığıdır (Average Precision-AP). AP, benzerlik sıralama pozisyonlarına göre duyarlılık değerlerinin ortalamasının hesaplanmasıdır.

$$AP = \frac{\sum_{k=1}^{n}(P(k) * rel(k))}{ilgili\ belge\ sayısı}$$

k, getirilen belgenin sırası

n, toplam getirilen belge sayısı

$P(k)$, getirilen sıralı benzer belge listesinde k da kesilen duyarlılık

$rel(k)$, k. sıradaki belge sorgu ile benzerse 1 değilse 0 değerini alır

Ortalama Duyarlılık (Mean Average Precision-MAP) ortalaması ise birden fazla sorgu için sistemin hesaplanan duyarlılık değerlerinin ortalamasıdır ve aşağıdaki şekilde hesaplanır:

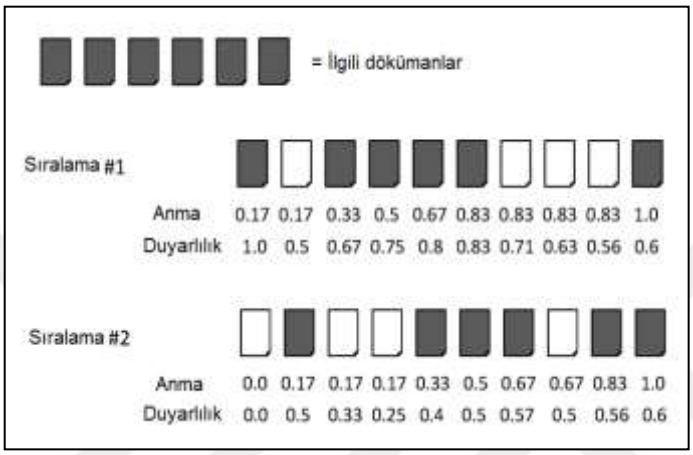
$$MAP = \frac{1}{s}\sum_{i=1}^{s} AP_i$$

s toplam sorgu sayısını, AP_i ise i. sorgunun duyarlılığını gösterir.

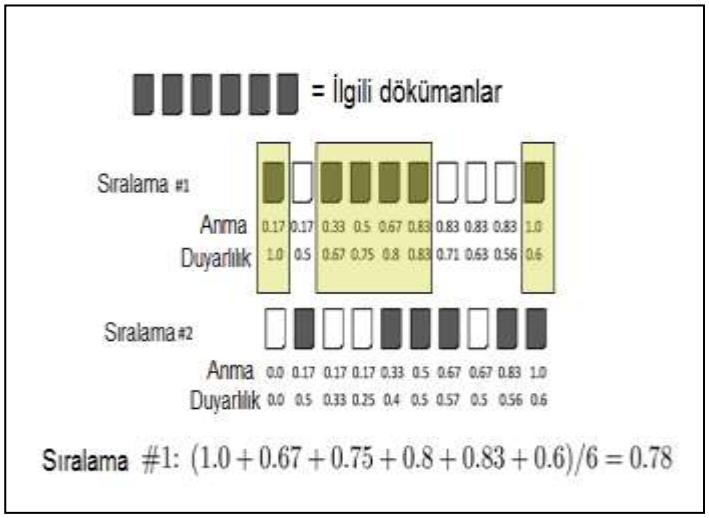
Anma (Recall-R) ölçüsü ise, sorguya uygun belgelerin geri getirilen belgeler arasında bulunma olasılığı olarak ifade edilir.

$$R = \frac{|\{ilgili\ belge\} \cap \{getirilen\ belge\}|}{|\{ilgili\ belge\}|}$$

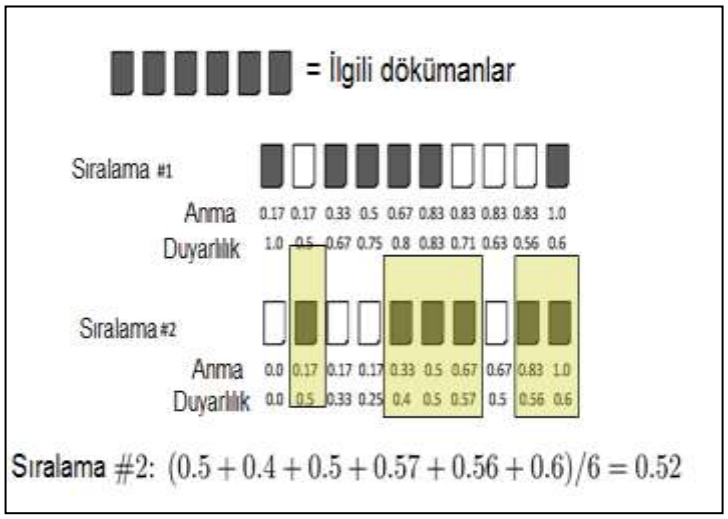
Aşağıdaki örnekte (Şekil 2.10), hesaplanan Duyarlılık değerleri üzerinden AP değerinin hesaplanması gösterilmiştir. Duyarlılık hesaplanırken, toplam benzer belge sayısı n=6 ise sıralama1 deki 3. Belge için: Anma 2/6 = 0.33 ve Duyarlılık 2/3 = 0.67 olacak şekilde hesaplama yapılmıştır.



(a)



(b)



(c)

Şekil 2.10 AP değerlerinin hesaplanması; (a) Örnek iki sıralama için anma ve duyarlılık değerleri (b) Sıralama1 için ortalama duyarlılık hesabı (c) Sıralama 2 için ortalama duyarlılık hesabı

Şekil 2.10 daki her iki sıralama için MAP değeri: (0.78+0.52) /2 = 0.65 olarak hesaplanır. Değerlendirme yapılırken MAP değeri yüksek olan sistemin, daha duyarlı olduğu söylenir. Yukarıdaki örnekte AP değerlerine göre sıralama1 sıralama 2 den daha iyidir sonucuna varılır.

2.6.3 Hipotez testleri

İçerik tabanlı arama sonucunda alınan sonuçların istatistiksel olarak anlamlı olduğunu test etmek gereklidir, bilgi erişim sistemlerinde bu amaçla kullanılan testler: t-test, randomizasyon test (randomization), Anova1 test, wilcoxon test, Bootstrap test ve sign test.

Hipotez, bir durum ile ilgili iddia edilen doğruluğu henüz ispatlanmamış varsayımdır. Hipotezler üzerinde farklı işlemler yapılarak iddianın doğruluğu veya yanlışlığı araştırılır. Önerilen hipotezin kabul edilip edilmeyeceğinin belirlenmesi hipotez testleri ile yapılır. Denemeye veya gözlem sonucu elde edilen sonuçların rastlantısal olup olmadığını inceleyen istatistiksel testlere Hipotez testleri denir. Hipotez testleri, özünde bir seçim ve karşılaştırma işlemidir. Bu sebepler birden fazla hipoteze ihtiyaç duyulur: alternatif hipotez /Null hipotez. Öne sürülen ana iddia null hipotezdir (H_0), bazı kaynaklarda sıfır hipotezi, başlangıç hipotezi olarak da anılır. Yapılan testler sonucunda null hipotezin doğruluğu hakkında emin olunmadığı bir durum ortaya çıktığında karşılaştırmak yapmak için önerilen ikinci hipotez alternatif / araştırma hipotezdir (H_A). İşlemler sonucunda H_0'ın doğruluğunda süphe edilmesi H_A'nın kabulü anlamı taşır. Bizim

için H_0 her iki sistemin aynı olduğudur. Bu durumda H_A iki sistemin farklı olması olarak kurulur. Hesaplanan p-değerleri, risk derecesine göre değerlendirilir. Risk derecesi temelde doğru olan null hipotezinin reddedilme olasılığını gösterir. Yaptığımız çalışmada genel risk derecesi olarak %5=0,05 kullanılmaktadır (Mendeş vd., 2005).

Hipotez testleri parametrik ve parametrik olmayan testler olarak ikiye ayrılır. Parametrik testler verinin normal dağılımdan geldiğini ve gruplar için aynı varyans özelliğine sahip olduğunu varsayar. Eğer bu varsayımlar doğru değilse parametrik olmayan testler uygulanmalıdır. Popüler birçok parametrik testten bazıları; t-test, welch t-test, anova testidir. Parametrik olmayan testler verilerin herhangi bir dağılımdan gelmediğini varsayar. Bazı popüler parametrik olmayan testler; randomizasyon testi, wilcoxon sıralama testidir. Bu testlerden randomizasyon, Anova1 ve t-test anlamlılığı ölçmek için yeterlidir (Bandyopadhyay et al., 2014; Smucker et al., 2007).

2.6.3.1 t-test

En yaygın kullanılan hipotez testidir. İki örneklem ortalamaları arasındaki farkı inceleyen parametrik bir hipotez testidir. t-test ile iki örneklemin ortalamaları karşılaştırılarak, aradaki farkın tesadüfi mi, yoksa istatistiksel olarak anlamlı mı olduğu belirlenir. t-testinin araştımacılar tarafından sık tercih edilmesinin sebebi küçük örneklemlerle (n<30) çalışmaya olanak sağlamasıdır (Seltman, 2015).

Bahsedilen örneklemlerin türüne göre t-testi 2 farklı durumda çalışılabilir: bir-örneklem t-test (one sample t-test) ve iki örneklem t-test (two-sample t-test).

İki örneklem t-testlerini, bağımsız örneklem t-testi (independent samples t-test) ve ilişkili eşli örneklem (paired samples t-test) t-testi olarak incelemek mümkündür.

- İki örneklemin birbiri ile ilişkisiz iki grup olması durumu; bağımsız örneklem t-testi. Standart sapma bilinmiyor ve örnekteki veri sayısı 30 dan az ise t-istatistiği aşağıdaki formül ile hesaplanır.

$$S_p^2 = \frac{((n_1 - 1)s_1^2) + ((n_2 - 1)s_2^2)}{n_1 + n_2 - 1}$$

$$t = \frac{\mu_1 - \mu_2}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

n_i -> i. örnekteki veri sayısı (i=1,2)
μ_i -> i. örnek ortalaması (i=1,2)
S_i^2 -> i. örneğin varyansı (i=1,2)
S_p^2 -> popülasyon varyansının birleştirilmiş tahmini

- Örneklemlerin önce ve sonra olarak gruplandırılması, aynı örneklemin belirli bir olaydan veya dönemden önceki ve sonraki verileri arasındaki farklılıkların incelenmesi durumu; ilişkili eşli örneklem t-testi (paired samples t-test).
  Nesneler arasındaki varyasyonu ortadan kaldırıp iki popülasyonun da normal dağılımlı olduğu varsayılır. Aşağıdaki gibi hesaplanır.

$$\bar{D} = \frac{\sum D}{n}, \quad S_D = \sqrt{\frac{\sum D^2 - \frac{(\sum D)^2}{n}}{n-1}}$$

$$t = \frac{\bar{D}}{S_D / \sqrt{n}}$$

$D = x_1 - x_2$ aynı örneğin 1.durumu ve 2. durumu arasındaki fark
n -> örneklemdeki gözlem sayısı

- Aynı örneklemin belirli bir değişkene ilişkin ölçülen ortalama değeri ile tahmin edilen yada bilinen ortalama değerinin karşılaştırılması durumu; Tek örneklem t-test (one sample t-test).

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

$$t = \frac{X - \mu}{S}\sqrt{n}$$

t-> tek örneklem t-test değeri

S -> Standart sapma,

$\bar{X}$ -> Örnek ortalamsı

μ -> popülasyon ortalaması

n-> örneklemdeki gözlem sayısı

2.6.3.2 Randomizasyon testi (Randomization test/ permutation test)

Randomizasyon testinde, eğer Sistem A ve Sistem B aynı ise bu sistemlerin sonuçlarını üreten bir N sistemi olduğu varsayılmaktadır. Bu sebeple, N sisteminden *k* tane sonuç ürettirilir, yarısının A sisteminden diğer yarısının B sisteminden geldiği şeklinde etiketlenir. Fakat sonuçları yokluk hipotezi altında etiketlemenin 2^k farklı yolu vardır. Yokluk hipotezi altında permutasyonları etiketleme eşit olasılıklıdır. Her bir permutasyon için A ve B farkını ölçebiliriz. Eğer 2^k permutasyonun herbiri üretilirse, MAP'ler arası farklardan orijinal sistemler arasındaki MAP farkından büyük veya büyük eşit olanlar hesaplanabilir. Hesaplanan bu değer 2^k ya bölündüğünde yokluk hipotezi için kesin tek taraflı p-değeri veya önemlilik seviyesi bulunmuş olur. Fakat 2^k permutasyon hesaplamak oldukça zaman alıcıdır bunun yerine sınırlı sayıda rasgele permutasyon üretilerek p-değeri hesaplanmalıdır. (Smucker et al., 2007)

2.6.3.3 Anova testi (Varyans Analizi)

Varyans analiz olarak da bilinen ANOVA testi, iki ya da daha fazla örneklem ortalaması arasında farkın anlamlı olup olmadığı ile ilgili hipotezi test etmek için kullanılmaktadır. Bu test temel olarak, iki örneklem grubu için yapılan t-test'ini ikiden fazla grup için genelleştirmektedir. Testin uygulandığı modellerde bir veya daha fazla bağımsız değişken söz konusu olabilir. Eğer bir bağımsız değişken var ise bu analiz tek bir değişken ile ilgili ikiden fazla grubun ortalamalarını karşılaştırarak belirli bir anlamlılık oranında bir farkın olup olmadığını sorgular (One-way/Tek yönlü). Tek yönlü varyans analizinde örneklerin elde edildiği populasyonlar normal ya da normale yakın dağılım gösterdikleri, örneklerin bağımsız olduğu ve populasyon varyanslarının eşit olduğu varsayılmaktadır. Analizde iki bağımsız değişken var ise, iki bağımsız değişkenin bağımlı değişken üzerinde ortak etkileri belirlenir (Two-way). Ayrıca iki bağımsız değişkene ilişkin farklı grupların bağımlı değişkene göre ortalamaları karşılaştırılır ve anlamlı bir fark olup olmadığı sınanır. (Bandyopadhyay et al., 2014) .

Varyans analizinde amaç, ikiden fazla örnek için $\bar{X}_i$ 'lerin genel ortalama $\bar{X}$'dan sapmalarının kareler toplamını, bölümlere ayırmak ve analiz etmektir. Analiz sonunda, örneklerin aynı anakütleye ait birer şans örneği olup olmadıkları açığa çıkmış olur. Örneklerdeki bütün X_{ij} değerlerinin genel ortalamadan gösterdikleri sapmaların kareler toplamının gruplar arası değişkenlik ve gruplar içi değişkenliğin toplamını ifade eder;

$$GKT = GAKT + GİKT$$

Genel Kareler Toplamı (GKT): $\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2$

Gruplar arası kareler toplamı (GAKT): $n\sum_{i=1}^{k}(X_i-\bar{X})^2$

Gruplar içi kareler toplamı (GİKT): $\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{X}_\iota)^2$

GAKT, örnek ortalamalarının genel ortalamadan gösterdiği sapmalar, *GİKT* ise her bir örnekteki değerlerin kendi örnek ortalamalarından gösterdiği sapmaları ifade etmektedir.

$$S_1^2=\frac{GAKT}{k-1},\qquad S_2^2=\frac{G\dot{I}KT}{k(n-1)}$$

S_1^2 ->Gruplar arası kareler ortalaması

S_2^2 -> Gruplar içi kareler ortalaması

k-> Örnek sayısı

n-> Örnek büyüklüğü

Varyans analizinin test istatistiği olan " f değeri" :

$$f=\frac{S_1^2}{S_2^2}$$

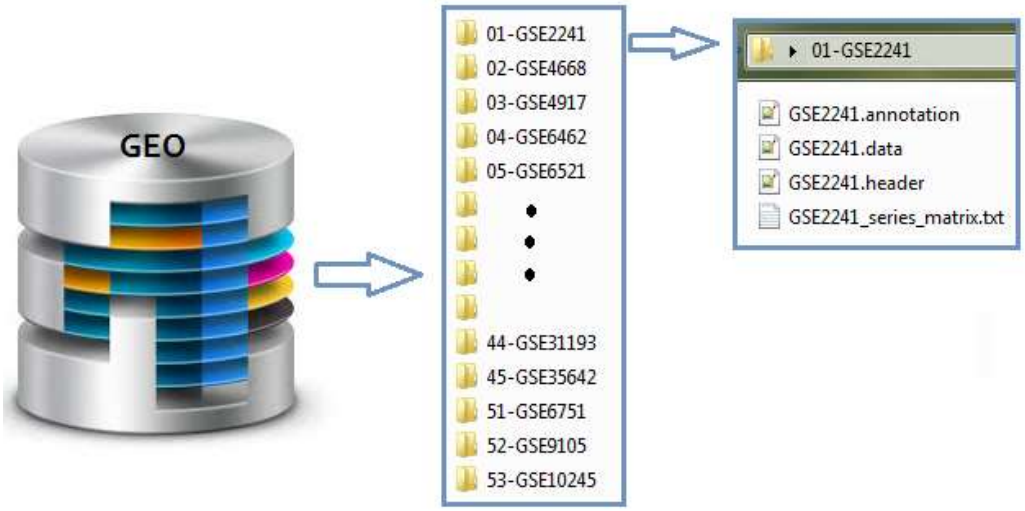
2.7 Test Verileri ve Organizasyonu

Geliştirilecek yöntemi test etmek için GEO veritabanından örnek deneyler elde edilmiştir. Toplam 51 GEO veri serisinden (GSM) 134 deney indirilmiştir. İndirilen verilerden ilk 46 tanesi meme kanseri ile ilgili yapılmış zaman serisi deneylerine aitken diğer veriler rahim kanseri, prostat kanseri, beyin kanseri, kolon kanseri, pankreas kanseri, lösemi ve diyabet gibi farklı hastalıklar ile ilgili yapılmış zaman serisi deneylerine aittir. İndirilen veri setlerindeki deneyler insan ve fareler (Mus musculus-23 adet) üzerinde yapılmıştır.

Üzerinde içerik tabanlı arama testleri yapılan veri setleri yöntemlere göre farklılık göstermektedir. İnsan ve fare genomundan deneylerle sadece FİO gen tabanlı parmak izi yöntemi test edilmiştir. İnsan ve farelerde ortak gen kümesi arama uzayını daralttığı için sadece insan genomu üzerinde iki farklı şekilde içerik tabanlı arama yöntemleri denenmiştir.

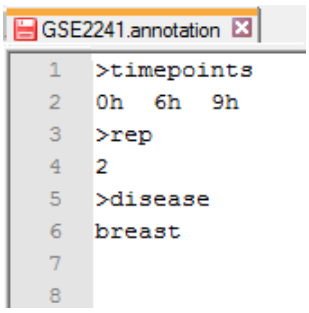
İlk olarak deneylerde ölçümü yapılan tüm genler (birleşim gen kümesi) dikkate alınarak yapılan testler 46 meme kanseri deneyi 65 diğer hastalıklar üzerinde yapılmış deneylerden oluşmuştur. İkinci olarak insan genomunda yapılan deneylerde ölçümü yapılan ortak genler (kesişim gen kümesi-Ek2) dikkate

alınarak yapılan testler 41 meme kanseri 58 diğer deneyler (Veri seti-Ek1) üzerinde yapılmıştır.



Şekil 2.11 Veri Organizasyonu

GEO'dan indirilen her bir matris serisi dosyası parçalanarak anotasyon (annotation), veri ve başlık (header) dosyaları olarak kaydedilmiştir (Şekil 2.11). Anotasyon dosyası deneyde kaç tane zaman noktası olduğunu, her bir zaman noktası için kaç tekrar olduğunu ve deneyin hangi hastalıkla ilgili olduğu bilgilerini içermektedir (Şekil 2.12).



Şekil 2.12 Veri Organizasyonu- GSE2241 için anotasyon dosyası içeriği

Header dosyası, deneyle ilgili, ne zaman yapıldığı, özeti gibi detaylı bilgileri içerir (Şekil 2.13).

GSE2241.header

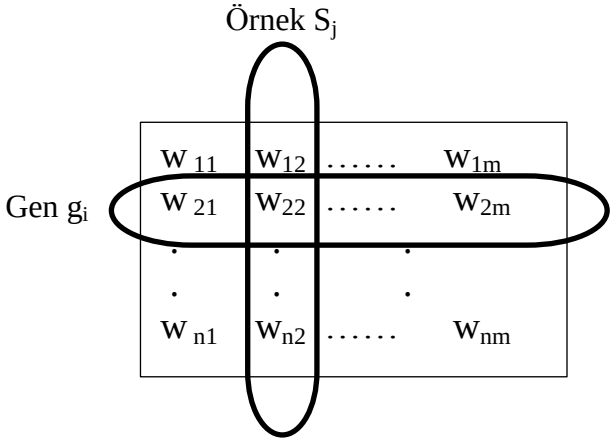
```
!Series_title   "HOXA5 transcriptional targets"
!Series_geo_accession   "GSE2241"
!Series_status  "Public on Feb 05 2005"
!Series_submission_date "Feb 03 2005"
!Series_last_update_date    "Nov 14 2012"
!Series_pubmed_id   "15757903"
!Series_summary "HOXA5 inducible cell line Hs578T was established and cultu
!Series_summary "Keywords: time-course"
!Series_type    "Expression profiling by array"
!Series_contributor "Hexin,,Chen"
!Series_contributor "Ethel,,Rubin"
!Series_contributor "Huiping,,Zhang"
!Series_contributor "Seung,,Chung"
!Series_contributor "Charles,C,Jie"
!Series_contributor "Elizabeth,,Garrett"
!Series_contributor "Shyam,,Biswal"
!Series_contributor "Saraswati,,Sukumar"
!Series_sample_id   "GSM41162 GSM41163 GSM41164 GSM41165 GSM41166 GSM41167
!Series_contact_name    "Hexin,,Chen"
!Series_contact_email   "hchen15@jhmi.edu"
!Series_contact_phone   "410-614-4075"
!Series_contact_department  " "
!Series_contact_institute   "Johns Hopkins University"
!Series_contact_address " "
!Series_contact_city    "Baltimore"
!Series_contact_state   "MD"
```

Şekil 2.13 Veri Organizasyonu- GSE2241 için basşlık dosyası içeriği

Mikrodizi deneylerinden elde edilen gen ifadesi verisi ifade-matrisleri olarak sunulur. Matrisin satırları, genlerin ifade desenlerini, sütunları örneklerin ifade profilllerini göstermektedir.

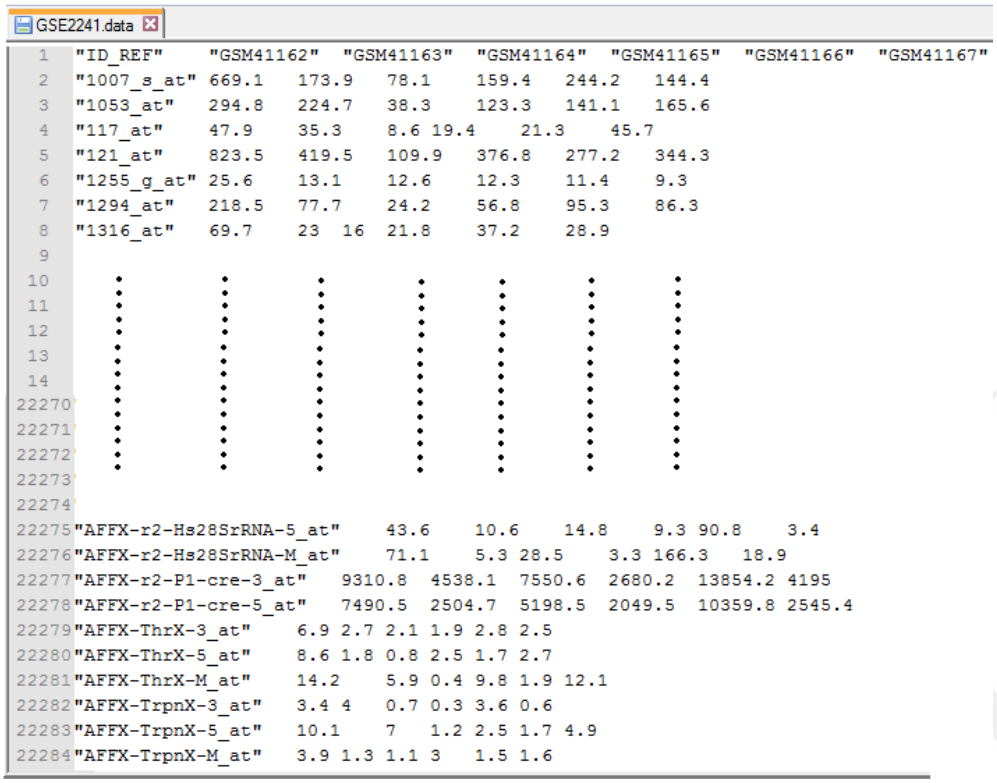
Gen ifade matrisi $M = \{ w_{ij} | 1 \leq i \leq n, 1 \leq j \leq m \}$

n-> gen sayısı, m-> örnek sayısı /zaman noktası sayısı, w_{ij} -> gen ifade matrisindeki herbir hücre başka bir ifade ile i. genin j. örneğindeki ifade ölçüm değerini gösterir.



Şekil 2.14 Gen ifade matrisi–M

İfade matrislerinin tutulduğu dosya (veri dosyası), probeID (g_i) ve GSM (GEO SaMple- S_j) numaralarına göre genlerin ifade değerlerini içermektedir (Şekil 2.15).



Şekil 2.15 Veri Organizasyonu- GSE2241 için veri dosyası içeriği

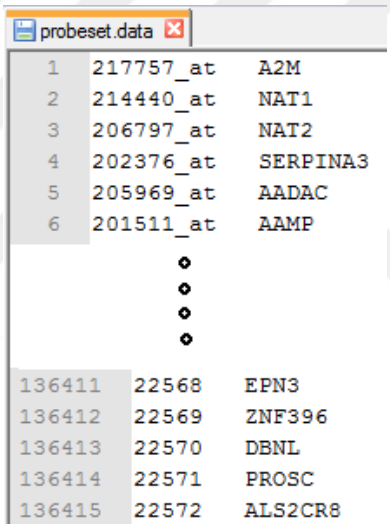
Veriler Affymetrix, Agilent Technology ve Illumina platformunun farklı çeşitlerinden alınmıştır. Bu platformların GEO ID'leri Çizelge 2.4 de verilmiştir.

Çizelge 2.4 Veri kümelerinin ait olduğu platform ve GEO ID'si

GEO ID	**PLATFORM**
GPL91	Affymetrix
GPL96	Affymetrix
GPL570	Affymetrix
GPL571	Affymetrix
GPL887	Agilent Technology
GPL6102	Illumina

GPL6883	Illumina
GPL6884	Illumina
GPL8300	Affymetrix
GPL9419	Affymetrix

Kullanılan platformlardaki ProbID'leri ve o probe'a karşılık gelen gen sembolü platformalar arası değişiklik göstermektedir. Tüm probeID'leri ve gen sembollerini içeren bir dosya oluşturulmuştur. Tüm veriler için tek bir gösterim sağlanması amacıyla herbir deneydeki probID'ler gen sembollerine dönüştürülmüştür. (Şekil 2.16).



Şekil 2.16 Kullanılan platformlardaki probeID ve karşılık gelen gen sembolleri listesi

3. BULGULAR

3.1 Meme kanseri ile ilintili sorgular

Farklı ifade olmuş gen tabanlı (Hayran vd., 2014) ve kümeleme tabanlı içerik tabanlı arama yöntemleri GEO'dan elde edilen test veri kümesi üzerinde denenmiştir. Başlangıçta deneyler meme kanseri ile ilgili ise 1 değilse 0 olarak etiketlenmiştir. Sonrasında, her bir meme kanseri deneyi test veritabanında aranarak benzerliğe göre sıralanmıştır. Bu bölümde öncelikle FİO tabanlı yöntemin sonra kümeleme tabanlı yöntemlerin içerik-tabanlı arama sonuçları verilecektir.

FİO tabanlı yöntemde test verilerindeki herbir deney için Nudge yöntemi ile deneydeki genlerin FİO olma olasılıkları hesaplanmış ve sonuçlar parmak izi dosyalarına kaydedilmiştir. Deneyler zaman serisi deneylerine ait olduğu için FİO tabanlı yöntemde iki farklı yol denenmiştir. İlk olarak, deneyde bulunan zaman noktalarından ilk ve son nokta arasındaki fark esas alınarak Nudge yöntemi ile FİO olma olasılıkları hesaplanmıştır (Son_FİO). İkinci yol ise ilk zaman noktası ve diğer tüm zaman noktaları arasındaki farklardan maksimum olan FİO gen olma olasılığı hesaplanmıştır (Max_FİO). Her iki yol da Nudge tabanlıdır.

FİO tabanlı yöntemlerde deneylerin benzerliği hesaplanırken FİO genler iki farklı şekilde belirlenmiştir. Birincisi eşik değer atama yöntemi; hesaplanan FİO gen olma olasılığı belirlenen bir eşiğin üzerinde ise o gen FİO olarak tanımlanmıştır. İkincisi; deneydeki tüm genler için hesaplanan FİO gen olma olasılıkları sıralanır, sıralanan genlerin yüzde olarak belirli bir kısmı FİO gen olarak kabul edilir. Yöntemlerde deneylerin benzerliği ölçülürken; Öklit uzaklığı, Pearson korelasyon katsayısı, Spearman sıralama katsayısı ve Tanimoto katsayısı ölçütleri kullanılmıştır. Yöntemlerin arama performansını ölçmek ve karşılaştırmak için ROC skorları ve eğrinin altında kalan alanlar hesaplanmıştır (Şekil 3.1- 3.2).

Bir deneydeki ifade ölçümü yapılan tüm genler hesaba katıldığında (Birleşim) en iyi ROC skoru Son_FİO yönteminde, FİO gen olma olasılığı sıralandığında listenin %0.9'unun FİO gen olarak hesaplandığı duruma aittir. Bu durumda kullanılan benzerlik ölçütü Tanimoto katsayısıdır (Çizelge 3.1). İçerik-tabanlı arama uygulamaları farklı eşik ve yüzde değerleri ile yapılmıştır. Çizelge 3.1 de en iyi ROC skorları alınan eşik ve yüzde değerleri verilmiştir. Çizelge 3.2 de birleşim verisi üzerinde farklı benzerlik ölçütleri ile hesaplanan ortalama ROC değerleri verilmiştir.

Çizelge 3.1 Birleşim gen verisinde Tanimoto katsayısı benzerliği uygulandığında hesaplanan ROC skorları

	Yüzde			**Eşik**	
	Son_FİO(%0.9)	**Max_FİO(%0.9)**		**Son_FİO(0.2)**	**Max_FİO(0.2)**
1	0,8191	0,7706		0,7659	0,7331
2	0,7629	0,7565		0,7298	0,7331
3	0,7599	0,7388		0,7204	0,7067
4	0,7525	0,7381		0,7167	0,7033
5	0,7502	0,7144		0,7134	0,7010
6	0,7391	0,7097		0,7060	0,6893
7	0,7318	0,7010		0,7060	0,6866
8	0,7284	0,6983		0,7027	0,6813
9	0,7234	0,6936		0,7020	0,6809
10	0,7211	0,6890		0,7003	0,6799
11	0,7171	0,6793		0,6967	0,6779
12	0,7171	0,6729		0,6863	0,6769
13	0,7164	0,6689		0,6839	0,6726
14	0,7107	0,6682		0,6839	0,6605
15	0,7087	0,6666		0,6833	0,6505
16	0,6997	0,6632		0,6829	0,6492
17	0,6946	0,6612		0,6709	0,6482
18	0,6940	0,6579		0,6709	0,6448
19	0,6866	0,6505		0,6696	0,6428
20	0,6846	0,6492		0,6672	0,6415
21	0,6786	0,6455		0,6652	0,6378
22	0,6766	0,6431		0,6642	0,6348
23	0,6749	0,6405		0,6642	0,6314
24	0,6652	0,6361		0,6599	0,6288

25	0,6498	0,6338		0,6545	0,6151
26	0,6488	0,6334		0,6538	0,6114
27	0,6418	0,6278		0,6448	0,6114
28	0,6351	0,6264		0,6448	0,6107
29	0,6271	0,6191		0,6365	0,6067
30	0,6221	0,6177		0,6355	0,5997
31	0,6204	0,6100		0,6338	0,5943
32	0,6117	0,5926		0,6144	0,5863
33	0,6067	0,5833		0,6060	0,5833
34	0,5993	0,5605		0,5960	0,5789
35	0,5980	0,5525		0,5739	0,5632
36	0,5819	0,5385		0,5682	0,5589
37	0,5796	0,5294		0,5609	0,5365
38	0,5609	0,5281		0,5605	0,5258
39	0,5495	0,5221		0,5532	0,5231
40	0,5298	0,5033		0,5391	0,5224
41	0,5274	0,5027		0,5351	0,4813
42	0,5251	0,4686		0,4846	0,4753
43	0,5010	0,4043		0,4739	0,4391
44	0,4318	0,4037		0,4472	0,4284
45	0,4043	0,4037		0,3776	0,4023
46	0,3843	0,3809		0,3579	0,3696

Çizelge 3.2 Birleşim verileri üzerinde farklı benzerlik ölçütleri için hesaplanan ortalama ROC skorları

Benzerlik Ölçütü	**Ortalama ROC değerleri**
Öklid Uzaklığı	0.501
Spearman Sıralama Korelasyon Katsayısı	0.604
Tanimoto Katsayısı	0.644
Pearson Korelasyon Katsayısı	0.634

Her bir deney için ifade değeri ölçülen genler farklılık göstermektedir. Bu sebeple test verileri, her deneyde ortak bulunan gen kümesine göre (7078 adet) tekrar düzenlemiştir. Kesişim verilerinin bulunduğu test verileri üzerinde de Nudge yöntemi (Son_FİO) ve Maksimum Nudge yöntemi (Max_FİO) farklı eşik ve yüzde değerleri için test edilmiştir. (Çizelge 3.3). En yüksek ROC skoru Son_FİO yönteminde sıralanmış FİO gen olasılıklarının %5'i FİO gen olarak kabul edildiğinde hesaplanmıştır (Şekil 3.1).

Çizelge 3.3 Kesişim gen verisinde Tanimoto katsayısı benzerliği uygulandığında hesaplanan ROC skorları

	Yüzde			**Eşik**	
	Son_FİO (%5)	**Max_FİO(%3)**		**Son_FİO (0.2)**	**Max_FİO (0.2)**
1	0,7819	0,8013		0,7599	0,7258
2	0,7819	0,7552		0,7144	0,7023
3	0,7749	0,7134		0,7137	0,6980
4	0,7441	0,6993		0,7114	0,6906
5	0,7274	0,6963		0,7114	0,6880
6	0,7241	0,6903		0,7080	0,6880
7	0,7207	0,6896		0,6973	0,6799
8	0,7027	0,6809		0,6910	0,6756
9	0,6960	0,6796		0,6906	0,6676
10	0,6957	0,6773		0,6870	0,6609

11	0,6886	0,6706		0,6860	0,6595
12	0,6873	0,6635		0,6853	0,6482
13	0,6870	0,6552		0,6853	0,6468
14	0,6783	0,6518		0,6836	0,6468
15	0,6726	0,6512		0,6819	0,6462
16	0,6709	0,6492		0,6816	0,6441
17	0,6639	0,6448		0,6766	0,6405
18	0,6632	0,6375		0,6732	0,6388
19	0,6625	0,6341		0,6722	0,6361
20	0,6609	0,6341		0,6689	0,6331
21	0,6592	0,6321		0,6639	0,6284
22	0,6562	0,6308		0,6612	0,6274
23	0,6488	0,6298		0,6612	0,6271
24	0,6482	0,6298		0,6582	0,6268
25	0,6441	0,6278		0,6528	0,6237
26	0,6405	0,6241		0,6492	0,6197
27	0,6398	0,6197		0,6472	0,6194
28	0,6358	0,6187		0,6361	0,6120
29	0,6261	0,6110		0,6288	0,6084
30	0,6224	0,6057		0,6147	0,6070
31	0,6191	0,6047		0,6124	0,5916
32	0,6144	0,5997		0,6117	0,5890
33	0,5983	0,5940		0,5883	0,5689
34	0,5863	0,5913		0,5786	0,5672
35	0,5659	0,5873		0,5773	0,5629
36	0,5455	0,5595		0,5548	0,5391
37	0,5452	0,5508		0,5522	0,5391
38	0,5338	0,5431		0,5468	0,5358

39	0,5274	0,5040		0,5395	0,5301
40	0,5194	0,4595		0,5154	0,5191
41	0,5047	0,4274		0,5100	0,4953
42	0,4378	0,4114		0,5054	0,4602
43	0,4368	0,3946		0,4140	0,4385
44	0,3876	0,3860		0,3876	0,4234
45	0,3475	0,3813		0,3622	0,3920
46	0,2926	0,3057		0,1094	0,2816

Ayrıca yöntemler, Pearson korelasyon katsayısı, Spearman sıralama korelasyon katsayısı, Tanimoto katsayısı ve Öklit uzaklığı gibi farklı benzerlik ölçütleri ile de test edilmiştir. Kesişim test verileri ile en iyi sonuç veren benzerlik ölçütü Nudge yöntemi (Son_FİO) için Pearson korelasyon katsayısı, Maksimum Nudge yöntemi (Max_FİO) için Tanimoto katsayısı olmuştur (Çizelge 3.4-3.5).

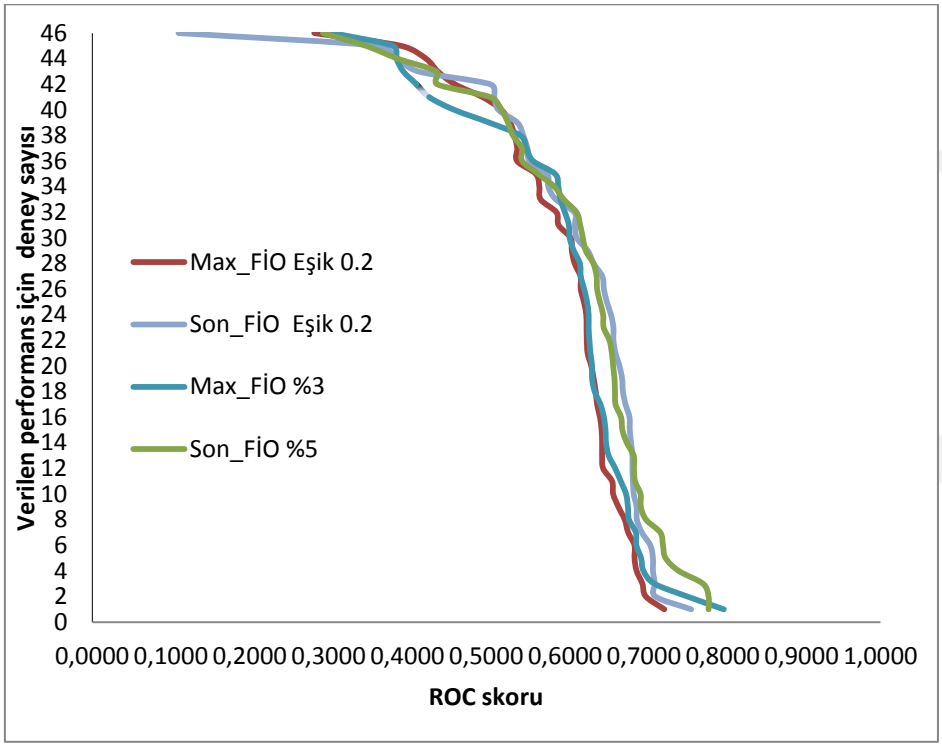
Çizelge 3.4 Kesişim verileri üzerinde farklı benzerlik ölçütleri için hesaplanan ortalama ROC skorları (Son_FİO)

Benzerlik Ölçütü	**Ortalama ROC değerleri**
Öklid Uzaklığı	0.504
Spearman Sıralama Korelasyon Katsayısı	0.592
Tanimoto Katsayısı	0.621
Pearson Korelasyon Katsayısı	0.626

Çizelge 3.5 Kesişim verileri üzerinde farklı benzerlik ölçütleri için hesaplanan ortalama ROC skorları (Max_FİO)

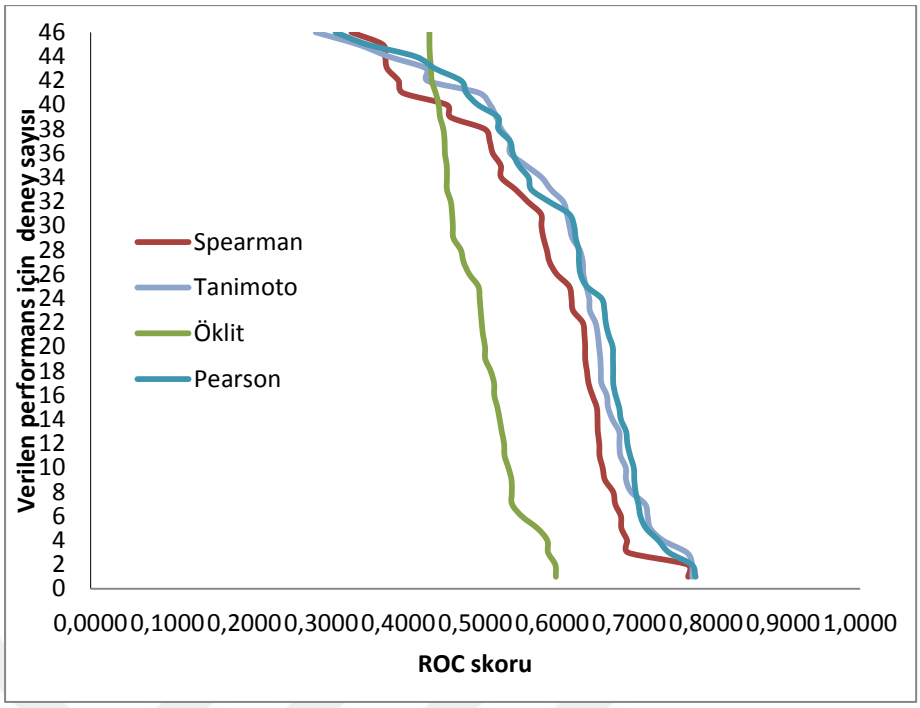
Benzerlik Ölçütü	**Ortalama ROC değerleri**
Öklid Uzaklığı	0.504
Spearman Sıralama Korelasyon Katsayısı	0.574
Tanimoto Katsayısı	0.602
Pearson Korelasyon Katsayısı	0.598

Sistemin genel arama performansının sunulması için ortalama ROC skorları yerine sistemin yüksek ROC skoruna sahip olduğu deney sayısının verilmesi daha uygun olacaktır. Bu durumda verilen eğrinin yüksekliği arama performansının etkililiği ile doğru orantılıdır. Başka bir deyişle eğrinin altında kalan alan (AUC) ne kadar büyük ise sistemin geri getirim performansı o kadar iyidir denilebilir. Şekil 3.1 de kesişim verisi kullanıldığında Tanimoto katsayısı ile hesaplanan ROC skorları görülmektedir. Eğrinin altında kalan alan bakımından Son_FİO yöntemi Max_FİO yöntemine göre daha iyi arama performansına sahiptir.



Şekil 3.1Tanimoto katsayısı benzerlik ölçütü olarak kullanıldığında geri getirim performansı

Şekil 3.2 de ise kesişim verisi kullanıldığı durumdaki sistemin arama performansı görülmektedir. Eğrinin altında kalan alana göre Pearson katsayısının diğer benzerlik ölçütlerine göre geri getirim performansı daha iyidir.



Şekil 3.2 Farklı benzerlik ölçütleri ile geri getirim performansı

Tez çalışmasında önerilen kümeleme-tabanlı içerik-tabanlı arama kesişim verileri üzerinde test edilmiştir. Kümeleme-tabanlı yöntemde FİO gen tabanlı yöntemden farklı olarak genlerin iki zaman noktasındaki ifade ölçüm değerlerinin farkları (Nudge yöntemi) yerine tüm zaman noktalarındaki ifade ölçüm değerleri hesaba katılmıştır. Tüm deneylerde ortak bulunan genler (Kesişim verileri) belirli model seçilerek (verinin karakteristiğine uygun model kümeleme algoritması tarafından belirlenmiştir) kümelenmiştir. Kümeleme tabanlı yöntemde içerik-tabanlı arama üç farklı kümeleme algoritması denenerek yapılmıştır; Mclust, Funfem ve Kmeans.

Çizelge 3.6 da 3 farklı kümeleme yöntemi ile oluşturulan parmak izleri ile yapılan içerik tabanlı arama sonucunda hesaplanan ROC ve AP değerleri görülmektedir. Çizelge 3.6 da verilen FİO tabanlı yöntem en yüksek ROC sonuçlarının elde edildiği, benzerlik ölçütü olarak Pearson katsayısının kullanıldığı Son_FİO yöntemine aittir. Son_FİO yönteminde, Nudge yöntemi ile genlerin FİO gen olma olasılıkları -ilk ve son zaman noktası arasındaki ifade değişim oranı- hesaplanmaktadır. Çizelgede verilen tüm hesaplamalar kesişim verisi üzerinde yapılmıştır.

Çizelge 3.6 Kümeleme-tabanlı içerik tabanlı arama sonuçları; ROC ve AP değerleri

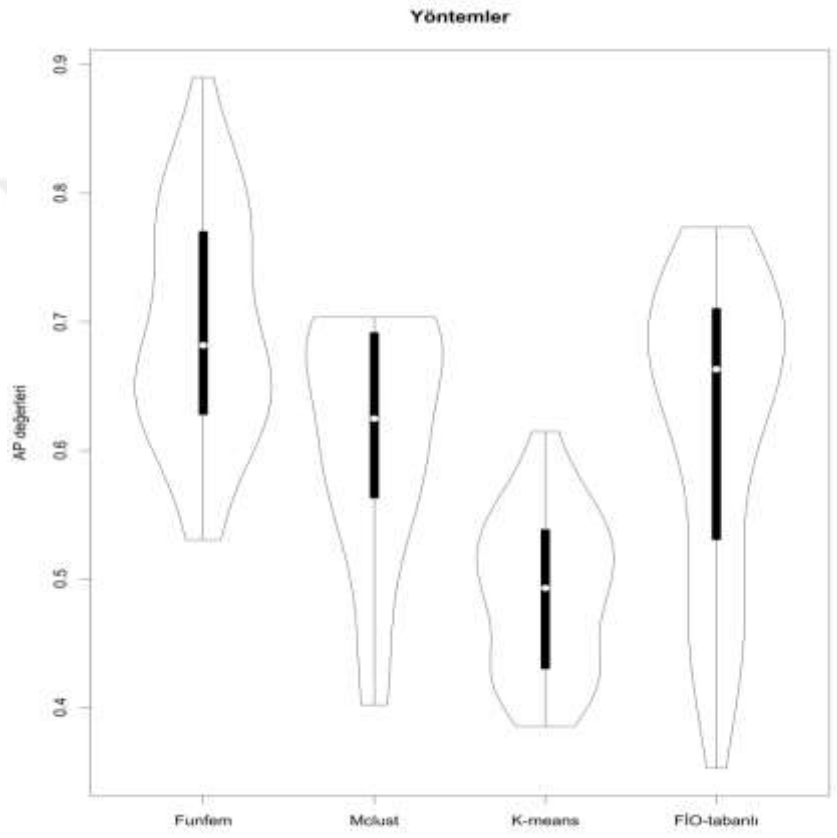
	ROC				AP			
	Funfem	Mclust	Kmeans	FİO-tabanlı	Funfem	Mclust	Kmeans	FİO-tabanlı
1	0,89908	0,62972	0,62237	0,79773	0,89008	0,70330	0,61430	0,77340
2	0,85366	0,61987	0,60345	0,79142	0,82975	0,70061	0,59675	0,74941
3	0,85114	0,61905	0,60261	0,75778	0,82877	0,69721	0,58411	0,73481
4	0,84735	0,61576	0,57948	0,75189	0,82323	0,69562	0,56636	0,73131
5	0,84609	0,61535	0,55299	0,74979	0,82252	0,69460	0,55231	0,73013
6	0,83894	0,61207	0,54836	0,74895	0,79802	0,69419	0,55199	0,72522
7	0,83221	0,61043	0,54584	0,74811	0,79501	0,69412	0,54996	0,72468
8	0,83011	0,61043	0,53070	0,74054	0,78018	0,69343	0,54944	0,72415
9	0,83011	0,60837	0,52565	0,73802	0,78018	0,69286	0,54530	0,71752
10	0,82843	0,60509	0,51934	0,73507	0,77936	0,69211	0,54297	0,71160
11	0,81960	0,60304	0,51892	0,72876	0,77034	0,69151	0,53822	0,71029
12	0,81161	0,60181	0,51808	0,72288	0,76128	0,69108	0,53130	0,70803
13	0,80908	0,60181	0,51808	0,71741	0,75579	0,68957	0,52692	0,70230
14	0,80782	0,60099	0,51556	0,71573	0,75446	0,68951	0,52032	0,69519
15	0,80151	0,60016	0,51430	0,71362	0,74560	0,68769	0,51953	0,67544
16	0,79857	0,59934	0,51388	0,71320	0,72305	0,67714	0,51348	0,67350
17	0,79605	0,59852	0,51388	0,71236	0,69752	0,67152	0,51131	0,66947
18	0,79100	0,59688	0,51220	0,70563	0,69559	0,65350	0,51050	0,66802
19	0,78385	0,59688	0,49916	0,70479	0,69227	0,63582	0,50030	0,66644
20	0,78301	0,59483	0,49664	0,69344	0,68326	0,62491	0,49774	0,66481
21	0,74811	0,59113	0,49579	0,69302	0,68155	0,62454	0,49267	0,66297
22	0,74012	0,59072	0,49117	0,67956	0,67385	0,60929	0,49198	0,65035
23	0,73886	0,58908	0,48234	0,67157	0,65786	0,60622	0,48720	0,64956
24	0,73886	0,57800	0,48108	0,66064	0,65776	0,60490	0,48459	0,64940

25	0,73129	0,57594	0,48024	0,64929		0,65192	0,60481	0,47627	0,63863
26	0,72918	0,57389	0,46384	0,64508		0,64960	0,59898	0,47249	0,63650
27	0,72792	0,57348	0,45038	0,62700		0,64892	0,59248	0,47078	0,60102
28	0,72456	0,57020	0,45038	0,61480		0,64746	0,57751	0,46032	0,57443
29	0,71489	0,56609	0,44996	0,60681		0,63299	0,57643	0,43178	0,57136
30	0,70648	0,56281	0,44912	0,58242		0,63106	0,57353	0,43176	0,56975
31	0,70059	0,54187	0,44323	0,56013		0,62758	0,56270	0,42975	0,53033
32	0,68251	0,53202	0,43902	0,53995		0,62375	0,55441	0,42253	0,50716
33	0,67998	0,51437	0,43566	0,53659		0,62221	0,54207	0,42178	0,48377
34	0,67914	0,51149	0,41758	0,53490		0,62057	0,53784	0,41967	0,47770
35	0,67746	0,50041	0,41295	0,53490		0,61846	0,51822	0,41950	0,47629
36	0,67536	0,50041	0,40580	0,52607		0,61742	0,50743	0,41828	0,45197
37	0,66905	0,49425	0,40412	0,51220		0,61405	0,47529	0,41736	0,44185
38	0,66190	0,49138	0,40328	0,48276		0,56545	0,45325	0,41625	0,42730
39	0,65854	0,48440	0,39907	0,44239		0,55812	0,41909	0,41487	0,42217
40	0,65181	0,42693	0,36712	0,43692		0,53048	0,40532	0,38531	0,41437
41	0,61733	0,39286	0,31034	0,34020		0,53005	0,40158	0,38471	0,35244

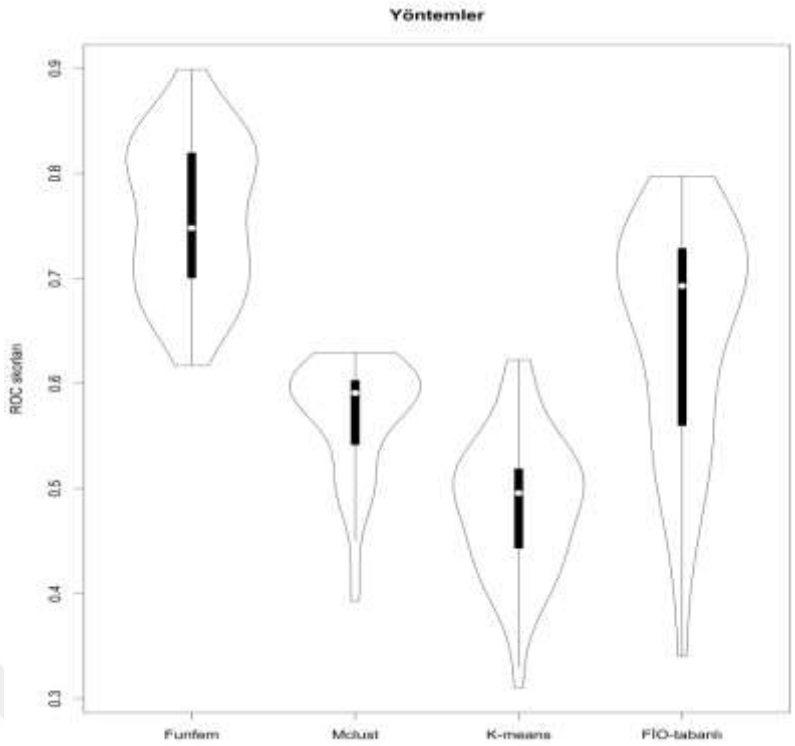
Şekil 3.3 ve Şekil 3.4 de ROC skorları ve AP değerlerinin olasılık yoğunlukları görülmektedir. Violin grafiğinde beyaz nokta ortalama değeri ve şişkin bölgeler tekrar eden değerleri göstermektedir. Bir yöntem için viyolin grafiğinde verilen şekil (Violin şekli) 1 noktasında (AP değeri /ROC skoru) yığılma yaptığı durum, yöntemin arama performansının mükemmel olduğunu işaret etmektedir. Tekrar eden değerlerin biriktiği viyolinin şişkin bölgesi 1 değerinden ne kadar aşağıda ise yöntemin performansının o kadar kötü olduğundan bahsedilebilir.

Şekil 3.3 de hesaplanan AP değerlerine göre en yüksek değerler (0.5-0.9) kümeleme-tabanlı yöntemde Funfem algoritması kullanıldığı duruma aittir. Değerlerin sık tekrar ettiği aralık 0.6-0.7 dir. En düşük değerler ise FİO-tabanlı yöntem ve Kmeans kümeleme algoritmasının kullanıldığı yönteme aittir. FİO-tabanlı yöntem her nekadar en düşük değerlere sahip olsada viloinin şekli itibari ile 0.6-0.8 aralığında değerler çok tekrar etmiş ve ortalama değerler alınan en

yüksek ortalama değere (Funfem) yakın çıkmıştır. Fakat Kmeans kümeleme algoritmasının kullanıldığı durumda değerler 0.4-0.6 arasında değişmiş ortalama değer olarakta FİO-tabanlı yönteme göre oldukça geride kalmıştır. Şekil 3.4 de ise hesaplanan ROC skorları verilmiştir. Ortalama ROC skoru ve MAP değeri birbirleriyle uyumlu olarak aynı sonucu vermiştir. Başka bir deyişle en iyi geri getirim performansı Funfem algoritması kullanıldığında en düşük performans Kmeans kümeleme algoritması uygulandığında hesaplanmıştır.



Şekil 3.3 Yöntemlerin AP değerlerinin violin grafiği

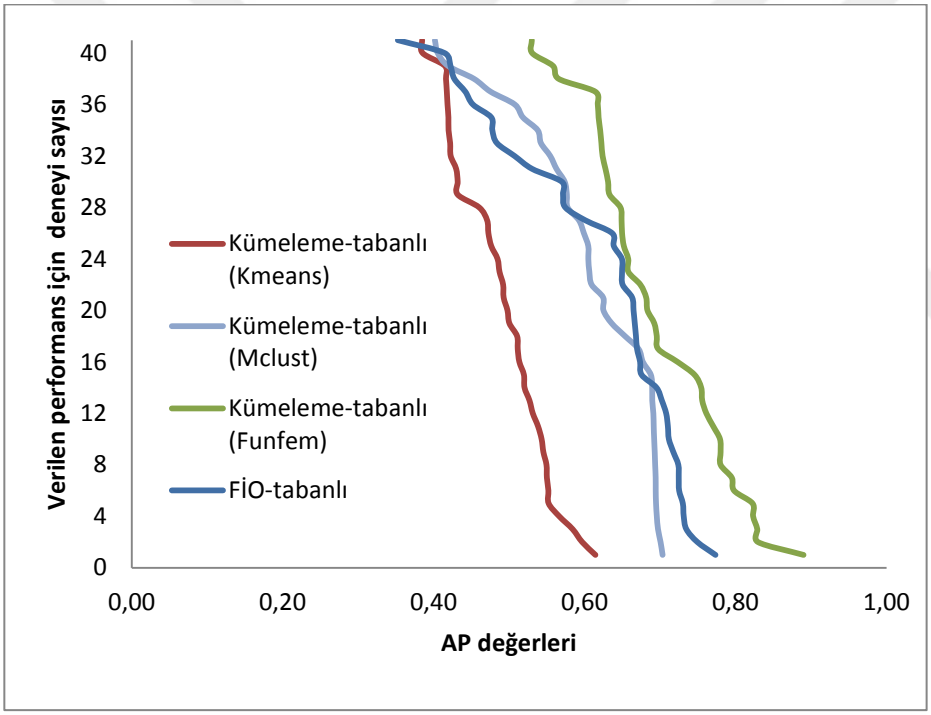


Şekil 3.4 Kümeleme-tabanlı yöntemlerin ROC değerlerinin violin grafiği

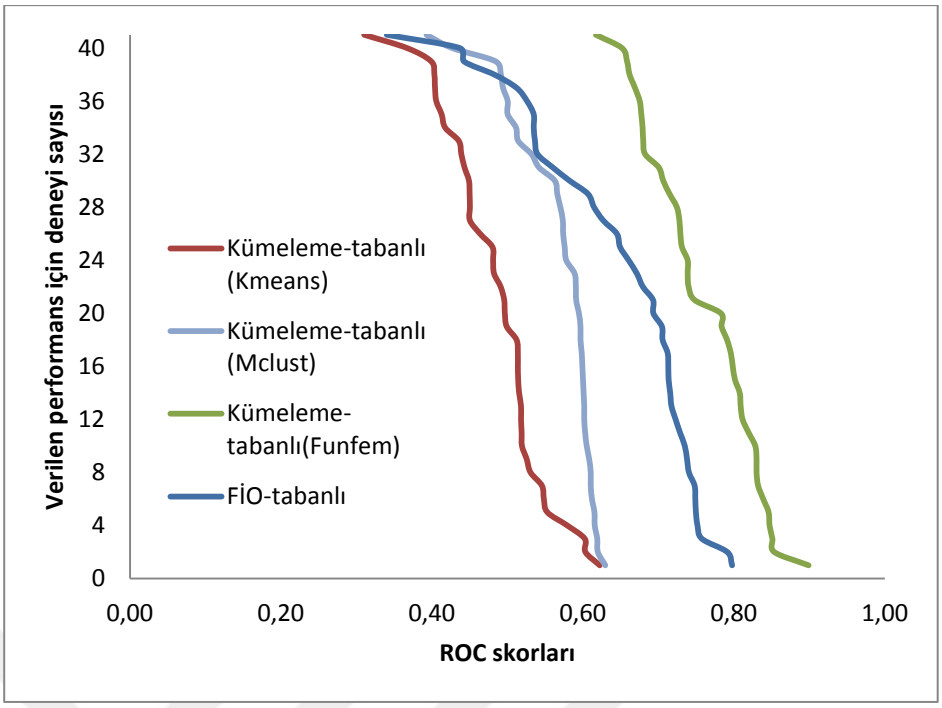
Yüksek ROC skoru ve MAP değerleri yöntemin sorguya benzer sonuçları öncelikli olarak getirdiğini gösterir. Çizelge 3.7 de verilen MAP ve ROC değerleri birbirleri ile tutarlıdır ve en yüksek skoru Funfem kümeleme yöntemi vermektedir. Şekil 3.5 ve 3.6 da test verilerine uygulanan içerik tabanlı arama yöntemlerinin arama performansları, ROC skorları ve AP değerleri üzerinden verilmiştir. Şekil 3.5 de AUC' a göre en iyi performans kümeleme algoritması olarak Funfem kullanıldığı yönteme aitken en kötü performans kümeleme algoritması olarak Kmeans kullanıldığı duruma aittir. FİO-tabanlı yöntem ile Mclust tabanlı kümeleme yöntemi arasında grafikte AP değerleri ile elde edilen eğrinin altında kalan alan açısından çok fark yoktur. Şekil 3.6 da verilen ROC skorları ise AP değerleri (Şekil 3.5) ile elde edilen performans karşılaştırması ile birebir örtüşmektedir.

Çizelge 3.7 Tüm yöntemler için hesaplanan Ortalama ROC skorları ve MAP değerleri

Yöntem	Ortalama ROC skor	MAP
FİO-tabanlı	0,64791	0,61817
Mclust (Kümeleme-tabanlı)	0,56834	0,61015
Funfem (Kümeleme-tabanlı)	0,75886	0,69433
Kmeans (Kümeleme-tabanlı)	0,48497	0,48958



Şekil 3.5 Yöntemlerin AP değerleri grafiği



Şekil 3.6 Yöntemlerde hesaplanan ROC skorları grafiği

Kümeleme algoritmalarının kullanıldığı içerik-tabanlı arama ile alınan sonuçlar öznitelik seçimi algoritmaları kullanılarak iyileştirilmeye çalışılmıştır. Funfem ve Mclust kümeleme yöntemleri ile oluşturulan parmak izlerine öznitelik seçme algoritmaları uygulanmış ve her bir gen için ait oldukları deneyi tanımlama ağırlıkları hesaplanmıştır. Gen ağırlıkları, deneylerin benzerliğini karşılaştırırken kullanılmıştır. Eğer iki deneyde aynı gen aynı kümeye atanmış ise ilgili genin ağırlığı, (iki deneydeki aynı kümeye atanan aynı genlerin toplam ağırlıkları) deneylerin benzerliği belirlenirken kullanılmıştır. Öznitelik seçimi algoritması kullanılmadığı durumda, deneylerin benzerliği hesaplanırken her bir genin ağırlığı sabit (1) alınmıştır. Çizelge 3.8 de farklı öznitelik seçimi algoritmaları kullanılarak hesaplanan gen ağırlıkları ile yapılan arama sonuçlarındaki duyarlılık değerleri görülmektedir. Çizelgede Funfem ve Mclust olarak verilen değerler ağırlıklandırılmamış yöntemlere aittir. Diğer kolonlar kullanılan öznitelik algoritmasının isimleriyle adlandırılmıştır. Çizelge 3.9 da ağırlıklandırılmış gen hesabı ile elde edilen ortalama duyarlılık değerleri verilmiştir.

Çizelge 3.8.Öznitelik seçimi algoritmaları sonucu yapılan aramaların AP değerleri

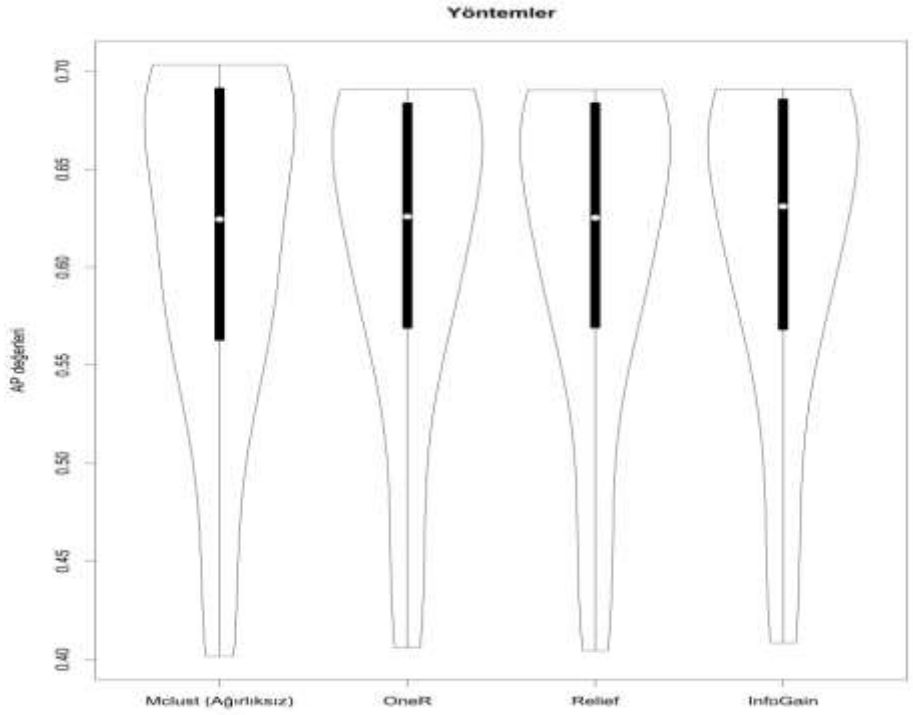
AP								
	Funfem					**Mclust**		
	Funfem	**Relief**	**InfoGain**	**OneR**	**Mclust**	**Relief**	**InfoGain**	**OneR**
1	0,8901	0,7806	0,7602	0,8702	0,7033	0,6905	0,6910	0,6907
2	0,8297	0,7667	0,7588	0,8149	0,7006	0,6897	0,6896	0,6899
3	0,8288	0,7638	0,7585	0,8137	0,6972	0,6895	0,6890	0,6897
4	0,8232	0,7549	0,7578	0,8116	0,6956	0,6874	0,6881	0,6872
5	0,8225	0,7278	0,7577	0,8033	0,6946	0,6863	0,6877	0,6856
6	0,7980	0,7113	0,7547	0,7956	0,6942	0,6859	0,6876	0,6855
7	0,7950	0,7095	0,7547	0,7769	0,6941	0,6857	0,6874	0,6852
8	0,7802	0,7060	0,7538	0,7769	0,6934	0,6856	0,6866	0,6850
9	0,7802	0,7060	0,7513	0,7710	0,6929	0,6851	0,6866	0,6844
10	0,7794	0,7033	0,7495	0,7675	0,6921	0,6851	0,6862	0,6843
11	0,7703	0,6979	0,7439	0,7666	0,6915	0,6840	0,6859	0,6839
12	0,7613	0,6976	0,7299	0,7611	0,6911	0,6832	0,6846	0,6836
13	0,7558	0,6850	0,7299	0,7594	0,6896	0,6825	0,6829	0,6836
14	0,7545	0,6830	0,7114	0,7549	0,6895	0,6822	0,6827	0,6825
15	0,7456	0,6339	0,6577	0,7530	0,6877	0,6809	0,6794	0,6806
16	0,7230	0,6298	0,6464	0,7104	0,6771	0,6704	0,6650	0,6662
17	0,6975	0,6131	0,6379	0,6892	0,6715	0,6621	0,6616	0,6619
18	0,6956	0,6090	0,6340	0,6892	0,6535	0,6466	0,6494	0,6461
19	0,6923	0,6052	0,6331	0,6822	0,6358	0,6371	0,6416	0,6372
20	0,6833	0,6048	0,6312	0,6766	0,6249	0,6368	0,6344	0,6367
21	0,6816	0,6040	0,6303	0,6639	0,6245	0,6253	0,6311	0,6259
22	0,6739	0,6037	0,6293	0,6522	0,6093	0,6244	0,6304	0,6234
23	0,6579	0,6027	0,6291	0,6521	0,6062	0,6184	0,6190	0,6176
24	0,6578	0,6024	0,6288	0,6507	0,6049	0,6169	0,6141	0,6159

25	0,6519	0,6021	0,6269	0,6450	0,6048	0,6096	0,6116	0,6104
26	0,6496	0,6009	0,6235	0,6441	0,5990	0,6082	0,6084	0,6066
27	0,6489	0,5976	0,6060	0,6438	0,5925	0,5940	0,6040	0,5956
28	0,6475	0,5953	0,6050	0,6402	0,5775	0,5920	0,5966	0,5953
29	0,6330	0,5925	0,6035	0,6189	0,5764	0,5915	0,5951	0,5916
30	0,6311	0,5776	0,6032	0,6151	0,5735	0,5804	0,5813	0,5781
31	0,6276	0,5634	0,5875	0,6133	0,5627	0,5691	0,5681	0,5689
32	0,6238	0,5324	0,5529	0,6046	0,5544	0,5638	0,5658	0,5643
33	0,6222	0,5305	0,5473	0,5980	0,5421	0,5591	0,5655	0,5586
34	0,6206	0,5085	0,5436	0,5952	0,5378	0,5411	0,5377	0,5416
35	0,6185	0,5053	0,5428	0,5936	0,5182	0,5219	0,5222	0,5257
36	0,6174	0,4756	0,5400	0,5922	0,5074	0,4849	0,5010	0,4925
37	0,6141	0,4742	0,5397	0,5681	0,4753	0,4795	0,4791	0,4799
38	0,5654	0,4733	0,5360	0,5539	0,4532	0,4522	0,4568	0,4525
39	0,5581	0,4732	0,5308	0,5503	0,4191	0,4310	0,4308	0,4316
40	0,5305	0,4727	0,5112	0,5209	0,4053	0,4120	0,4134	0,4125
41	0,5301	0,4311	0,4894	0,5124	0,4016	0,4045	0,4085	0,4059

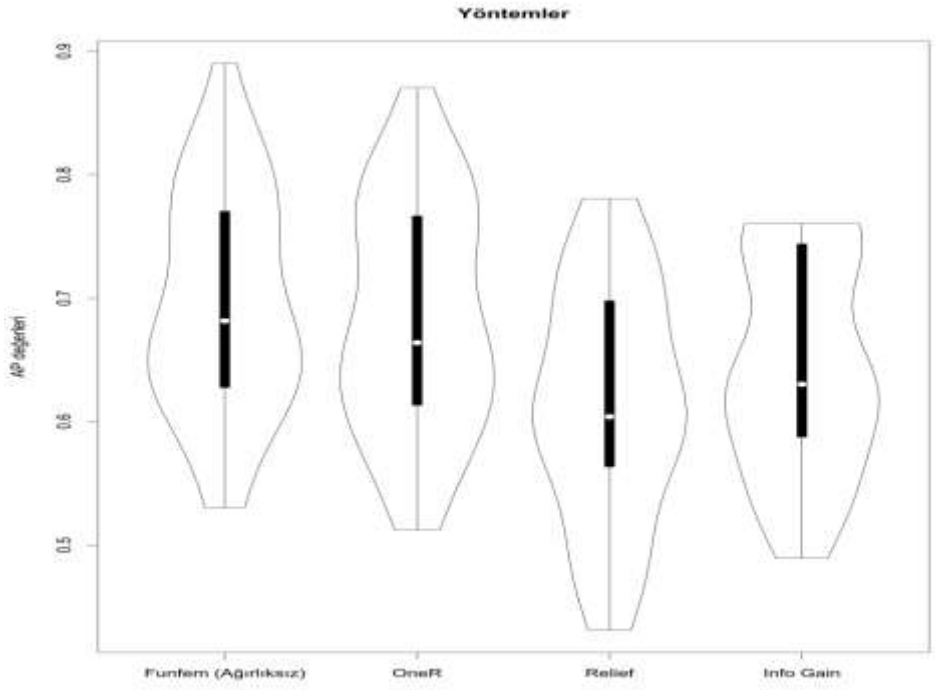
Çizelge 3.9 Öznitelik seçimi algoritmaları sonucu yapılan aramaların MAP değerleri

Kümeleme Yöntemi	**MAP (Ağırlıksız)**	**MAP (OneR)**	**MAP (Relief)**	**MAP (InfoGain)**
Mclust	0,6102	0,6104	0,6102	0,6119
Funfem	0,6943	0,6823	0,6148	0,6444

Şekil 3.7 ve Şekil 3.8 de ağırlıklı ve ağırlıksız yöntemlerle yapılan içerik tabanlı arama sonucunda elde edilen AP değerlerinin violin çizgesi verilmiştir. Ortaya çıkan viyolinlerden de anlaşıldığı gibi öznitelik algoritmaları ağırlıksız yöntemlerin aldığı sonuçları iyileştirmemiş yaklaşık aynı sonuçları vermiştir.



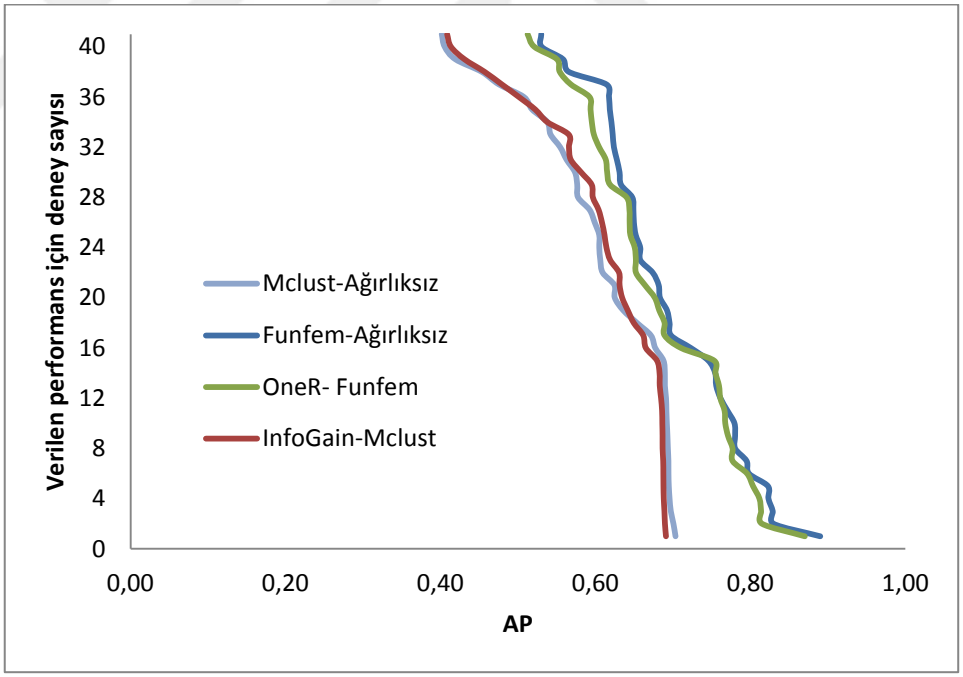
Şekil 3.7 Mclust kümeleme algoritması ve farklı öznitelik seçimi ile yapılan arama için hesaplanan AP değerlerinin Violin çizgesi.



Şekil 3.8 Funfem kümeleme algoritması ve farklı öznitelik seçimi ile yapılan arama için hesaplanan AP değerlerinin Violin çizgesi.

Yüksek MAP değeri, arama sonucunda ortaya çıkan listenin üst sıralarında sorgu deneye daha çok benzer deneylerin bulunduğunu göstermektedir. Çizelge 3.8 de verilen AP değerlerine göre Funfem kümeleme yöntemi Mclust'a göre daha iyi getirme performansı göstermiştir. Ayrıca sonuçlar gösteriyor ki özellik seçimi algoritmaları ile hesaplanan ağırlıkların arama sonuçlarının iyileştirilmesinde bir etkisi yoktur.(Şekil 3.7, Şekil 3.8): kullanılan öznitelik seçimi algoritmaları, ağırlıksız yönteme göre ya aynı ya da çok yakın MAP değeri vermiştir. Violin çizgeleri MAP değerlerini beyaz nokta şeklinde göstermektedir. Violindeki şişkin bölgeler AP değerlerinin diğerlerine göre daha çok tekrar ettiği aralıkları verir.

Şekil 3.9 de ise, en iyi sonuç vermiş ağırlıklı ve ağırlıksız yöntemlerin arama performansı bir arada gösterilmiştir. Benzerlik ölçütü olarak ağırlıksız çakışma yönteminin ve parmak izi için Funfem kümeleme yönteminin kullanıldığı model en iyi sonuçları vermiştir. Parmak izi için Mclust kümeleme yönteminin kullanıldığı model ise en kötü performansı göstermiştir.



Şekil 3.9 En iyi sonuç alınmış öznitelik seçimi algoritmalarının (OneR Funfem, InfoGain Mclust) ve ağırlıklandırılmamış yöntemlerin (Funfem, Mclust) karşılaştırılması.

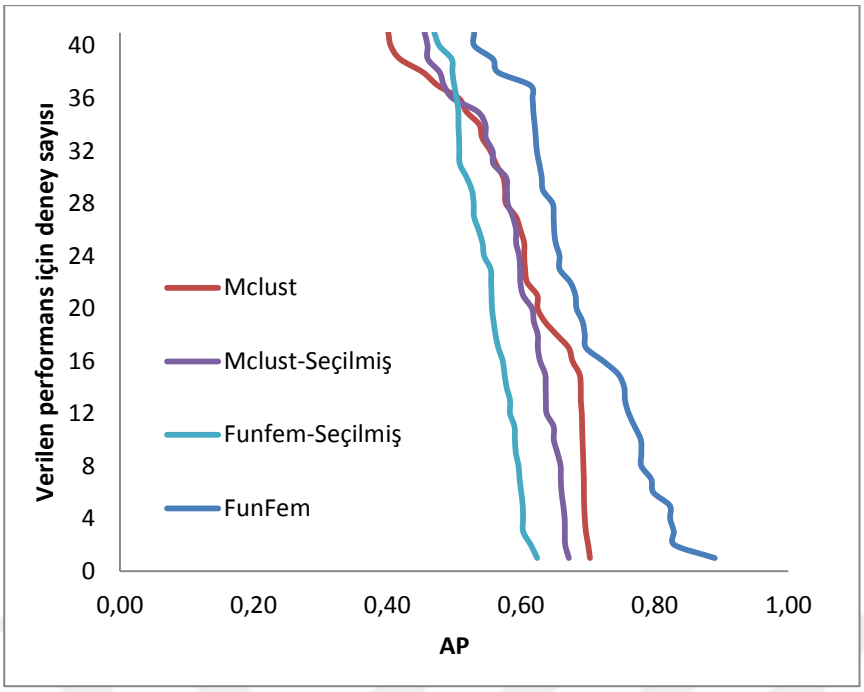
Diğer bir yaklaşımda, ortak gen kümesindeki tüm genlere ağırlık vermek yerine ağırlığı en büyük olan ilk x adet gen (x=200, x=500, x=700,.. x<7078) ile benzerlik hesabı yapılmış, diğer genler her iki deneyde ortak kümeye ait olsa bile

hesaplamaya katılmamıştır. Bu durumda da yöntemin geri getirim performansında herhangi bir iyileşme görülmemiştir. OneR attribute öznitelik seçme algoritması ile seçilen 700 genin deneylerin benzerliği hesabında kullanıldığı durumda hesaplanan AP değerleri aşağıda (Çizelge 3.10) verilmiştir. Sonuçlar Şekil 3.10 da ağırlıklandırılmamış Mclust ve Funfem yöntemleri ile karşılaştırılmalı olarak verilmiştir. Ağırlıklandırılmamış yöntemlerden arama performansı olarak daha iyi sonuç alındığı görülmektedir.

Çizelge3.10 OneR öznitelik seçimi algoritması ile ağırlıklandırılmış (şeçilmiş 700 gen) yöntemlerle ağırlıklandırılmamış yöntemlerin(Funfem, Mclust) karşılaştırılması

	AP			
	Funfem	**Mclust**	**Seçilmiş Mclust**	**Seçilmiş Funfem**
1	0,8901	0,7033	0,6717	0,6243
2	0,8297	0,7006	0,6662	0,6147
3	0,8288	0,6972	0,6657	0,6034
4	0,8232	0,6956	0,6654	0,6032
5	0,8225	0,6946	0,6634	0,6027
6	0,7980	0,6942	0,6608	0,6005
7	0,7950	0,6941	0,6596	0,5980
8	0,7802	0,6934	0,6594	0,5965
9	0,7802	0,6929	0,6546	0,5924
10	0,7794	0,6921	0,6492	0,5910
11	0,7703	0,6915	0,6490	0,5904
12	0,7613	0,6911	0,6384	0,5837
13	0,7558	0,6896	0,6373	0,5837
14	0,7545	0,6895	0,6370	0,5786
15	0,7456	0,6877	0,6362	0,5754
16	0,7230	0,6771	0,6286	0,5728
17	0,6975	0,6715	0,6253	0,5663

18	0,6956	0,6535	0,6252	0,5620
19	0,6923	0,6358	0,6194	0,5592
20	0,6833	0,6249	0,6167	0,5567
21	0,6816	0,6245	0,6033	0,5560
22	0,6739	0,6093	0,5992	0,5554
23	0,6579	0,6062	0,5987	0,5547
24	0,6578	0,6049	0,5973	0,5449
25	0,6519	0,6048	0,5928	0,5425
26	0,6496	0,5990	0,5927	0,5365
27	0,6489	0,5925	0,5878	0,5298
28	0,6475	0,5775	0,5802	0,5293
29	0,6330	0,5764	0,5790	0,5267
30	0,6311	0,5735	0,5777	0,5186
31	0,6276	0,5627	0,5590	0,5086
32	0,6238	0,5544	0,5575	0,5078
33	0,6222	0,5421	0,5475	0,5075
34	0,6206	0,5378	0,5468	0,5064
35	0,6185	0,5182	0,5352	0,5064
36	0,6174	0,5074	0,4983	0,5038
37	0,6141	0,4753	0,4848	0,5003
38	0,5654	0,4532	0,4783	0,4975
39	0,5581	0,4191	0,4606	0,4965
40	0,5305	0,4053	0,4604	0,4781
41	0,5301	0,4016	0,4559	0,4705



Şekil 3.10 OneR öznitelik seçimi algoritması ile ağırlıklandırılmış (şeçilmiş 700 gen) yöntemlerle ağırlıklandırılmamış yöntemlerin(Funfem, Mclust) karşılaştırılması.

Zaman serisi mikrodizi deneylerinde 2 den fazla zaman noktası için ölçümler alınmaktadır. Fakat iki nokta arasındaki zaman deneyden deneye değişmektedir. Bazı deneylerde dakika, bazı deneylerde ise aylar sonra ikinci ölçümler yapılmaktadır. Bu durumun arama sonuçları üzerinde etkisini araştırmak için, öncelikle deneyler kategorize edilerek gruplara ayrılmıştır: saat,dakika, gün, ay. Fakat gruplardaki deney sayıları test yapılabilmesi için yeterli olmadığı için saat ve dakika grupları birleştirilere tek grup oluşturulmuştur. Bu durumda 21'i meme kanseri olmak üzere 56 deney üzerinde kümeleme yöntemleri ile içerik tabanlı arama yapılmıştır.

Alınan sonuçlara göre her ne kadar en iyi ROC skoru karma gruba göre daha iyi olsa da averaj ROC skorları karma gruba oranla daha kötü çıkmıştır (Çizelge 3.11). Bunun yanı sıra Mclust yöntemi ortalama ROC skorlarında homojen veri üzerinde karma veriye göre daha iyi sonuç vermiştir.

Çizelge 3.11. Homojen ve karma veriler üzerinde denenen yöntemlerin ROC skorları

	En iyi ROC skor		Averaj ROC skor	
Yöntem	Homojen veri	Karma veri	Homojen veri	Karma veri
Funfem	0,912925	0,899075	0,697635	0,758857
Mclust	0,757823	0,629721	0,651312	0,568345

Bütün yapılan analizlerin ötesinde, meme kanseri ile ilgili deneyler için arama yapıldığında, Funfem kümeleme-tabanlı yöntemin getirdiği listenin üst sıralarında yer alan ve doğrudan meme kanseri deneyi ile ilgili olmayan deneyler için bir inceleme yapılmıştır. Bu çalışmadan birkaç örnek verecek olursak: GSE9936 (meme kanseri ile ilgili) deneyi sorgulandığında, getirilen sonuç listenin 6. sırasında prostat kanseri ile ilgili bir deney GSE7868 bulunmaktadır. Meme ve prostat kanserlerden biri östorojen diğeri androjen hormonuna bağlı endokrin kanseri türü olduğu görülmektedir. Aynı durum GSE11506 ile GSE8702 deneyleri için de bulunmuştur. Diğer bir örnek, meme kanseri GSE18494 deneyi sorgulandığında yöntem GSE8066 deneyini ilgili olarak getirmektedir fakat bu deney Neuroblastomas ile ilgili bir deneydir. Detaylara inildiğinde GSE18494 deneyinde U87 glioma kültürü ile ifade seviyelerindeki değişimi inceledikleri görülmüştür ve glioma ile Neuroblastomas sinir sistemi kaynaklı habasettir.

Başka bir deyişle kümeleme-tabanlı yöntemin benzer deneyler olarak belirlediği fakat bizim etiketimize göre benzer olmayan ve ROC/ MAP hesabı yaparken yanlış olarak değerlendirdiğimiz sonuçlar da mevcuttur.

3.2 Diğer kanserler ile ilintili sorgular

Bu bölüme kadar yapılan testlerde sorgu olarak meme kanseri deneyleri kullanılmış ve arama sonuçlarının getirdiği deneylerde meme kanseri deneylerinin öncelikli olarak getirilip getirilmediğine göre sistemin duyarlılığı hesaplanmıştır. Burada sorgu olarak test verileri içinde meme kanserinden sonra sayı olarak en çok bulunan periodontal bir deneyi kullanarak yöntemler test edilmiştir. Test verilerinin (toplam 99 deney) içinden 15 adeti (GSE6751) periodontal deneyler oluşturmaktadır. Yapılan arama sonucunda 15 deney için ortalama duyarlılıklar hesaplanmış ve dağılımları violin grafikte verilmiştir (Şekil 3.11). Geliştirilen üç yöntem içinde yapılan arama sonuçlarına göre hesaplanan AP değerleri Çizelge-3.12 de verilmiştir.

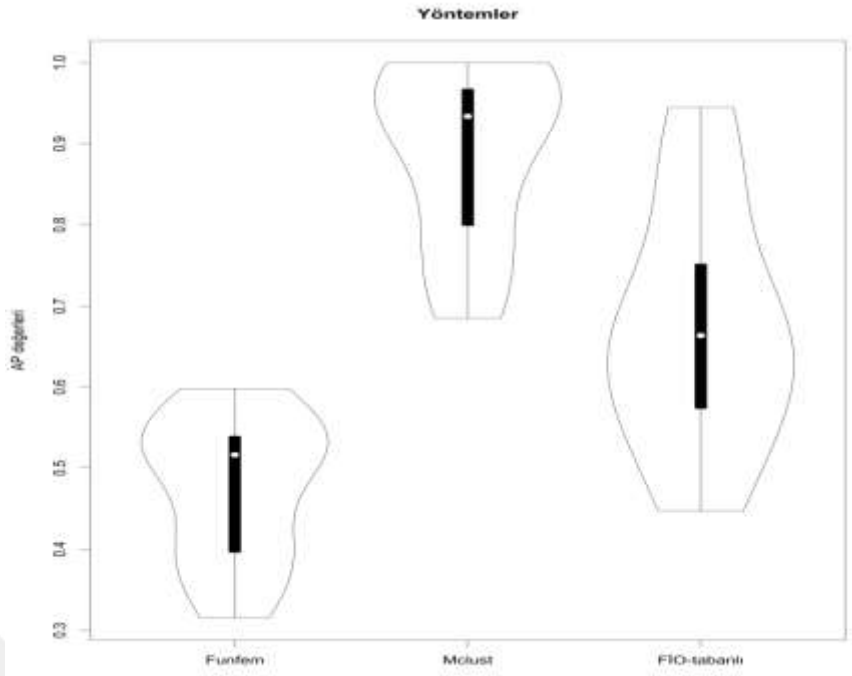
Çizelge 3.12 Periodontal deney verileri sorgulandığında hesaplanan AP değerleri

AP		
Funfem	**Mclust**	**FİO-tabanlı**
0,5978	1,0000	0,9452
0,5540	1,0000	0,9151
0,5449	0,9741	0,8461

0,5448	0,9690	0,8108
0,5343	0,9667	0,6930
0,5296	0,9667	0,6699
0,5288	0,9667	0,6659
0,5165	0,9341	0,6637
0,5005	0,9125	0,6426
0,4186	0,8761	0,6192
0,3986	0,8045	0,6040
0,3936	0,7937	0,5430
0,3848	0,7348	0,5418
0,3461	0,7165	0,4782
0,3152	0,6855	0,4473

Çizelge 3.13 Periodontal deney verileri sorgulandığında hesaplanan MAP değerleri

Yöntem	**MAP**
FİO tabanlı	0,6724
Mclust (Kümeleme-tabanlı)	0,8867
Funfem (Kümeleme-tabanlı)	0,4739



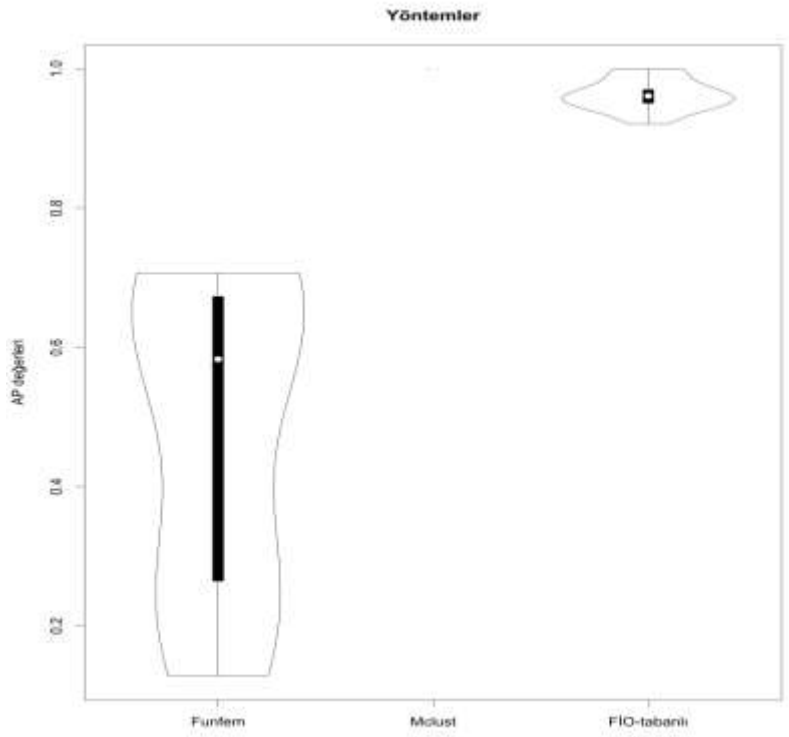
Şekil 3.11 Periodontal deney verileri sorgulandığında AP değerlerinin dağılımı

Çizelge 3.13 de periodontal deneyi aradığımız durumda en iyi sonucu bir kümeleme yöntemi olan Mclust'ın verdiği en kötü sonucun genleri Funfem algoritması ile kümelediğimiz yöntemden alındığı görülmektedir. Şekil 3.11 de verilen violin şekillerinde Mclust yöntemi geri getirim performansı açısından mükemmel sonuca çok yakındır. FİO-tabanlı yöntemde ise değerler 0.8-0.5 aralığında sıklıkla tekrar etmektedir. Funfem yöntemi ise 0-6 ile 0.3 arasında değişen AP değerleri ile periodontal deneyler üzerinde en kötü duyarlılığa sahip yöntem olmuştur.

Meme kanseri ve periodontal deneylerden sonra en fazla sayıdaki (12 adet) test verisi diabet deneylerine aittir. Yapılan arama sonucunda 12 deney için ortalama duyarlılıklar hesaplanmış ve dağılımları violin grafikte verilmiştir (Şekil 3.12). Geliştirilen üç yöntem içinde yapılan arama sonuçlarına göre hesaplanan AP değerleri Çizelge-3.14 de verilmiştir. Bu değerlere göre Mclust yöntemi, tüm diğer diabet deneyleri ilk sırada geri getirmiş ve duyarlılık 1 olarak hesaplanmıştır. En kötü geri getirim performansı ise 0.1 ve 0.7 arasında değişen değerlerle Funfem kümeleme tabanlı yönteme aittir.

Çizelge 3.14 Diabet verileri sorgulandığında hesaplanan AP değerleri

AP		
Funfem	**Mclust**	**FİO-tabanlı**
0,707285	1	1
0,707053	1	1
0,69986	1	0,988095238
0,664287	1	0,964285714
0,639523	1	0,964285714
0,61592	1	0,964285714
0,55015	1	0,958333333
0,283499	1	0,956666667
0,268492	1	0,952380952
0,251926	1	0,946078431
0,181837	1	0,944204981
0,128377	1	0,921245421



Şekil 3.12 Diabet verileri sorgulandığında AP değerlerinin dağılımı

Şekil 3.12 de test verilerinde sorgu olarak diabet deneyleri arandığında hesaplanan AP değerlerinin violin çizgesi görülmektedir. Çizelge 3.14 de de görüldüğü gibi Mclust yönteminde diabet deneyi sorgu olarak kullanıldığında tüm diğer diabet deneylerini benzer deney olarak listelemektedir. Bu sebeple violin çizgesinde Mclust 1.0 noktasında görülmektedir. Kümeleme-tabanlı Mclust yöntemi Diabet verileri için mükemmel geri getirim performansına sahiptir.

Yöntemler prostat kanseri (7 adet) verileri ile denendiğinde de Mclust yönteminin Funfem yönteminden daha iyi sonuç verdiği görülmektedir. Bunun sebebi test verilerinde meme kanseri olmayan deneylerin (periodontal, prostat, diabet) 6 zaman noktasından daha az sayıda zaman noktasına sahip olması ve veri kaybının yaşanmamış olması olarak yorumlanabilir.

3.3 İstatistiksel Önem Testleri

Yöntemlerin istatistiksel açıdan anlamlı olup olmadığını test etmek için Anova1, t-test ve randomizasyon testleri yapılmıştır. Testlerin detayları Ek-3 de verilmiştir.

ROC skorları ile hesaplanan Anova1 testi p-değerleri Çizelge 3.15'de özetlenmiştir.

Çizelge 3.15 Anova1 testi ile hesaplanan p-değerleri (ROC skorlar)

Yöntem	FİO-tabanlı	Kmeans (Kümeleme-tabanlı)	Funfem (Kümeleme-tabanlı)
Mclust (Kümeleme-tabanlı)	<2e-16	1.45e-14	1.4e-15
Funfem (Kümeleme-tabanlı)	<2e-16	<2e-16	*
K-means (Kümeleme-tabanlı)	<2e-16	*	*

Dört farklı yöntemde hesaplanan AP değerleri tüm ikili kombinasyonlarda Anova1 testindeki p-değeri 2e-16 değerinden küçük çıkmıştır.

ROC skorları ile hesaplanan t-test p-değeri sonuçları Çizelge 3.16 de AP değerleri ile hesaplanan p-değeri sonuçları ise Çizelge 3.17 de verilmiştir.

Çizelge 3.16 ROC skorları ile hesaplanan t-test p-değerleri

Yöntem	FİO-tabanlı	K-means (Kümeleme-tabanlı)	Funfem (Kümeleme-tabanlı)
Mclust (Kümeleme-tabanlı)	9.791e-05	2.242e-08	< 2.2e-16
Funfem (Kümeleme-tabanlı)	6.98e-07	<2.2e-16	*
K-means (Kümeleme-tabanlı)	1.198e-11	*	*

Çizelge 3.17 AP değerleri ile hesaplanan t-test p-değerleri

Yöntem	FİO-tabanlı	K-means (Kümeleme-tabanlı)	Funfem (Kümeleme-tabanlı)
Mclust (Kümeleme-tabanlı)	0.7228	6.38e-10	5.341e-50
Funfem (Kümeleme-tabanlı)	0.001147	<2.2e-16	*
K-means (Kümeleme-tabanlı)	2.389e-08	*	*

Randomizasyon testi için 100.000 rasgele permütasyon üretilmiştir. AP değerleri ile hesaplanan randomizasyon testi p-değerleri Çizelge 3.18 de verilmiştir. Çizelge 3.19 de ROC değerleri ile hesaplanan randomizasyon testi p-değerleri verilmiştir.

Çizelge 3.18 AP değerleri ile hesaplanan randomizasyon testi p-değerleri

Yöntem	FİO-tabanlı	K-means (Kümeleme-tabanlı)	Funfem Kümeleme-tabanlı)
Mclust (Kümeleme-tabanlı)	3e-05	0	1e-04
Funfem (Kümeleme-tabanlı)	1e-05	0	*
K-means (Kümeleme-tabanlı)	0	*	*

Çizelge 3.19 ROC değerleri ile hesaplanan randomizasyon testi p-değeri

Yöntem	FİO-tabanlı	K-means (Kümeleme-tabanlı)	Funfem (Kümeleme-tabanlı)
Mclust (Kümeleme-tabanlı)	0.72197	0	5e-05
Funfem (Kümeleme-tabanlı)	0.00121	0	*
K-means (Kümeleme-tabanlı)	0	*	*

Hesaplanan p-değerlerini risk derecesine -temelde doğru olan null hipotezinin reddedilme olasılığı- göre değerlendirecek olursak: yukarıda verilen tüm tablolarda (Mclust ve FİO tabanlı yöntemin t-test ve randomizasyon testi hariç) p-değerleri 0.05 den -yaptığımız çalışmada genel risk derecesi olarak %5=0.05 kullanılmaktadır- daha düşük bir değer çıkmıştır. Bu durumda Null hipotezi H_0 –her iki sistemin aynı olduğu- red edilmiştir. Başka bir ifade ile herhangi iki sistem için hesaplanan değerler raslantısal olarak elde edilmemiş ve istatistiksel olarak anlamlıdır sonucuna varılmaktadır.

4. SONUÇ

Mikrodizi gen ifade veri tabanlarında sunulan veriler (deneyler) birçok akademik çalışmanın temelini oluşturmaktadır. Bu bağlamda bilim adamlarının ihtiyaç duyduğu verinin hızlı ve güvenilir biçimde geri getiriminin sağlanması önem taşımaktadır. Birçok mikrodizi veritabanının arama için özel fonksiyonu bulunmkatadır. Bu arama fonksiyonları üst-veri tabanlıdır ve deneylerin eksik kategorizasyonu, etiketlerin yanlış verilmesi gibi nedenlerden amacına ulaşmakta zorluk çekmektedir. Bu sebeple, deneyleri anahtar kelime ile aramak yerinde bir deneyin tamamını sorgu olarak kullanan ve içerik olarak benzerlerinin geri getirimini hedefleyen yöntemler önerilmiştir. Önerilen içerik-tabanlı yöntemlerin çoğunluğunda farklı ifade olmuş genler kullanılarak benzer deneylerin bulunması yoluna gidilmiştir. Bu tez çalışmasında ise her bir deneyde bulunan genler model tabanlı kümeleme yöntemi ile kümelenmiştir. Deneylerin benzerliği ölçülürken iki farklı deneyde aynı kümeye atanan aynı genler dikkate alınmıştır. Ayrıca tez çalışmasında kullanılan veriler, mikrodizi gen ifade veritabanlarının çoğunluğunu içeren zaman serisi deneyleridir.

Bu tez çalışması için Hayran ve ark. (2014) tarafından zaman serileri deneyleri için geliştirilmiş farklı ifade olmuş gen tabanlı (FİO-tabalı) yöntem temel alınmış ve önerilen kümeleme-tabanlı yöntem ile karşılaştırmalı olarak verilmiştir. Her iki çalışma da GEO'dan toplanan test verileri üzerinde denenmiştir.

FİO-tabanlı yöntemde hedef deneylerdeki FİO genleri tanımlamalarını doğru yapmaktadır. FİO gen tanımı yapılırken Nudge yöntemi kullanılarak her bir genin FİO olma olasılıkları hesaplanarak bir parmak izi dosyası oluşturulmuştur. Parmak izleri iki farklı şekilde hesaplanmıştır. Birincisinde zaman serisi deneylerinde bulunan zaman noktalarından ilk ve son zaman noktasında ölçülen ifade değeri hesaplamaya dâhil edilmiştir (Son_FİO). Diğerinde ise ilk zaman noktası ile tüm zaman noktaları arasında en yüksek FİO gen olma olasılığı kullanılmıştır (Max_FİO). FİO genler belirlendikten sonra deneylerin benzerlikleri farklı benzerlik ölçütleri ile sınanmıştır; Pearson katsayısı, Spearman sıralama katsayısı, Öklit uzaklığı ve Tanimoto katsayısı. Test verisinde bulunan deneylerdeki tüm genler hesaplamaya dâhil edildiği durumda hesaplanan ROC skorlarına göre en iyi sonuç Son_FİO yaklaşımı ile benzerlik ölçütü olarak Tanimoto katsayısı kullanılan duruma aittir. Hesaplamalarda deneydeki tüm genler yerinde, sadece test verisindeki deneylerde ifade değeri ölçülen ortak genler (Kesişim gen kümesi-7078 gen) kullanıldığı durumda ise en yüksek ROC

skoru Son_FİO yaklaşımı ile benzerlik ölçütü olarak Pearson korelasyon katsayısı kullanıldığında alınmıştır.

Kümeleme-tabanlı ve FİO-tabanlı içerik tabanlı arama yöntemi için aynı test verileri kullanılmıştır. Kümeleme-tabanlı yöntemde parmak izleri kesişim gen kümesi esas alınarak oluşturulmuştur. Bu sebeple iki yöntem karşılaştırılırken FİO-tabanlı yöntemde kesişim verisi için en iyi sonuç alınan, benzerlik ölçütü olarak Pearson korelasyon katsayısının kullanıldığı Son_FİO yaklaşımı temel alınmıştır.

Kümeleme-tabanlı yöntemde genler hiyerarşik model tabanlı Mclust ve Funfem algoritmaları ile kümelenmiştir. Bu algoritmalara ek olarak en temel kümeleme algoritmalarından K-means algoritması ile de genler kümelenerek sonuçlar karşılaştırılmıştır. Kümeleme-tabanlı oluşturulan parmak izi dosyasına gen sembolleri ve o genin atandığı küme kaydedilmiştir. Oluşturulan parmak izleri ile iki deneyin benzerliği için deneylerde bulunan aynı genin atandığı kümeye bakılmıştır. Örneğin benzerliği hesaplanacak iki parmak izinde bulunan A geni her iki parmak izinde de X kümesine atanmış ise ve bu şekilde aynı genin aynı kümeye atanma sayısı ne kadar fazla ise deneyler o kadar birbirine benzerdir yaklaşımı uygulanmıştır. Farklı kümeleme algoritmaları ile oluşturulan parmak izleri ile içerik-tabanlı arama yapılmış, sorgu olarak verilen bir meme kanseri deneyi için test verisinde bulunan diğer meme kanseri ile ilgili deneylerin ne kadar ön sırada getirildiğine bakılmıştır. Sonuçlar ROC skorları ve MAP değerleri ile değerlendirilmiştir.

Kümeleme neticesinde alınan sonuçlar yine aynı test verileri üzerinde denenen FİO gen tabanlı içerik tabanlı arama yönteminde alınan sonuçlarla karşılaştırıldığında Funfem kümeleme yönteminin diğerlerine göre daha iyi sonuç verdiği görülmüştür. Funfem'in daha iyi sonuç vermesindeki en büyük etken, deneyleri kümelerken tüm zaman noktalarındaki ölçümlerin hesaba katılmasıdır. Oysaki Mclust yönteminde sadece 6 zaman noktası için kümeleme yapılmış ve veri kaybına sebep olunmuştur. Kmeans kümeleme algoritması ile yapılan içerik-tabanlı arama yöntemi en kötü sonuçları vermiştir bunun temel sebebi K-means algoritmasının zaman serisi verileri için uygun olmayışı ve Mclsut yönteminde olduğu gibi veri kaybının yaşanmış olmasıdır.

Test verilerinde meme kanseri olmayan bir deney arandığında ise yine kümeleme tabanlı yaklaşım olan Mclust diğerlerine göre iyi sonuç vermiştir. Bunun sebebi ise test verilerinde meme kanseri olmayan deneylerin (periodontal, prostat, diabet) 6 zaman noktasından daha az sayıda olması ve veri kaybının yaşanmamış olmasıdır.

Bunlara ek olarak genlerin, deneyler üzerindeki ağırlığını belirlemek ve içerik-tabanlı arama sonuçlarını iyileştirmek için özellik seçimi algoritmaları kullanılmıştır. Özellik seçimi algoritmaları (OneR nitelik algoritması, Relief algoritması ve Bilgi kazanımı algoritması) ile hesaplanan ağırlıklar iki deneyin benzerliği karşılaştırılırken kullanılmıştır. Ağırlıksız yöntemde aynı kümeye düşen aynı gen için ağırlık 1 olarak hesaplanırken burada algoritmanın hesapladığı ağırlık esas alınmıştır. Alınan sonuçların, ağırlıksız yönteme göre bir iyileştirme sağlamadığı görülmüştür.

Bu çalışma gen ifade veri tabanlarında bulunan zaman serisi deneyleri üzerinde içerik-tabanlı geri getirimi yapan ilk kümeleme-tabanlı yöntemdir. Ayrıca bilgi geri getirimi ve biyoenformatik alanlarında çalışan araştırmacılar için önemli bir tartışma açmaktadır.

DNA metilasyonu, gen ifadesi ile beraber biyolojik senaryonun anlamlılığını ölçen önemli bir epigenetik faktördür (Maulik et. al., 2015). Bu tez çalışmasında gen ifade veritabanlarında kullanılmak için tasarlanan kümeleme-tabanlı içerik-tabanlı arama yöntemi metilasyon veritabanı için de uyarlanabilir. Bunun için geri getirim modelinin parametrelerinin tekrar yapılandırılması gerekmektedir. Gelecek çalışmalar kapsamında sistemin metilasyon veri tabanından elde edilen temsili bir veri kümesi üzerinde uygulanması yeralmaktadır. Gen ifade verileri ve metilasyon verilerini eşleştirerek platformlar arası arama yapan içerik-tabanlı deney geri getirimi çalışmalarının pratikte uygulamasının yararlı olacağı düşünülmektedir.

KAYNAKLAR DİZİNİ

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data*, pp 94–105.

Aktürk, Z., 2009, Atatürk Üniversitesi Tıp Fakültesi 2009-2010 dönem biyoistatistik ders notları http://aile.atauni.edu.tr/ogrenciler/

Ball C. A., Brazma A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J. C., Parkinson, H., Quackenbush, J., Ringwald, M., Sansone, S. A., Sherlock, G., Spellman, P., Stoeckert, C., Tateno,Y., Taylor, R., White, J., and Winegarden, N., 2004 Submission of microarray data to public repositories, *PLoS Biol* 2: e317 pp.

Bandyopadhyay, S., Mallik, S. and Mukhopadhyay, A., 2014. A survey and comparative study of statistical tests for identifying differential expression from microarray data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, *11*(1), pp.95-115.

Banfield, J.D. and Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pp.803-821.

Barrett, D.B., Troup, S.E., Wilhite, P., Ledoux, D.,Rudnev, C., Evangelista, I.F.,Kim, A., Soboleva, M. Tomashevsky, K.A., Marshall, K.H., Phillippy, P.M.,Sherman, R.N., Muertter, R. Edgar, 2009 NCBI GEO: archive for high-throughput functional genomic data, *Nucleic Acids Res*, 37 Database: D885-90.

Beal, M. and Krishnamurthy, P., 2012. Gene expression time course clustering with countably infinite hidden Markov models. *arXiv preprint arXiv:1206.6824*.

Beal, M.J. and Krishnamurthy, P., 2006, Clustering Gene Expression Time Course Data with Countably Infinite Hidden Markov Models, Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence UAI, 13-16 July, MIT.

Bell, F. and Sacan, A., 2012, April. Content based searching of gene expression databases using binary fingerprints of differential expression profiles. In *Health Informatics and Bioinformatics (HIBIT), 2012 7th International Symposium on* 107-113 pp.

KAYNAKLAR DİZİNİ (devam)

Ben-Dor A., Shamir R. and Yakhini Z., 1999. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297.

Berber T., Alpkoçak A., 2009, İçerik Tabanlı Tıbbi Görüntü Erişimi Için Bütünleştirilmiş Bilgi Erişimi Modeli, *VI. Ulusal Tıp Bilişimi Kongresi*,.344-351ss.

Berkhin, P., 2006. A survey of clustering data mining techniques. In*Grouping multidimensional data* Springer Berlin Heidelberg.,25-71 pp.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, pp.993-1022.

Bouveyron, C., Côme, E. and Jacques, J., 2014 The discriminative functional mixture model for the analysis of bike sharing systems. Preprint HAL n.01024186, University Paris Descartes,.

Bouveyron, C. and Brunet, C., 2012. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, *22*(1), pp.301-324.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C. and Gaasterland, T., 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature genetics*, *29*(4), 365-371 pp.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G. and Oezcimen, A., 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic acids research*, *31*(1), 68-71 pp.

Buntine, W. and Jakulin, A., 2004, July. Applying discrete PCA in data analysis. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 59-66). AUAI Press.

Caldas, J., Gehlenborg, N., Faisal, A., Brazma, A. and Kaski, S., 2009. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *BMC Bioinformatics*, *10*(Suppl 13), p.P1.

Celeux, G. and Govaert, G., 1995. Gaussian parsimonious clustering models. *Pattern recognition*, *28*(5), pp.781-793.

KAYNAKLAR DİZİNİ (devam)

Chen, R., Mallelwar, R., Thosar, A., Venkatasubrahmanyam, S. and Butte, A.J., 2008. GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed. *BMC bioinformatics*,*9*(1), 548 p.

Cheng Y., Church GM. 2000.Biclustering of expression data. *Proceedings of the Eighth International Conferenceon Intelligent Systems for Molecular Biology (ISMB)*, 8: pp 93–103,

Das, R., Kalita, J. and Bhattacharyya, D.K., 2009. A new approach for clustering gene expression time series data. *International Journal of Bioinformatics Research and Applications*, *5*(3), pp.310-328.

Dean, N. and Raftery, A.E., 2005. Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC bioinformatics*, *6*(1), p.173.

Dirican, A., "Tanı testi performanslarının değerlendirilmesi ve kıyaslanması", http://www.ctf.istanbul.edu.tr/dergi/online/2001v32/s1/011a4.htm (Erişim tarihi 29 Ocak 2016)

EBI- European Bioinformatics Institute -,"ArrayExpress" https://www.ebi.ac.uk/arrayexpress/ (Erişim tarihi: 14 Mayıs 2016)

Engreitz, J.M., Morgan, A.A., Dudley, J.T., Chen, R., Thathoo, R., Altman, R.B. and Butte, A.J., 2010a. Content-based microarray search using differential expression profiles. *BMC bioinformatics*, *11*(1), p.603.

Engreitz, J.M., Daigle, B.J., Marshall, J.J. and Altman, R.B., 2010b. Independent component analysis: mining microarray data for fundamental human gene expression modules. *Journal of biomedical informatics*, *43*(6), pp.932-944.

Engreitz, J.M., Chen, R., Morgan, A.A., Dudley, J.T., Mallelwar, R. and Butte, A.J., 2011. ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics*, *27*(23),3317-3318 pp.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), pp.861-874.

Feng, C., Araki, M., Kunimoto, R., Tamon, A., Makiguchi, H., Niijima, S., Tsujimoto, G. and Okuno, Y., 2009. GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. *BMC genomics*,*10*(1), 411 p.

KAYNAKLAR DİZİNİ (devam)

Fraley, C. and Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), pp.611-631.

Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H. and Horton, P., 2007. CellMontage: similar expression profile search server. *Bioinformatics*, *23*(22), 3103-3104 pp.

Getz G., Levine E. and Domany E., 2000**.** Coupled two-way clustering analysis of gene microarray data..*Proc. Natl. Acad. Sci. USA*, Vol. 97(22):12079–12084,

Georgii, E., Salojärvi, J., Brosché, M., Kangasjärvi, J. and Kaski, S., 2012. Targeted retrieval of gene expression measurements using regulatory models. *Bioinformatics*, *28*(18), pp.2349-2356.

Griffiths, D.M.B.T.L. and Tenenbaum, M.I.J.J.B., 2004. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, *16*, p.17.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeek M., Mesirov J.P., Coller H., Loh M.L.,Downing J.R., Caligiuri M.A., Bloomfield D.D., and Lander E.S. 1999.Molecular classification of cancer:Class discovery and class prediction by gene expression monitoring. Science, Vol. 286(15):531–37,

Halkidi, M., Batistakis, Y. and Vazirgiannis, M., 2001. On clustering validation techniques. *Journal of intelligent information systems*, *17*(2), pp.107-145.

Hayran, A., Ogul, H. and Ozkoc, E., 2014, September. Content-Based Search on Time-Series Microarray Databases. In *Database and Expert Systems Applications (DEXA), 2014 25th International Workshop on* (pp. 89-93). IEEE.

Holte, R.C., 1993. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, *11*(1), pp.63-90.

Horton, P.B., Kiseleva, L. and Fujibuchi, W., 2006. RaPiDS: an algorithm for rapid expression profile database search. *Genome Informatics*, *17*(2), 67-76 pp.

KAYNAKLAR DİZİNİ (devam)

Hunter, L., Taylor, R.C., Leach, S.M. and Simon, R., 2001. GEST: a gene expression search tool based on a novel Bayesian similarity metric.*Bioinformatics*, *17*(suppl 1), S115-S122 pp.

Jaakkola, T. and Haussler, D., 1999. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pp.487-493.

Jiang, D., Pei, J. and Zhang, A., 2003, March. DHC: a density-based hierarchical clustering method for time series gene expression data. *Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on* IEEE. pp. 393-400.

Jiang, D., Tang, C. and Zhang, A., 2004. Cluster analysis for gene expression data: a survey. *Knowledge and Data Engineering, IEEE Transactions* on, 16(11), pp.1370-1386.

Kass, R.E. and Raftery, A.E., 1995. Bayes factors. *Journal of the american statistical association*, *90*(430), pp.773-795.

Kohonen, T., 1985. Self-organization and associative memory. *Applied Optics*, *24*, pp.145-147.

Kuenzel, L., 2010. Gene clustering methods for time series microarray data. *Biochemistry*, *218*.

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. and Reich, M., 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, *313*(5795), pp.1929-1935.

Lazzeroni, L. and Owen A., 2002. Plaid models for gene expression data. *Statistica Sinica*, 12(1), pp 61–86.

MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* Vol. 1, No. 14,. 281-297 pp.

Mahmood, A. and Lei, W., 2009. One-RM: An Improved One-Rule Classifier.*Bangladesh Journal of Scientific and Industrial Research*, *44*(2), pp.171-180.

KAYNAKLAR DİZİNİ (devam)

Maulik U, Mallik S, Mukhopadhyay A, Bandyopadhyay S., 2015. Analyzing Large Gene Expression and Methylation Data Profiles Using StatBicRM: *Statistical Biclustering-Based Rule Mining. PLoS ONE* 10(4): e0119448. doi:10.1371/

McLachlan, G.J. and Basford, K.E., 1988. Mixture models. Inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, *1*.

Mendeş, M., Subaşı, S. and Başpınar, E., 2005. Bilimsel Çalışmalarda P-Değerinin Rapor Edilmesi. Tarım bilimleri dergisi, 11(4),.359-363ss.

Mergen, H., 2015, Microarray teknolojisi, *Hacettepe Üniversitesi, Biyoloji Anabilimdalı, BIO775 Proteomik ve Genomik Ders notları*

Miller, H. and Han, J., 2001. Spatial clustering methods in data mining: a survey. *Geographic data mining and knowledge discovery, Taylor and Francis*.

Molina, L.C., Belanche, L. and Nebot, À., 2002. Feature selection algorithms: A survey and experimental evaluation. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on* (pp. 306-313). IEEE.

Moller-Levet, C.S., Cho, K.H., Yin, H. and Wolkenhauer, O., 2003. *Clustering of gene expression time-series data*. Tech. rep., Department of Computer Science, University of Rostock, Germany.

Murtagh, F., Starck, J.L. and Berry, M.W., 2000. Overcoming the curse of dimensionality in clustering by means of the wavelet transform. *The Computer Journal*, *43*(2), pp.107-120.

National Human Genome Research Institute (NHGRI) "DNA Microarray Technology" http://www.genome.gov/10000533 (Erişim tarihi 07 Mart 2016)

NCBI, "*About GEO DataSets*", http://www.ncbi.nlm.nih.gov/geo/info/datasets.html, (Erişim tarihi: 14 Mayıs 2016)

Novakovic, J., 2009. Using information gain attribute evaluation to classify sonar targets. In *17th Telecommunications forum TELFOR* (pp. 24-26).

Pearson, K., 1896, Mathematical contributions to the theory of evolution. *III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society Ser.* A; 187: 253–318pp.

KAYNAKLAR DİZİNİ (devam)

Raghavan, V.V. and Sever, H., 1995. On the reuse of past optimal queries. *In: Proceedings of 18th ACMInternational Conference on Research and Development in Information Retrieval (SIGIR'95),* Seattle, WA,USA, pp. 344-351.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270 (5235): 467-70pp.

Seltman, H.J., 2015. Experimental design and analysis. Online at: http://www. stat. cmu. edu/, hseltman/309/Book/Book. pdf.

Shamir R. and Sharan R.,2000. Click: A clustering algorithm for gene xpression analysis. *In Proceedingsof the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00). AAAI,Press*

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A. and Eisen, M.B., 2001. The stanford microarray database. *Nucleic Acids Research*,*29*(1), 152-155 pp.**Spearman C.E**., 1904, The proof and measurement of association between two things. American Journal of Psychology, 15: 72–101pp.

Smucker, M.D., Allan, J. and Carterette, B., 2007, November. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 623-632). ACM.

Soydal, İ.,Al.,U. Ve Sezen, U., 2005.İçerik tabanlı görüntü erişim sistemleri, *Bilgi Dünyası* 6(2), 155-170ss.

Stormo,G. D., 2000, DNA Binding sites: representation and discovery, *Bioinformatics* 16 : 16-23pp.

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. and Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), pp.15545-15550.

Syed, A.A., 2004. Performance analysis of K-means algorithm and Kohonen networks. *Florida Atlantic University,* Master of Science.

KAYNAKLAR DİZİNİ (devam)

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, *96*(6), pp.2907-2912.

Tomak,L., Bek Y., 2010, İşlem karakteristik eğrisi analizi ve eğri altında kalan alanların karşılaştırılması, *19 Mayıs Üniversitesi Deneysel ve Klinik Tıp Dergisi* ; 27:58-65ss.

Yoltas, A., Karaboz, İ., 2010, DNA Mikroarray Teknolojisi ve Uygulama Alanları, *Elektronik Mikrobiyoloji Dergisi* 08 : 01-19s.

Zhong, S. and Ghosh, J., 2003. A unified framework for model-based clustering. *The Journal of Machine Learning Research*, *4*, pp.1001-1037.

ÖZGEÇMİŞ

Esma Ergüner Özkoç
Bağlıca mah. 1107 sok. Platform sitesi A blok D3
Etimesgut, Ankara
E-mail : eerguner@gmail.com

KİŞİSEL BİLGİLER

Doğum Tarihi : 9.9.1982
Medeni Durum: Evli

İŞ TECRÜBESİ

2004-2005	Doğu Akdeniz Üniversitesi -KKTC Öğrenci Asistan
2006 -2007	Gebze Yüksek Teknoloji Enstitüsü, Bilgisayar Mühendisliği - Kocaeli Araştırma Görevlisi
2007-2010	Yaşar Üniversitesi, Yazılım Mühendisliği/ Bilgisayar Mühendisliği- İzmir Araştırma Görevlisi
2010- 2015	Başkent Üniversitesi, Bilgisayar Mühendisliği -Ankara Ders saati ücretli Öğretim Görevlisi
2015-	Başkent Üniversitesi, Yönetim Bilişim Sistemleri -Ankara

EĞİTİM BİLGİLERİ

Devam	Ege Üniversitesi Uluslarası Bilgisayar Enstitüsü Doktora, Bilgi Teknolojileri
2008	Gebze Yüksek Teknoloji Enstitüsü Yüksek Lisans, Bilgisayar Mühendisliği
2005	Doğu Akdeniz Üniversitesi Lisans, Bilgi Teknolojileri

YAYINLAR

Erguner, E.; Sogukpinar, I. "Smart card based remote authentication scheme with strengthened security via ECDH" Computer and Information Sciences, 2007. ISCIS 2007. 22nd International Symposium on Computer Information Sciences 7-9 Nov. 2007 Page(s):1- 6, Ankara / Turkey

Özkoç Ergüner, E; Ercan, T. "Yoon- Ryu- Yoo Uzaktan Kimlik Doğrulama Yönteminin Kripto Analizi" 2. Ağ ve Bilgi Güvenliği Ulusal Sempozyumu, 16-18 Mayıs 2008. Girne/KKTC

Ercan, T; Koyuncu, M; Özkoç Ergüner, E. "Grid (Şebeke) Ağlardaki Riskler ve Dağıtık Erişim Denetimi" 2. Ağ ve Bilgi Güvenliği Ulusal Sempozyumu, 16-18 Mayıs 2008. Girne/KKTC

Özkoç Ergüner, E; "Smart Card Based Remote User Authentication Scheme", First JRC-Turkey workshop on ICT security, 1-2 December 2009, İzmir

Hayran, A., Ergüner Özkoç, E., Oğul, H.: "Content-based search in time-series microarray databases." 5th International Workshop on Biological Knowledge Discovery and Data Mining (BIOKDD'14),2014, Munich, Germany

Ergüner Özkoç, E., Oğul, H.: Effect of feature selection on time-series microarray experiment search, 2nd Bionengineering Conferece, (BIOENG'15), İstanbul.

E.E.Özkoç, H.Oğul, (2016), Content-based search on time-series microarray databases using clustering-based fingerprints, Current Bioinformatics (accepted).

KURSLAR

IT Essentials 03.2007- 05.2007 / GYTE
CCNA-1 Network Basics 05.2007- 07.2007 / GYTE

ÖDÜLLER

Doğu Akdeniz Üniversitesi (School of computing and Technology) – 2005 Okul birincisi

HOBİLER

Blog yazarlığı: http://hobipediaofema.blogspot.com/

EKLER

Ek 1 Test Veri Seti

Ek 2 Kesişim Gen Kümesi Gen Sembolleri Listesi

Ek 3 İstatistiksel Önem Testi Sonuçları

Ek 1 Test Veri Seti

	Deney	Hastalık	Zaman Noktası	Tekrar
1	GSE6751_08	non-breast cancer	1h, 2h, 3h, 4h	1
2	GSE6751_11	non-breast cancer	1h, 2h, 3h	1
3	GSE6751_17	non-breast cancer	1h, 2h, 3h, 4h	1
4	GSE6751_23	non-breast cancer	1h, 2h, 3h, 4h	1
5	GSE6751_26	non-breast cancer	1h, 2h, 3h, 4h	1
6	GSE6751_29	non-breast cancer	1h, 2h, 3h, 4h	1
7	GSE6751_32	non-breast cancer	1h, 2h, 3h, 4h	1
8	GSE6751_35	non-breast cancer	1h, 2h, 3h, 4h	1
9	GSE6751_36	non-breast cancer	1h, 2h, 3h, 4h	1
10	GSE6751_38	non-breast cancer	1h, 2h, 3h, 4h	1
11	GSE6751_40	non-breast cancer	1h, 2h, 3h, 4h	1
12	GSE6751_43	non-breast cancer	1h, 2h, 3h, 4h	1
13	GSE6751_45	non-breast cancer	1h, 2h, 3h, 4h	1
14	GSE6751_54	non-breast cancer	1h, 2h, 3h, 4h	1
15	GSE6751_55	non-breast cancer	1h, 2h, 3h, 4h	1
16	GSE9105_S1	non-breast cancer	0h, 30h, 240h	1
17	GSE9105_S10	non-breast cancer	0h, 30h, 240h	1
18	GSE9105_S11	non-breast cancer	0h, 30h, 240h	1
19	GSE9105_S12	non-breast cancer	0h, 30h, 240h	1
20	GSE9105_S2	non-breast cancer	0h, 30h, 240h	1

21	GSE9105_S3	non-breast cancer	0h, 30h, 240h	1
22	GSE9105_S4	non-breast cancer	0h, 30h, 240h	1
23	GSE9105_S5	non-breast cancer	0h, 30h, 240h	1
24	GSE9105_S6	non-breast cancer	0h, 30h, 240h	1
25	GSE9105_S7	non-breast cancer	0h, 30h, 240h	1
26	GSE9105_S8	non-breast cancer	0h, 30h, 240h	1
27	GSE9105_S9	non-breast cancer	0h, 30h, 240h	1
28	GSE35642_50nm	non-breast cancer	0w, 1w, 4w	3
29	GSE35642_5nM	non-breast cancer	0w, 1w, 4w	3
30	GSE31193_IFN	non-breast cancer	0h, 6h, 24h	3
31	GSE31193_IL28B	non-breast cancer	0h, 6h, 24h	3
32	GSE29828_3uM	non-breast cancer	2d, 4d, 6d	3
33	GSE29828_Vehicle	non-breast cancer	2d, 4d, 6d	3
34	GSE27524	non-breast cancer	0h, 18h, 36h, 53h, 72h, 96h	2
35	GSE25417	non-breast cancer	5d, 10d, 15d	3
36	GSE22611_MOCK	non-breast cancer	0h, 2h, 6h	3
37	GSE22611_NOD2-L1007	non-breast cancer	0h, 2h, 6h	3
38	GSE22611_NOD2wt	non-breast cancer	0h, 2h, 6h	3
39	GSE17708	non-breast cancer	0h, 0.5h, 1h, 2h, 4h, 8h, 16h, 24h, 72h	2

40	GSE14429	non-breast cancer	0h, 1h, 6h	4
41	GSE12445_BL_Simut10	non-breast cancer	0d, 1d, 2d	3
42	GSE12445_BL_Simut12	non-breast cancer	0d, 1d, 2d	3
43	GSE12445_BL_siSC	non-breast cancer	0d, 1d, 2d	3
44	GSE12445_224Simut10	non-breast cancer	0d, 1d, 2d, 3d	3
45	GSE12445_224siSC	non-breast cancer	0d, 1d, 2d, 3d	3
46	GSE4655	non-breast cancer	1d, 3d, 5d, 7d, 9d, 11d	2
47	GSE12662	non-breast cancer	0h, 2h, 4h, 8h, 16h, 24h	2
48	GSE8702	non-breast cancer	0w, 3w, 4w, 20w, 44w, 52w	2
49	GSE8066_IMR32c63	non-breast cancer	0h, 24h, 72h, 288h	1
50	GSE8066_IMR32c66	non-breast cancer	0h, 24h, 72h, 288h	1
51	GSE8066_SHEP	non-breast cancer	0h, 8h, 24h, 120h	1
52	GSE8066_SKNAS	non-breast cancer	0h, 24h, 72h, 144h	1
53	GSE7868	non-breast cancer	0h, 4h, 16h	3
54	GSE7259_sample1	non-breast cancer	0d, 5d, 10d	1
55	GSE7259_sample2	non-breast cancer	0d, 5d, 10d	1
56	GSE7223	non-breast cancer	0h, 6h, 24h	3
57	GSE3113	non-breast cancer	0h, 1h, 2h, 4h,	1

			6h, 8h, 12h, 24h	
58	GSE2189_+MGd	non-breast cancer	0h, 4h, 12h, 24h	3
59	GSE28305	breast cancer	4h, 16h, 48h	2
60	GSE20361	breast cancer	0g, 3g, 15g, 30g, 90g, 120g, 150g, 180g	1
61	GSE18494_HEPG2	breast cancer	0h, 4h, 8h, 12h	3
62	GSE18494_MB231	breast cancer	0h, 4h, 8h, 12h	3
63	GSE18494_U87	breast cancer	0h, 4h, 8h, 12h	3
64	GSE11506_e2	breast cancer	3h, 6h	3
65	GSE11352_e2	breast cancer	0h, 12h, 24h, 48h	3
66	GSE11324	breast cancer	0h, 3h, 6h, 12h	3
67	GSE9936_Ad+E2	breast cancer	4h, 24h	3
68	GSE9936_Ad+EQ	breast cancer	4h, 24h	3
69	GSE9936_Ad+HG	breast cancer	4h, 24h	3
70	GSE9936_Ad+IG	breast cancer	4h, 24h	3
71	GSE9936_Ad+LG	breast cancer	4h, 24h	3
72	GSE9936_Ad+veh	breast cancer	4h, 24h	3
73	GSE9936_AdERb+E2	breast cancer	4h, 24h	3
74	GSE9936_AdERb+EQ	breast cancer	4h, 24h	3

75	GSE9936_AdERb+HG	breast cancer	4h, 24h	3
76	GSE9936_AdERb+IG	breast cancer	4h, 24h	3
77	GSE9936_AdERb+LG	breast cancer	4h, 24h	3
78	GSE9936_AdERb+veh	breast cancer	4h, 24h	3
79	GSE7848_actein 20	breast cancer	0h, 6h, 24h	2
80	GSE7848_actein 40	breast cancer	0h, 6h, 24h	2
81	GSE7561_igf	breast cancer	0h, 3h, 24h	3
82	GSE7561_sfm	breast cancer	3h, 24h	2
83	GSE6521_AG1478	breast cancer	5m, 10m, 15m, 30m, 45m, 60m, 90m	1
84	GSE6521_HRG	breast cancer	5m, 10m, 15m, 30m, 45m, 60m, 90m	1
85	GSE6521_HRG_AG1478	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
86	GSE6521_HRG_U0126	breast cancer	0m, 5m, 10m 15m, 30m, 45m, 60m, 90m	1
87	GSE6521_U0126	breast cancer	5m, 10m, 15m, 30m, 45m, 60m, 90m	1
88	GSE6462_EGF_0_1	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1

89	GSE6462_EGF_0_5	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
90	GSE6462_EGF_10_0	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
91	GSE6462_EGF_1_0	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
92	GSE6462_HRG_0_1	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
93	GSE6462_HRG_0_5	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
94	GSE6462_HRG_10_0	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
95	GSE6462_HRG_1_0	breast cancer	0m, 5m, 10m, 15m, 30m, 45m, 60m, 90m	1
96	GSE4917_DEX	breast cancer	30m, 120m, 240m, 1440m	3
97	GSE4917_ETHANOL	breast cancer	30m, 120m, 240m, 1440m	3
98	GSE4668	breast cancer	0d, 1d, 2d	3
99	GSE2241	breast cancer	0h, 6h, 9h	2

Ek 2 Kesişim Gen Kümesi Gen Sembolleri Listesi

AADAC	CHN2	GCLM	MAPK3	PRAME	ST6GALNAC4
AAK1	CHP	GCM1	MAPK4	PRB1	ST7
AAMP	CHPF	GCNT1	MAPK6	PRCC	ST8SIA1
AANAT	CHRD	GCNT2	MAPK7	PRCP	ST8SIA3
AARS	CHRNA1	GDF10	MAPK8	PRDM1	ST8SIA4
AASDHPPT	CHRNA2	GDF11	MAPK8IP1	PRDM2	ST8SIA5
AASS	CHRNA3	GDF15	MAPK8IP2	PRDX1	STAB1
AATF	CHRNA4	GDF5	MAPK8IP3	PRDX2	STAC
AATK	CHRNA5	GDI1	MAPK9	PRDX3	STAG1
ABAT	CHRNA6	GDI2	MAPKAP1	PRDX4	STAG2
ABCA1	CHRNB1	GDPD5	MAPKAPK2	PRDX6	STAM
ABCA12	CHRNB2	GEM	MAPKAPK3	PRELP	STAM2
ABCA2	CHRNB3	GEMIN4	MAPKAPK5	PREP	STAMBP
ABCA3	CHRND	GFAP	MAPKBP1	PREPL	STAR
ABCA4	CHRNE	GFER	MAPRE1	PRF1	STARD13
ABCA5	CHST1	GFI1	MAPRE2	PRG2	STARD3
ABCA6	CHST10	GFPT1	MAPRE3	PRIM1	STARD5
ABCA8	CHST2	GFPT2	MAPT	PRKAA1	STARD7
ABCB1	CHST3	GFRA1	MARCKS	PRKAA2	STARD8
ABCB11	CHST7	GFRA2	MARCKSL1	PRKAB1	STAT1
ABCB4	CHSY1	GFRA3	MARCO	PRKAB2	STAT2
ABCB6	CHUK	GGA2	MARK2	PRKACA	STAT3
ABCB7	CIAPIN1	GGA3	MARK3	PRKACB	STAT4
ABCB8	CIB1	GGCX	MARS	PRKACG	STAT5A
ABCB9	CIB2	GGH	MAS1	PRKAG1	STAT5B
ABCC1	CIC	GGPS1	MASP1	PRKAR1A	STAT6
ABCC10	CIITA	GH2	MASP2	PRKAR2A	STATH
ABCC2	CILP	GHITM	MAST1	PRKAR2B	STAU1
ABCC3	CIRBP	GHR	MAST2	PRKCA	STAU2
ABCC4	CIT	GHRH	MAT1A	PRKCD	STC1
ABCC5	CITED1	GHRHR	MAT2A	PRKCE	STC2
ABCC6	CITED2	GIF	MATK	PRKCG	STEAP1

ABCC8	CIZ1	GIP	MATN1	PRKCH	STIM1
ABCC9	CKAP4	GIPC1	MATN2	PRKCI	STIP1
ABCD1	CKAP5	GIPR	MATN3	PRKCQ	STK10
ABCD2	CKB	GIT2	MATN4	PRKCSH	STK11
ABCD4	CKM	GJA1	MATR3	PRKCZ	STK16
ABCE1	CKMT2	GJA4	MAX	PRKD1	STK17A
ABCF1	CKS2	GJA5	MAZ	PRKD2	STK17B
ABCF3	CLASP2	GJA8	MB	PRKD3	STK19
ABCG1	CLC	GJB1	MBD1	PRKDC	STK24
ABCG2	CLCA1	GJB3	MBD2	PRKG1	STK25
ABHD14A	CLCA2	GJB5	MBD3	PRKG2	STK3
ABHD2	CLCC1	GJC1	MBD4	PRKX	STK38
ABHD3	CLCN1	GK	MBL2	PRL	STK38L
ABHD5	CLCN2	GK2	MBNL1	PRLR	STK39
ABI1	CLCN3	GLA	MBNL2	PRM1	STK4
ABI2	CLCN4	GLB1	MBP	PRM2	STMN1
ABL1	CLCN5	GLCE	MBTPS1	PRMT1	STMN2
ABL2	CLCN6	GLDC	MBTPS2	PRMT2	STOM
ABLIM1	CLCN7	GLG1	MC1R	PRNP	STOML1
ABLIM3	CLDN10	GLI1	MC2R	PROC	STOML2
ABO	CLDN14	GLI2	MC5R	PROCR	STRAP
ABP1	CLDN18	GLI3	MCAM	PRODH	STRN
ABR	CLDN3	GLIPR1	MCC	PRODH2	STRN3
ABTB2	CLDN4	GLMN	MCCC2	PROL1	STS
ACAA1	CLDN5	GLO1	MCF2	PROM1	STT3A
ACAA2	CLDN7	GLP1R	MCF2L	PROP1	STX10
ACACA	CLDN8	GLRA1	MCF2L2	PROS1	STX11
ACACB	CLDN9	GLRA2	MCFD2	PROSC	STX12
ACADL	CLDND1	GLRA3	MCL1	PROX1	STX16
ACADM	CLEC10A	GLRB	MCM2	PROZ	STX1A
ACADS	CLEC11A	GLRX	MCM3	PRPF18	STX2
ACADSB	CLEC3B	GLS	MCM3AP	PRPF19	STX6
ACADVL	CLGN	GLS2	MCM4	PRPF3	STX7
ACAT1	CLIC1	GLT25D2	MCM5	PRPF31	STX8

ACAT2	CLIC2	GLT8D1	MCM6	PRPF4	STXBP1
ACBD3	CLIC4	GLUD1	MCM7	PRPF4B	STXBP2
ACCN1	CLIC5	GLUD2	MCRS1	PRPF8	STXBP3
ACCN2	CLK1	GLUL	MDC1	PRPH	SUB1
ACCN3	CLK2	GLYAT	MDFI	PRPS1	SUCLA2
ACD	CLK3	GM2A	MDH1	PRPS1L1	SUCLG2
ACE	CLMN	GMDS	MDH2	PRPS2	SULF1
ACHE	CLN3	GMFB	MDK	PRPSAP1	SULT1C2
ACIN1	CLN5	GMFG	MDM1	PRPSAP2	SULT2A1
ACLY	CLNS1A	GML	MDM2	PRR3	SULT2B1
ACO1	CLOCK	GMPR	MDM4	PRR4	SULT4A1
ACO2	CLPP	GNA11	MDN1	PRRG1	SUMO1
ACOT11	CLPS	GNA12	ME1	PRRG2	SUOX
ACOT7	CLPTM1	GNA13	ME2	PRRX1	SUPT4H1
ACOT8	CLPX	GNA15	ME3	PRSS12	SUPT5H
ACOX1	CLSTN1	GNAI1	MEA1	PRSS16	SUPT6H
ACOX2	CLSTN3	GNAI2	MECP2	PRSS22	SUPT7L
ACOX3	CLTA	GNAI3	MED12	PRSS23	SUPV3L1
ACP1	CLTB	GNAL	MED6	PRSS3	SURF1
ACP2	CLTC	GNAO1	MED8	PRSS8	SURF2
ACP5	CLTCL1	GNAQ	MEF2A	PRTN3	SUZ12
ACPP	CLU	GNAS	MEF2C	PRUNE	SV2A
ACRV1	CLUAP1	GNAT1	MEF2D	PSAP	SV2B
ACSBG1	CLUL1	GNAZ	MEGF8	PSCA	SVIL
ACSL1	CMA1	GNB1	MEIS2	PSD	SWAP70
ACSL4	CMKLR1	GNB2	MELK	PSD3	SYCP1
ACSL6	CNBP	GNB2L1	MEN1	PSD4	SYCP2
ACSM1	CNGA1	GNB3	MEOX1	PSEN1	SYDE1
ACSM3	CNGA3	GNB5	MEOX2	PSEN2	SYF2
ACTA2	CNGB1	GNE	MEP1A	PSG1	SYK
ACTB	CNIH	GNG11	MEP1B	PSG11	SYMPK
ACTG1	CNIH3	GNG12	MERTK	PSG3	SYN1
ACTG2	CNKSR1	GNG4	MEST	PSG4	SYN2
ACTL6A	CNKSR2	GNG5	MET	PSG9	SYN3

ACTL6B	CNN1	GNG7	METAP1	PSIP1	SYNCRIP
ACTL8	CNN3	GNGT1	METAP2	PSKH1	SYNE1
ACTN1	CNNM2	GNL1	METTL1	PSMA1	SYNE2
ACTN2	CNOT1	GNL2	METTL3	PSMA2	SYNGR1
ACTN3	CNOT2	GNLY	METTL7A	PSMA3	SYNGR2
ACTN4	CNOT3	GNMT	MFAP1	PSMA4	SYNGR3
ACTR1A	CNOT4	GNPAT	MFAP2	PSMA5	SYNGR4
ACTR1B	CNOT8	GNPDA1	MFAP3	PSMB1	SYNJ1
ACTR2	CNP	GNRH1	MFAP3L	PSMB10	SYNJ2
ACTR3	CNR1	GNRH2	MFAP4	PSMB2	SYP
ACVR1	CNR2	GNRHR	MFAP5	PSMB3	SYPL1
ACVR1B	CNTFR	GNS	MFGE8	PSMB4	SYT1
ACVR2A	CNTN1	GOLGA1	MFHAS1	PSMB5	SYT11
ACVR2B	CNTN2	GOLGA2	MFI2	PSMB6	SYT17
ACVRL1	CNTN5	GOLGA3	MFN1	PSMB7	SYT5
ACYP1	CNTN6	GOLGA4	MFN2	PSMB8	T
ACYP2	CNTNAP2	GOLGB1	MFNG	PSMB9	TAC1
ADA	COASY	GORASP2	MFSD5	PSMC2	TACC1
ADAM10	COBL	GOSR1	MFSD7	PSMC3	TACC2
ADAM11	COBLL1	GOSR2	MGAT1	PSMC3IP	TACR1
ADAM12	COBRA1	GOT1	MGAT2	PSMC4	TACR2
ADAM15	COCH	GOT2	MGAT3	PSMC5	TACR3
ADAM17	COG2	GP1BA	MGAT4C	PSMC6	TACSTD2
ADAM18	COG4	GP9	MGAT5	PSMD1	TAF1
ADAM19	COG5	GPA33	MGEA5	PSMD10	TAF10
ADAM2	COG7	GPAA1	MGLL	PSMD11	TAF11
ADAM20	COIL	GPC1	MGMT	PSMD12	TAF13
ADAM22	COL10A1	GPC3	MGP	PSMD13	TAF15
ADAM23	COL11A1	GPC4	MGST2	PSMD14	TAF1A
ADAM28	COL11A2	GPC5	MGST3	PSMD2	TAF1B
ADAM7	COL13A1	GPD1	MIA	PSMD3	TAF1C
ADAM8	COL15A1	GPD1L	MICAL2	PSMD4	TAF2
ADAM9	COL16A1	GPD2	MICB	PSMD5	TAF4
ADAMDEC1	COL17A1	GPKOW	MID1	PSMD6	TAF5

ADAMTS2	COL18A1	GPLD1	MINA	PSMD7	TAF5L
ADAMTS3	COL19A1	GPM6A	MINK1	PSMD8	TAF6
ADAMTSL2	COL1A1	GPM6B	MINPP1	PSMD9	TAF6L
ADAMTSL3	COL1A2	GPNMB	MIPEP	PSME1	TAF7
ADAR	COL21A1	GPR1	MKI67	PSME2	TAF9
ADARB1	COL2A1	GPR107	MKL1	PSME3	TAGLN
ADCY1	COL3A1	GPR116	MKLN1	PSME4	TAGLN2
ADCY2	COL4A1	GPR12	MKNK1	PSMF1	TAGLN3
ADCY3	COL4A2	GPR125	MKNK2	PSPH	TAL1
ADCY6	COL4A3	GPR126	MKRN1	PSTPIP1	TANK
ADCY7	COL4A4	GPR135	MKRN3	PTAFR	TAOK2
ADCY9	COL4A5	GPR137B	MLANA	PTBP1	TAOK3
ADCYAP1R1	COL4A6	GPR143	MLC1	PTCRA	TAP1
ADD1	COL5A2	GPR15	MLF1	PTDSS1	TAP2
ADD2	COL6A1	GPR161	MLF2	PTEN	TAPBP
ADD3	COL6A2	GPR162	MLH1	PTGDR	TARBP1
ADH1A	COL6A3	GPR17	MLH3	PTGDS	TARBP2
ADH1B	COL7A1	GPR171	MLL	PTGER1	TARDBP
ADH1C	COL9A1	GPR176	MLL2	PTGER2	TARP
ADH5	COL9A2	GPR18	MLL4	PTGER3	TARS
ADH6	COL9A3	GPR19	MLLT1	PTGER4	TAT
ADIPOQ	COLEC10	GPR20	MLLT10	PTGES	TATDN2
ADIPOR2	COLQ	GPR3	MLLT11	PTGES3	TAX1BP1
ADK	COMMD4	GPR31	MLLT3	PTGFR	TAX1BP3
ADM	COMP	GPR35	MLN	PTGIR	TAZ
ADNP	COMT	GPR37	MLX	PTGIS	TBC1D1
ADORA1	COPA	GPR37L1	MLXIP	PTGS1	TBC1D22A
ADORA2A	COPB2	GPR39	MMD	PTGS2	TBC1D2B
ADORA2B	COPE	GPR4	MME	PTH	TBC1D4
ADORA3	COPS2	GPR44	MMP1	PTHLH	TBC1D5
ADPRH	COPS3	GPR45	MMP10	PTK2	TBC1D9
ADRA1A	COPS5	GPR50	MMP11	PTK2B	TBC1D9B
ADRA1B	COPS6	GPR56	MMP12	PTK6	TBCA
ADRA1D	COPS7A	GPR6	MMP13	PTK7	TBCC

ADRA2A	COPS8	GPR64	MMP14	PTMA	TBCD
ADRA2C	COQ2	GPR65	MMP15	PTMS	TBCE
ADRB1	COQ7	GPR68	MMP16	PTN	TBKBP1
ADRB2	COQ9	GPR98	MMP17	PTOV1	TBL1X
ADRB3	CORO1A	GPRASP1	MMP19	PTP4A1	TBL2
ADRBK1	CORO2A	GPRC5A	MMP2	PTP4A2	TBL3
ADRBK2	CORO2B	GPRC5B	MMP20	PTPLB	TBP
ADRM1	COX10	GPSM3	MMP24	PTPN1	TBPL1
ADSL	COX11	GPT	MMP3	PTPN11	TBR1
ADSS	COX17	GPX1	MMP7	PTPN12	TBX1
AEBP1	COX4I1	GPX2	MMP8	PTPN13	TBX19
AES	COX5A	GPX3	MMP9	PTPN14	TBX2
AFF1	COX5B	GPX4	MMRN1	PTPN18	TBX5
AFF2	COX6A1	GPX7	MN1	PTPN2	TBX6
AFF3	COX6A2	GRAP	MNAT1	PTPN21	TBXA2R
AFG3L2	COX6B1	GRAP2	MNDA	PTPN22	TBXAS1
AFM	COX6C	GRB10	MNT	PTPN3	TCAP
AFP	COX7A1	GRB14	MOAP1	PTPN4	TCEA2
AGA	COX7A2	GRB2	MOBP	PTPN6	TCEAL1
AGER	COX7B	GRB7	MOCS1	PTPN7	TCEAL4
AGGF1	COX7C	GREB1	MOCS3	PTPN9	TCEB1
AGL	COX8A	GREM1	MOG	PTPRA	TCEB2
AGPAT1	CP	GRHPR	MORC2	PTPRB	TCEB3
AGPAT2	CP110	GRIA1	MORC3	PTPRC	TCERG1
AGPS	CPA1	GRIA2	MOS	PTPRCAP	TCF12
AGR2	CPA2	GRIA3	MPDU1	PTPRD	TCF15
AGRP	CPA3	GRIK1	MPDZ	PTPRE	TCF21
AGT	CPB1	GRIK2	MPG	PTPRF	TCF3
AGTPBP1	CPB2	GRIK3	MPHOSPH10	PTPRG	TCF4
AGTR1	CPD	GRIK5	MPHOSPH6	PTPRH	TCF7
AGTR2	CPE	GRIN1	MPI	PTPRK	TCF7L2
AGXT	CPEB3	GRIN2A	MPL	PTPRM	TCFL5
AHCTF1	CPLX2	GRIN2B	MPO	PTPRN	TCIRG1
AHCY	CPM	GRIN2C	MPP1	PTPRN2	TCL1A

AHCYL1	CPN1	GRIN2D	MPP2	PTPRO	TCN1
AHDC1	CPN2	GRINA	MPP3	PTPRR	TCN2
AHNAK	CPNE1	GRK1	MPP6	PTPRS	TCOF1
AHR	CPNE3	GRK4	MPPE1	PTPRT	TCP1
AHSA1	CPNE6	GRK5	MPPED1	PTPRU	TCP11L1
AHSG	CPS1	GRK6	MPPED2	PTPRZ1	TCTA
AIF1	CPSF1	GRLF1	MPST	PTRF	TDO2
AIM1	CPSF4	GRM1	MPV17	PTS	TDRD3
AIM2	CPSF6	GRM2	MPZ	PTTG1	TDRD7
AIP	CPT1A	GRM3	MPZL1	PTTG1IP	TEAD1
AIPL1	CPVL	GRM4	MR1	PTTG2	TEAD3
AIRE	CPZ	GRM5	MRAS	PTX3	TEAD4
AJAP1	CR1	GRM7	MRC2	PUM1	TEC
AK1	CR2	GRM8	MRE11A	PUM2	TEF
AK2	CRABP1	GRN	MRFAP1L1	PURA	TEK
AKAP1	CRABP2	GRP	MRPL12	PVALB	TEKT2
AKAP10	CRADD	GRPEL1	MRPL19	PVR	TENC1
AKAP11	CRAT	GRPR	MRPL23	PVRL1	TEP1
AKAP12	CREB1	GRSF1	MRPL28	PVRL2	TERF1
AKAP13	CREB3	GSK3A	MRPL3	PWP1	TERF2
AKAP3	CREB3L1	GSK3B	MRPL33	PXN	TERF2IP
AKAP4	CREB3L2	GSN	MRPL40	PYCR1	TERT
AKAP5	CREB5	GSPT1	MRPL49	PYGB	TES
AKAP6	CREBBP	GSR	MRPL9	PYGL	TESK1
AKAP8	CREBL2	GSS	MRPS11	PYGM	TESK2
AKAP8L	CREG1	GSTA1	MRPS12	PYY	TEX261
AKAP9	CRELD1	GSTA4	MRPS14	PZP	TF
AKR1A1	CREM	GSTM1	MRPS18B	ProSAPiP1	TFAM
AKR1B1	CRH	GSTM3	MRPS27	QDPR	TFAP2A
AKR1B10	CRHBP	GSTM5	MRPS31	QKI	TFAP2B
AKR1C1	CRHR1	GSTO1	MS4A1	QPCT	TFAP2C
AKR1C3	CRHR2	GSTP1	MS4A2	QPRT	TFAP4
AKR1C4	CRIM1	GSTT1	MS4A3	QRICH1	TFCP2
AKR1D1	CRIP1	GSTZ1	MSC	R3HDM1	TFDP1

AKR7A2	CRIP2	GTF2A1	MSH2	R3HDM2	TFDP2
AKR7A3	CRISP1	GTF2A2	MSH3	RAB11A	TFE3
AKT1	CRISP2	GTF2B	MSH4	RAB11B	TFEB
AKT2	CRISP3	GTF2E1	MSH6	RAB11FIP2	TFEC
AKT3	CRK	GTF2E2	MSI1	RAB11FIP3	TFF1
ALAD	CRKL	GTF2F1	MSLN	RAB11FIP5	TFF2
ALAS1	CRLF1	GTF2F2	MSMB	RAB13	TFF3
ALAS2	CRLF3	GTF2H1	MSN	RAB14	TFIP11
ALB	CRMP1	GTF2H4	MSR1	RAB1A	TFPI
ALCAM	CROCC	GTF2I	MST1R	RAB21	TFPI2
ALDH1A1	CROT	GTF3C1	MSX1	RAB22A	TFR2
ALDH1A2	CRP	GTF3C2	MSX2	RAB27A	TFRC
ALDH1A3	CRTAM	GTPBP1	MT1E	RAB27B	TG
ALDH1B1	CRTAP	GTPBP6	MT1F	RAB28	TGDS
ALDH1L1	CRTC1	GTSE1	MT1G	RAB30	TGFA
ALDH2	CRX	GUCA1A	MT1H	RAB31	TGFB1
ALDH3A1	CRY1	GUCA1B	MT1X	RAB32	TGFB1I1
ALDH3A2	CRY2	GUCA2A	MT2A	RAB33A	TGFB2
ALDH3B1	CRYAA	GUCA2B	MT3	RAB35	TGFB3
ALDH3B2	CRYAB	GUCY1A2	MT4	RAB36	TGFBI
ALDH4A1	CRYBA1	GUCY1A3	MTA1	RAB3A	TGFBR1
ALDH5A1	CRYBA4	GUCY1B3	MTAP	RAB3B	TGFBR2
ALDH6A1	CRYBB1	GUCY2C	MTDH	RAB3GAP1	TGFBR3
ALDH7A1	CRYBB2	GUCY2D	MTERF	RAB40A	TGFBRAP1
ALDH9A1	CRYBB3	GUCY2F	MTF1	RAB40B	TGIF2
ALDOA	CRYGA	GUK1	MTF2	RAB40C	TGM1
ALDOB	CRYGC	GULP1	MTFR1	RAB4A	TGM2
ALDOC	CRYGD	GUSB	MTG1	RAB4B	TGM3
ALG3	CRYM	GYG2	MTHFD1	RAB5A	TGM4
ALG8	CRYZ	GYPA	MTHFD2	RAB5B	TGM5
ALK	CS	GYPB	MTHFR	RAB5C	TGOLN2
ALKBH1	CSDA	GYPC	MTHFS	RAB6A	TH
ALMS1	CSDC2	GYPE	MTIF2	RAB6B	THBD
ALOX12	CSE1L	GYS1	MTM1	RAB7L1	THBS1

ALOX12B	CSF1	GYS2	MTMR1	RAB8A	THBS2
ALOX15	CSF1R	GZMA	MTMR11	RAB9A	THBS3
ALOX15B	CSF2	GZMB	MTMR2	RABAC1	THBS4
ALOX5	CSF2RA	GZMH	MTMR3	RABEP1	THOC1
ALOX5AP	CSF3	GZMK	MTMR4	RABEPK	THOC2
ALPI	CSF3R	GZMM	MTMR6	RABGAP1	THOC5
ALPK3	CSH1	H1F0	MTMR9	RABGGTA	THOP1
ALPL	CSHL1	H1FX	MTNR1B	RABGGTB	THPO
ALPPL2	CSK	H2AFV	MTR	RABIF	THRA
ALX3	CSN1S1	H2AFY	MTRR	RABL3	THRB
AMACR	CSN2	H2AFZ	MTSS1	RAC1	THUMPD1
AMBP	CSN3	H6PD	MTTP	RAC2	THY1
AMD1	CSNK1A1	HAAO	MTUS1	RAC3	TIA1
AMELX	CSNK1D	HABP2	MTX2	RAD1	TIAL1
AMELY	CSNK1E	HABP4	MUC1	RAD17	TIAM1
AMFR	CSNK1G2	HADHA	MUC4	RAD21	TICAM1
AMH	CSNK2A1	HAGH	MUC7	RAD23A	TIE1
AMHR2	CSNK2A2	HAL	MUSK	RAD23B	TIMELESS
AMIGO2	CSPG4	HAO1	MUT	RAD50	TIMM17A
AMOTL2	CSPG5	HAP1	MUTYH	RAD51	TIMM17B
AMPD1	CSRP1	HAPLN1	MVD	RAD51AP1	TIMM44
AMPD2	CSRP2	HARS	MVK	RAD51C	TIMM8A
AMPD3	CSRP3	HARS2	MVP	RAD51L1	TIMP1
AMPH	CST1	HAS1	MX1	RAD51L3	TIMP2
AMT	CST3	HAS2	MX2	RAD52	TIMP3
ANAPC10	CST5	HAT1	MXD1	RAD54L	TIMP4
ANAPC5	CST6	HAX1	MXD4	RAD9A	TIPARP
ANG	CST7	HBB	MXI1	RAE1	TIPRL
ANGEL1	CSTA	HBD	MXRA5	RAF1	TJP1
ANGEL2	CSTB	HBE1	MYB	RAG1	TJP2
ANGPT1	CSTF1	HBEGF	MYBL2	RAG2	TJP3
ANGPT2	CSTF2	HBP1	MYBPC1	RAGE	TK1
ANGPTL2	CSTF2T	HBS1L	MYBPC2	RAI14	TK2
ANGPTL7	CSTF3	HBXIP	MYBPC3	RALA	TKT

ANK1	CTAG2	HBZ	MYBPH	RALB	TKTL1
ANK2	CTAGE5	HCCS	MYC	RALBP1	TLE1
ANK3	CTBP1	HCFC1	MYCBP	RALGDS	TLE2
ANKMY2	CTBP2	HCK	MYCBP2	RALGPS1	TLE3
ANKRD1	CTBS	HCLS1	MYCL1	RALY	TLE4
ANKRD12	CTCF	HCN2	MYCN	RAMP1	TLK1
ANKRD17	CTDP1	HCN4	MYD88	RAMP2	TLK2
ANKRD26	CTDSP2	HCP5	MYEF2	RAMP3	TLL2
ANKRD28	CTDSPL	HCRT	MYF5	RANBP1	TLN1
ANKRD46	CTF1	HCRTR1	MYF6	RANBP2	TLN2
ANKRD6	CTGF	HCRTR2	MYH10	RANBP3	TLR1
ANKRD7	CTH	HDAC1	MYH11	RANBP6	TLR2
ANKS1A	CTNNA1	HDAC2	MYH2	RANBP9	TLR3
ANP32B	CTNNA2	HDAC3	MYH3	RANGAP1	TLR5
ANP32E	CTNNAL1	HDAC4	MYH6	RAP1A	TLR6
ANPEP	CTNNB1	HDAC5	MYH7	RAP1GDS1	TLX1
ANXA1	CTNNBIP1	HDAC6	MYH8	RAP2A	TLX2
ANXA10	CTNND1	HDAC9	MYH9	RAP2B	TM2D1
ANXA11	CTNND2	HDC	MYL1	RAP2C	TM4SF1
ANXA13	CTNS	HDDC2	MYL2	RAPGEF1	TM4SF4
ANXA2	CTPS	HDGF	MYL4	RAPGEF3	TM4SF5
ANXA3	CTR9	HDGFRP3	MYL7	RAPGEF4	TM7SF2
ANXA4	CTRC	HDLBP	MYL9	RAPSN	TM9SF1
ANXA5	CTRL	HEBP2	MYLK	RARA	TM9SF2
ANXA6	CTSB	HECW1	MYLPF	RARB	TM9SF4
ANXA7	CTSC	HELZ	MYO10	RARG	TMC6
AOAH	CTSD	HEPH	MYO18A	RARRES1	TMCC1
AOC2	CTSE	HERC3	MYO1A	RARRES2	TMCC2
AOC3	CTSF	HERC4	MYO1B	RARRES3	TMCO1
AOX1	CTSG	HERPUD1	MYO1C	RARS	TMED1
AP1B1	CTSH	HES1	MYO1D	RASA1	TMED10
AP1G1	CTSK	HESX1	MYO1E	RASA2	TMED2
AP1G2	CTSL2	HEXA	MYO1F	RASA3	TMED3
AP1S1	CTSO	HEXB	MYO5A	RASAL2	TMED5

AP1S2	CTSS	HEXIM1	MYO6	RASGRF1	TMED9
AP2A2	CTSW	HFE	MYO7A	RASGRP1	TMEFF1
AP2B1	CTSZ	HGF	MYO9B	RASGRP2	TMEM106C
AP2M1	CTTN	HGFAC	MYOC	RASGRP3	TMEM109
AP2S1	CUBN	HGS	MYOD1	RASSF1	TMEM11
AP3B1	CUL1	HHEX	MYOG	RASSF2	TMEM115
AP3B2	CUL2	HIBCH	MYOM1	RASSF7	TMEM123
AP3D1	CUL3	HIC1	MYOM2	RASSF8	TMEM159
AP3M2	CUL4A	HINT1	MYOZ3	RB1	TMEM47
AP3S2	CUL4B	HIP1	MYRIP	RB1CC1	TMEM5
AP4E1	CUL5	HIPK1	MYST1	RBBP4	TMEM59
AP4M1	CUL7	HIPK2	MYST3	RBBP5	TMEM63A
AP4S1	CX3CL1	HIPK3	MYST4	RBBP6	TMEM66
APAF1	CX3CR1	HIRA	MYT1	RBBP7	TMEM87A
APBA1	CXADR	HIRIP3	N4BP1	RBBP8	TMF1
APBA2	CXCL1	HIST1H1A	N4BP3	RBCK1	TMOD1
APBA3	CXCL10	HIST1H1B	NAALAD2	RBL1	TMPO
APBB1	CXCL11	HIST1H1C	NAALADL1	RBL2	TMPRSS11D
APBB2	CXCL12	HIST1H1D	NAB1	RBM10	TMPRSS2
APBB3	CXCL13	HIST1H1E	NAB2	RBM14	TMPRSS6
APC	CXCL2	HIST1H1T	NACA	RBM15B	TMSB10
APC2	CXCL3	HIST1H2AK	NAGA	RBM17	TNC
APCS	CXCL5	HIST1H2BD	NAGLU	RBM25	TNF
APEX1	CXCL6	HIST1H2BK	NAGPA	RBM3	TNFAIP1
APEX2	CXCL9	HIST1H2BM	NAP1L1	RBM38	TNFAIP2
API5	CXCR3	HIST1H3I	NAP1L3	RBM39	TNFAIP3
APITD1	CXCR4	HIST1H4L	NAP1L4	RBM4B	TNFAIP6
APLP1	CXCR6	HIST2H2BE	NAPA	RBM5	TNFAIP8
APLP2	CXorf1	HIVEP1	NAPG	RBM6	TNFRSF10B
APOA1	CYB561	HIVEP2	NARS	RBM8A	TNFRSF10C
APOA4	CYB561D2	HK2	NASP	RBMS1	TNFRSF10D
APOB	CYB5A	HK3	NAT1	RBMS2	TNFRSF11A
APOBEC1	CYB5R1	HLA-A	NAT2	RBMS3	TNFRSF11B
APOBEC2	CYB5R3	HLA-DMA	NAT6	RBMX2	TNFRSF13B

APOBEC3B	CYBA	HLA-DMB	NAT8	RBP1	TNFRSF14
APOBEC3C	CYBB	HLA-DOA	NAT9	RBP4	TNFRSF17
APOBEC3F	CYC1	HLA-DOB	NAV3	RBPMS	TNFRSF1A
APOBEC3G	CYCS	HLA-DPA1	NBL1	RCBTB2	TNFRSF1B
APOC1	CYFIP1	HLA-DPB1	NBN	RCE1	TNFRSF21
APOC3	CYFIP2	HLA-DQA1	NCALD	RCHY1	TNFRSF25
APOD	CYHR1	HLA-DQB1	NCBP1	RCN2	TNFRSF4
APOE	CYLC2	HLA-DRA	NCBP2	RCOR1	TNFRSF8
APOF	CYLD	HLA-DRB1	NCDN	RDBP	TNFRSF9
APOH	CYP11A1	HLA-DRB4	NCF2	RDH11	TNFSF10
APOL1	CYP11B1	HLA-E	NCF4	RDH16	TNFSF11
APOM	CYP11B2	HLA-F	NCK1	RDH5	TNFSF12
APP	CYP17A1	HLA-G	NCKAP1	RECK	TNFSF14
APPBP2	CYP19A1	HLCS	NCKIPSD	RECQL	TNFSF4
APRT	CYP1A1	HLF	NCL	RECQL4	TNFSF8
AQP1	CYP1A2	HMBS	NCOA1	RECQL5	TNFSF9
AQP2	CYP1B1	HMG20B	NCOA2	REEP1	TNIP1
AQP3	CYP21A2	HMGA1	NCOA3	REEP2	TNK1
AQP4	CYP24A1	HMGA2	NCOA4	REEP5	TNK2
AQP6	CYP26A1	HMGB2	NCOA6	REG1A	TNKS
AQP7	CYP27A1	HMGB3	NCOR2	REG1B	TNNC1
AQP8	CYP27B1	HMGCL	NCR1	REG3A	TNNC2
AQP9	CYP2A13	HMGCR	NCR2	REL	TNNI1
AQR	CYP2A6	HMGCS1	NCR3	RELA	TNNI2
AR	CYP2A7	HMGCS2	NCSTN	RELB	TNNI3
ARAF	CYP2B6	HMGN3	NDEL1	RELN	TNNT1
ARC	CYP2C18	HMGN4	NDN	REM1	TNNT2
ARCN1	CYP2C19	HMHA1	NDP	REN	TNNT3
ARF1	CYP2C8	HMMR	NDRG1	RENBP	TNP1
ARF3	CYP2C9	HMOX1	NDRG2	REPS1	TNP2
ARF4	CYP2E1	HMOX2	NDRG4	REPS2	TNPO1
ARF5	CYP2J2	HMX1	NDST1	RER1	TNPO2
ARF6	CYP3A4	HNF4A	NDST2	RERE	TNPO3
ARFGAP3	CYP3A5	HNF4G	NDUFA1	REV3L	TNR

ARFGEF1	CYP3A7	HNMT	NDUFA5	RFC1	TNRC6B
ARFGEF2	CYP4A11	HNRPDL	NDUFA7	RFC2	TOB1
ARFIP1	CYP4B1	HOMER1	NDUFA9	RFC3	TOB2
ARFIP2	CYP4F11	HOMER2	NDUFAB1	RFC4	TOE1
ARG1	CYP4F12	HOMER3	NDUFAF1	RFC5	TOM1
ARG2	CYP4F2	HOXA1	NDUFB1	RFK	TOM1L1
ARHGAP1	CYP51A1	HOXA10	NDUFB3	RFNG	TOM1L2
ARHGAP11A	CYP7A1	HOXA11	NDUFB5	RFX1	TOMM20
ARHGAP12	CYP7B1	HOXA4	NDUFB6	RFX2	TOMM34
ARHGAP19	CYR61	HOXA5	NDUFB7	RFX3	TOMM40
ARHGAP22	D4S234E	HOXB1	NDUFB8	RFX5	TOMM70A
ARHGAP25	DAAM1	HOXB13	NDUFC1	RFXANK	TOP1
ARHGAP26	DAB2	HOXB2	NDUFS1	RFXAP	TOP2A
ARHGAP29	DACH1	HOXB3	NDUFS2	RGL1	TOP2B
ARHGAP4	DAD1	HOXB5	NDUFS3	RGL2	TOP3A
ARHGAP5	DAG1	HOXB6	NDUFS4	RGN	TOP3B
ARHGAP6	DAO	HOXB7	NDUFS6	RGR	TOPBP1
ARHGDIA	DAP	HOXC11	NDUFS7	RGS1	TOPORS
ARHGDIB	DAP3	HOXC4	NDUFS8	RGS10	TOR1A
ARHGEF1	DAPK1	HOXD1	NDUFV1	RGS11	TOR1AIP1
ARHGEF10	DAPK2	HOXD10	NDUFV2	RGS12	TOR1B
ARHGEF11	DAPK3	HOXD13	NEB	RGS13	TOX
ARHGEF12	DARC	HOXD3	NEBL	RGS14	TP53
ARHGEF15	DARS	HOXD9	NECAP1	RGS16	TP53BP1
ARHGEF16	DAXX	HPCA	NEDD4	RGS19	TP53BP2
ARHGEF17	DAZAP2	HPCAL1	NEDD4L	RGS2	TP53I11
ARHGEF18	DAZL	HPD	NEDD8	RGS20	TP53I3
ARHGEF2	DBC1	HPN	NEDD9	RGS3	TPD52
ARHGEF4	DBF4	HPRT1	NEFH	RGS4	TPD52L1
ARHGEF5	DBH	HPS1	NEFL	RGS5	TPD52L2
ARHGEF6	DBI	HPS5	NEK1	RGS6	TPI1
ARHGEF7	DBN1	HPX	NEK2	RGS7	TPM1
ARHGEF9	DBP	HR	NEK3	RGS9	TPM2
ARID1A	DBT	HRAS	NEK4	RHAG	TPM3

ARID3A	DCBLD2	HRC	NEK9	RHBDD3	TPMT
ARID4A	DCC	HRG	NELL1	RHBDF1	TPO
ARID5A	DCHS1	HRH1	NELL2	RHBDL1	TPP1
ARID5B	DCK	HRK	NENF	RHCE	TPP2
ARIH1	DCLRE1A	HRSP12	NEO1	RHEB	TPPP
ARIH2	DCP2	HS2ST1	NET1	RHO	TPR
ARL1	DCT	HS3ST1	NEU1	RHOA	TPST1
ARL2BP	DCTD	HSBP1	NEU3	RHOB	TPST2
ARL3	DCTN1	HSD11B1	NEURL	RHOBTB1	TPT1
ARL4C	DCTN2	HSD11B2	NEUROD1	RHOBTB2	TPX2
ARL4D	DCTN3	HSD17B1	NEUROD2	RHOBTB3	TRA2A
ARL6IP5	DCTN5	HSD17B2	NEUROG3	RHOC	TRADD
ARMC8	DCTN6	HSD17B3	NF1	RHOG	TRAF1
ARMCX2	DCX	HSD17B4	NF2	RHOH	TRAF2
ARNT	DDAH1	HSD17B6	NFASC	RHOQ	TRAF3
ARNT2	DDAH2	HSD17B8	NFAT5	RIBC2	TRAF3IP1
ARNTL	DDB1	HSD3B1	NFATC1	RIF1	TRAF3IP2
ARPC1A	DDB2	HSD3B2	NFATC2IP	RIMS2	TRAF3IP3
ARPC2	DDC	HSDL2	NFATC3	RIMS3	TRAF4
ARPC3	DDIT3	HSF1	NFATC4	RIN1	TRAF5
ARPC5	DDIT4	HSF2	NFE2	RIN2	TRAF6
ARR3	DDN	HSF2BP	NFE2L1	RING1	TRAFD1
ARRB2	DDO	HSF4	NFE2L2	RIOK3	TRAIP
ARSA	DDOST	HSP90AA1	NFE2L3	RIPK1	TRAK1
ARSB	DDR1	HSP90AB1	NFIB	RIPK2	TRAK2
ARSD	DDR2	HSP90B1	NFIC	RIT1	TRAM1
ARSE	DDT	HSPA1L	NFIL3	RIT2	TRAM2
ARSF	DDX1	HSPA2	NFIX	RLBP1	TRAP1
ART1	DDX10	HSPA4L	NFKB1	RLF	TRAPPC3
ART3	DDX11	HSPA5	NFKB2	RLN1	TRAPPC6A
ART4	DDX17	HSPA8	NFKBIB	RLN2	TRAT1
ARTN	DDX18	HSPB1	NFKBIE	RNASE1	TREH
ARVCF	DDX23	HSPB2	NFKBIL1	RNASE3	TRH
ASAH1	DDX27	HSPB3	NFRKB	RNASE4	TRHR

ASB1	DDX28	HSPB6	NFX1	RNASE6	TRIAP1
ASB4	DDX3X	HSPBP1	NFYA	RNASEH1	TRIB1
ASB9	DDX3Y	HSPE1	NFYB	RNASEH2A	TRIB2
ASCC2	DDX42	HSPG2	NFYC	RND1	TRIM10
ASCC3	DDX46	HTATIP2	NGFR	RND2	TRIM14
ASCL1	DDX49	HTATSF1	NHLH1	RND3	TRIM15
ASCL2	DDX5	HTN3	NHLH2	RNF10	TRIM2
ASF1A	DDX51	HTR1B	NHP2L1	RNF11	TRIM21
ASGR1	DDX52	HTR1D	NID1	RNF113A	TRIM22
ASGR2	DDX6	HTR1E	NID2	RNF126	TRIM23
ASH2L	DEAF1	HTR2A	NINJ1	RNF13	TRIM24
ASIP	DECR1	HTR2B	NIPA2	RNF139	TRIM25
ASL	DEDD	HTR2C	NIPBL	RNF14	TRIM26
ASMT	DEFA4	HTR3A	NIPSNAP1	RNF167	TRIM28
ASMTL	DEFA5	HTR4	NISCH	RNF185	TRIM29
ASNA1	DEFA6	HTR6	NIT1	RNF2	TRIM3
ASNS	DEFB1	HTR7	NKG7	RNF24	TRIM31
ASPA	DEGS1	HTRA1	NKRF	RNF4	TRIM32
ASPH	DEK	HTRA2	NKTR	RNF40	TRIM33
ASPHD1	DEPDC5	HUS1	NKX2-2	RNF41	TRIM37
ASTN2	DES	HUWE1	NKX2-5	RNF44	TRIM38
ASXL1	DEXI	HYAL1	NKX2-8	RNF5	TRIM44
ATF1	DFFA	HYAL2	NKX3-1	RNF6	TRIM52
ATF2	DFFB	HYOU1	NLE1	RNF8	TRIM58
ATF3	DFNA5	IAPP	NLGN1	RNGTT	TRIM9
ATF5	DGAT1	IARS	NLGN4Y	RNH1	TRIO
ATF6	DGCR14	IARS2	NMB	RNMT	TRIOBP
ATF7	DGCR2	IBSP	NMBR	RNPEP	TRIP10
ATG12	DGKA	IBTK	NME3	RNPS1	TRIP11
ATG4A	DGKB	ICA1	NME4	ROBO1	TRIP12
ATG4B	DGKD	ICAM1	NME5	ROCK1	TRIP13
ATG5	DGKE	ICAM2	NME6	ROCK2	TRIP4
ATG9A	DGKG	ICAM3	NMI	ROD1	TRIP6
ATIC	DGKQ	ICAM4	NMT1	ROM1	TRMT1

ATM	DGKZ	ICAM5	NMT2	ROR1	TRMU
ATN1	DGUOK	ICK	NMU	ROR2	TRO
ATOX1	DHCR24	ICMT	NNAT	RORA	TROAP
ATP10A	DHCR7	ICOS	NNMT	RORB	TROVE2
ATP10D	DHODH	ICOSLG	NNT	RORC	TRPA1
ATP11A	DHPS	ICT1	NOC2L	ROS1	TRPC1
ATP11B	DHRS1	ID1	NOL3	RP2	TRPC3
ATP12A	DHRS2	ID2	NOL7	RPA1	TRPC4AP
ATP1A1	DHRS3	ID3	NOLC1	RPA2	TRPC6
ATP1A2	DHRS7	ID4	NONO	RPA3	TRPM1
ATP1A3	DHX15	IDE	NOS1	RPE	TRPM2
ATP1B1	DHX16	IDH1	NOS1AP	RPE65	TRPV6
ATP1B2	DHX29	IDH2	NOS3	RPGR	TRRAP
ATP1B3	DHX30	IDH3A	NOTCH2	RPGRIP1	TSC1
ATP2A1	DHX34	IDH3B	NOTCH3	RPIA	TSC2
ATP2A2	DHX38	IDH3G	NOTCH4	RPL10	TSC22D1
ATP2A3	DHX57	IDS	NOV	RPL11	TSC22D2
ATP2B1	DHX8	IDUA	NOVA1	RPL12	TSC22D3
ATP2B2	DHX9	IER2	NOVA2	RPL13	TSC22D4
ATP2B3	DIAPH1	IER3	NOX1	RPL13A	TSFM
ATP2B4	DIAPH2	IFI16	NPAS1	RPL14	TSG101
ATP2C1	DICER1	IFI27	NPAS2	RPL17	TSHB
ATP4A	DIDO1	IFI30	NPAS3	RPL19	TSHR
ATP4B	DIO1	IFI35	NPAT	RPL22	TSN
ATP5A1	DIO2	IFI44	NPC1	RPL23	TSNAX
ATP5B	DIO3	IFI44L	NPC2	RPL27	TSPAN1
ATP5C1	DIP2C	IFI6	NPEPPS	RPL28	TSPAN2
ATP5D	DIRAS3	IFIH1	NPFF	RPL3	TSPAN3
ATP5F1	DISC1	IFIT1	NPHP1	RPL30	TSPAN31
ATP5G1	DKC1	IFIT2	NPHP4	RPL31	TSPAN4
ATP5G2	DKK1	IFIT3	NPHS1	RPL34	TSPAN5
ATP5G3	DKK3	IFIT5	NPM3	RPL36AL	TSPAN6
ATP5H	DKK4	IFITM1	NPPA	RPL37	TSPAN7
ATP5I	DLAT	IFITM2	NPPB	RPL37A	TSPAN8

ATP5J	DLC1	IFITM3	NPR1	RPL38	TSPAN9
ATP5J2	DLD	IFNA10	NPR2	RPL39L	TSPO
ATP5L	DLEC1	IFNA14	NPR3	RPL3L	TSPYL2
ATP5O	DLG1	IFNA16	NPTN	RPL4	TSPYL4
ATP5S	DLG2	IFNA2	NPTX1	RPL5	TSPYL5
ATP6AP1	DLG3	IFNA21	NPTX2	RPL8	TSR1
ATP6AP2	DLG4	IFNA5	NPTXR	RPLP2	TSSK2
ATP6V0A1	DLG5	IFNA6	NPY	RPN1	TST
ATP6V0A2	DLGAP1	IFNA8	NPY2R	RPN2	TSTA3
ATP6V0B	DLGAP2	IFNAR1	NPY5R	RPP14	TTBK2
ATP6V0C	DLGAP4	IFNAR2	NQO1	RPP30	TTC1
ATP6V0D1	DLK1	IFNB1	NQO2	RPP38	TTC15
ATP6V1A	DLX2	IFNG	NR0B1	RPP40	TTC3
ATP6V1B1	DLX4	IFNGR1	NR0B2	RPS11	TTF1
ATP6V1B2	DLX5	IFNGR2	NR1D1	RPS12	TTF2
ATP6V1C1	DMBT1	IFNW1	NR1D2	RPS13	TTK
ATP6V1D	DMC1	IFRD1	NR1H2	RPS14	TTLL1
ATP6V1E1	DMD	IFRD2	NR1H3	RPS16	TTLL12
ATP6V1F	DMP1	IFT140	NR1H4	RPS19	TTLL4
ATP6V1G1	DMPK	IFT88	NR1I2	RPS20	TTLL5
ATP6V1G2	DMTF1	IGBP1	NR1I3	RPS21	TTN
ATP6V1H	DMWD	IGF1	NR2C1	RPS23	TTPA
ATP7B	DMXL1	IGF1R	NR2C2	RPS24	TTR
ATP8A1	DMXL2	IGF2BP3	NR2E1	RPS25	TUB
ATP8A2	DNAH17	IGF2R	NR2E3	RPS29	TUBB2C
ATP8B1	DNAH7	IGFALS	NR2F1	RPS3	TUBB3
ATP8B3	DNAH9	IGFBP1	NR2F2	RPS4X	TUBB4
ATP9A	DNAJA1	IGFBP2	NR2F6	RPS5	TUBB4Q
ATP9B	DNAJA2	IGFBP3	NR3C1	RPS6	TUBG1
ATPAF2	DNAJA3	IGFBP4	NR3C2	RPS6KA1	TUBGCP2
ATR	DNAJB1	IGFBP5	NR4A1	RPS6KA2	TUBGCP3
ATRN	DNAJB12	IGFBP6	NR4A2	RPS6KA3	TUFM
ATRX	DNAJB2	IGFBP7	NR4A3	RPS6KA4	TULP1
ATXN1	DNAJB4	IGHMBP2	NR5A1	RPS6KA5	TULP2

ATXN10	DNAJB5	IGJ	NR5A2	RPS6KB1	TULP3
ATXN2	DNAJB6	IGLL1	NRAS	RPS6KB2	TUSC2
ATXN2L	DNAJB9	IGSF1	NRCAM	RPS9	TUSC3
ATXN3	DNAJC11	IGSF3	NRD1	RPUSD2	TWIST1
ATXN7	DNAJC13	IGSF6	NRF1	RQCD1	TWISTNB
AUH	DNAJC3	IHH	NRG1	RRAD	TXK
AURKA	DNAJC4	IK	NRG2	RRAGA	TXLNA
AURKB	DNAJC6	IKBKAP	NRGN	RRAGB	TXN
AURKC	DNAJC7	IKBKB	NRIP1	RRAGD	TXN2
AUTS2	DNAJC8	IKBKE	NRL	RRAS	TXNDC9
AVIL	DNAL4	IL10	NRP1	RRAS2	TXNIP
AVP	DNALI1	IL10RA	NRP2	RRBP1	TXNL1
AVPR1A	DNASE1	IL10RB	NRTN	RREB1	TXNL4A
AVPR1B	DNASE1L1	IL11	NRXN1	RRH	TXNRD1
AVPR2	DNASE1L2	IL11RA	NRXN2	RRM1	TXNRD2
AXIN1	DNASE1L3	IL12A	NRXN3	RRM2	TYK2
AXL	DNASE2	IL12B	NSDHL	RRN3	TYMS
AZGP1	DNM1	IL12RB1	NSF	RRS1	TYR
AZIN1	DNM1L	IL12RB2	NSFL1C	RS1	TYRO3
AZU1	DNM2	IL13	NSMAF	RSAD2	TYROBP
B2M	DNM3	IL13RA1	NT5C2	RSBN1	TYRP1
B3GALNT1	DNMT1	IL13RA2	NT5E	RSL1D1	U2AF1
B3GALT2	DNPEP	IL15	NTF3	RSU1	U2AF2
B3GALT4	DNTT	IL15RA	NTHL1	RTCD1	UAP1
B3GALT5	DNTTIP2	IL16	NTNG1	RTF1	UBAP2L
B3GAT3	DOC2A	IL18	NTRK1	RTN1	UBC
B3GNT1	DOC2B	IL18R1	NTRK2	RTN2	UBE2A
B3GNT3	DOCK1	IL18RAP	NTRK3	RTN4	UBE2B
B4GALT1	DOCK10	IL1A	NTS	RUFY3	UBE2C
B4GALT2	DOCK4	IL1B	NTSR1	RUNX1	UBE2D1
B4GALT3	DOCK9	IL1R1	NTSR2	RUNX1T1	UBE2D2
B4GALT4	DOK1	IL1R2	NUAK1	RUNX2	UBE2D3
B4GALT5	DOK2	IL1RAP	NUBP1	RUNX3	UBE2E1
B4GALT6	DOK4	IL1RL1	NUCB1	RUSC1	UBE2E3

BAAT	DOK5	IL1RL2	NUCB2	RUSC2	UBE2G1
BACH1	DOLPP1	IL1RN	NUDC	RUVBL1	UBE2G2
BAD	DOM3Z	IL2	NUDCD3	RUVBL2	UBE2H
BAG1	DOPEY2	IL23A	NUDT1	RXRA	UBE2I
BAG5	DOT1L	IL24	NUDT13	RXRB	UBE2J1
BAGE	DPAGT1	IL27RA	NUDT21	RXRG	UBE2L3
BAHD1	DPEP1	IL2RA	NUDT3	RYBP	UBE2L6
BAI1	DPF1	IL2RB	NUFIP1	RYK	UBE2M
BAI2	DPF2	IL2RG	NUMA1	RYR1	UBE2N
BAI3	DPH2	IL3	NUMB	RYR2	UBE2V2
BAIAP2	DPM1	IL32	NUP133	RYR3	UBE3A
BAIAP3	DPM2	IL3RA	NUP153	S100A1	UBE3B
BAK1	DPP4	IL4	NUP155	S100A10	UBE3C
BAMBI	DPP6	IL4R	NUP160	S100A11	UBE4A
BANF1	DPT	IL5	NUP210	S100A12	UBE4B
BAP1	DPYD	IL5RA	NUP214	S100A13	UBL3
BARD1	DPYS	IL6	NUP50	S100A2	UBL4A
BARX2	DPYSL2	IL6R	NUP62	S100A3	UBN1
BASP1	DPYSL3	IL6ST	NUP88	S100A4	UBOX5
BATF	DPYSL4	IL7	NUP93	S100A5	UBQLN2
BAX	DR1	IL7R	NUP98	S100A7	UBR2
BAZ1A	DRAP1	IL8	NUPL1	S100A8	UBTF
BAZ1B	DRD1	IL9	NUPL2	S100A9	UCHL1
BAZ2A	DRD2	ILF2	NUPR1	S100B	UCHL3
BAZ2B	DRD3	ILF3	NVL	S100G	UCK2
BBC3	DRD4	ILK	NXF1	S100P	UCKL1
BBOX1	DRD5	ILVBL	NXPH3	SAA4	UCN
BBS4	DRG1	IMMT	NXT2	SACM1L	UCP2
BBX	DRG2	IMP4	OAS1	SACS	UCP3
BCAM	DRP2	IMPA1	OAS2	SAFB	UFD1L
BCAP29	DSC1	IMPA2	OASL	SAFB2	UGCG
BCAP31	DSC2	IMPDH1	OAT	SAG	UGDH
BCAR3	DSC3	IMPDH2	OAZ1	SALL1	UGP2
BCAS1	DSCAM	IMPG1	OAZ3	SALL2	UGT2B4

BCAS2	DSCR3	INA	OCRL	SAMD14	UGT8
BCAT1	DSCR4	INADL	ODC1	SAMD4A	ULK1
BCAT2	DSG1	ING1	ODF1	SAMHD1	ULK2
BCHE	DSG2	ING2	ODF2	SAMM50	UMOD
BCKDHA	DSG3	ING3	OFD1	SAP18	UMPS
BCKDHB	DSP	INHA	OGDH	SAP30	UNC119
BCKDK	DST	INHBA	OGFR	SAR1A	UNC13B
BCL10	DSTN	INHBB	OGG1	SARDH	UNC50
BCL11A	DTNA	INHBC	OGT	SARS	UNC5B
BCL2	DTNB	INPP1	OIP5	SART1	UNC5C
BCL2A1	DUS4L	INPP4A	OLFM1	SART3	UNG
BCL2L1	DUSP1	INPP4B	OLFM4	SASH1	UPF2
BCL2L11	DUSP10	INPP5A	OLFML1	SATB1	UPF3A
BCL2L2	DUSP11	INPP5B	OLFML2A	SBF1	UPK1A
BCL3	DUSP14	INPP5D	OLFML2B	SC4MOL	UPK1B
BCL6	DUSP2	INPP5E	OLIG2	SC5DL	UPK2
BCL7A	DUSP3	INPP5F	OLR1	SCAMP1	UPK3A
BCL7B	DUSP4	INPPL1	OMD	SCAMP3	UPP1
BCL9	DUSP5	INS	OMG	SCAMP5	UQCRB
BCLAF1	DUSP6	INSIG1	ONECUT1	SCAP	UQCRC1
BCR	DUSP7	INSIG2	ONECUT2	SCARB1	UQCRC2
BCS1L	DUSP8	INSL3	OPA1	SCARB2	UQCRFS1
BDKRB1	DUSP9	INSL4	OPCML	SCARF1	UQCRH
BDKRB2	DUT	INSM1	OPLAH	SCCPDH	UROD
BDNF	DUX1	INSR	OPN1SW	SCD	UROS
BECN1	DVL3	INTS1	OPRD1	SCEL	USF2
BET1	DYNC1H1	INVS	OPRK1	SCFD1	USH1C
BFSP1	DYNC1I1	IPO13	OPRL1	SCG2	USH2A
BFSP2	DYNC1LI2	IPO8	OPRM1	SCG5	USP1
BGLAP	DYNC2LI1	IPO9	OPTN	SCGB1A1	USP10
BGN	DYNLT1	IQCB1	OR10H3	SCGB1D1	USP11
BHMT	DYRK1A	IQCE	OR2F1	SCGB1D2	USP13
BICD1	DYRK1B	IQCK	OR2F2	SCGB2A1	USP14
BICD2	DYRK2	IQGAP1	OR2H1	SCGB2A2	USP15

BID	DYRK3	IQGAP2	OR2J2	SCGN	USP2
BIK	DYRK4	IQSEC1	OR5I1	SCML2	USP20
BIN1	DZIP1	IRAK1	OR7A5	SCN1B	USP33
BIRC2	DZIP3	IRAK3	OR7E24	SCN2B	USP34
BIRC3	E2F1	IREB2	ORM1	SCN4A	USP4
BIRC5	E2F2	IRF1	OS9	SCN5A	USP46
BLCAP	E2F3	IRF2	OSBP	SCN7A	USP5
BLK	E2F4	IRF2BP1	OSBPL1A	SCN9A	USP6
BLM	E2F5	IRF3	OSBPL2	SCNN1A	USP6NL
BLMH	EBAG9	IRF4	OSBPL3	SCNN1B	USP7
BLNK	EBNA1BP2	IRF5	OSBPL8	SCNN1D	USP8
BLOC1S1	EBP	IRF6	OSGEP	SCO2	USP9X
BLVRA	ECD	IRF7	OSM	SCP2	USPL1
BLVRB	ECE1	IRF8	OSMR	SCRG1	UST
BMP1	ECE2	IRGC	OSR2	SCRIB	UTF1
BMP10	ECH1	IRS1	OSTF1	SCRN1	UTP14C
BMP2	ECHS1	IRS2	OTC	SCTR	UTP20
BMP2K	ECM1	IRS4	OTUB1	SCYL3	UTRN
BMP3	ECM2	IRX5	OTUD4	SDC1	UVRAG
BMP4	EDA	ISG15	OVGP1	SDC2	VAC14
BMP5	EDEM1	ISG20	OVOL1	SDC3	VAMP1
BMP6	EDIL3	ISG20L2	OVOL2	SDC4	VAMP2
BMP7	EDN2	ISLR	OXA1L	SDCBP	VAMP3
BMP8A	EDN3	ITCH	OXCT1	SDF2	VAMP4
BMPR1B	EDNRA	ITGA10	OXSR1	SDHB	VAMP5
BMPR2	EDNRB	ITGA2	OXT	SDHC	VAMP8
BMX	EEA1	ITGA2B	Oca.02	SDHD	VAPA
BNC1	EED	ITGA3	P2RX1	SDS	VAPB
BNIP1	EEF1A1	ITGA4	P2RX3	SEC14L1	VARS
BNIP2	EEF1A2	ITGA5	P2RX4	SEC14L2	VASH1
BNIP3	EEF1D	ITGA6	P2RX5	SEC23A	VASP
BNIP3L	EEF1E1	ITGA7	P2RX7	SEC23B	VAT1
BOP1	EEF2	ITGA8	P2RY1	SEC23IP	VAV1
BPGM	EFEMP1	ITGAE	P2RY10	SEC24B	VBP1

BPHL	EFEMP2	ITGAL	P2RY11	SEC24C	VCAM1
BPI	EFHA1	ITGAM	P2RY14	SEC24D	VCL
BRAP	EFHD1	ITGAV	P2RY2	SEC61A1	VCP
BRCA1	EFNA1	ITGAX	P2RY6	SEC61B	VDAC1
BRCA2	EFNA2	ITGB1	P4HA1	SEC61G	VDAC2
BRD1	EFNA3	ITGB1BP1	P4HA2	SEC63	VDAC3
BRD2	EFNA4	ITGB2	PA2G4	SECTM1	VDR
BRD3	EFNA5	ITGB3	PABPC1	SEL1L	VEGFB
BRD4	EFNB1	ITGB3BP	PABPN1	SELE	VEGFC
BRD8	EFNB2	ITGB5	PACRG	SELENBP1	VENTX
BRDT	EFNB3	ITGB6	PACSIN2	SELL	VGF
BRE	EFS	ITGB7	PAEP	SELP	VGLL1
BRF1	EGF	ITGB8	PAF1	SELPLG	VGLL4
BRMS1	EGFR	ITGBL1	PAFAH1B1	SEMA3A	VILL
BRP44	EGR1	ITIH1	PAFAH1B2	SEMA3B	VIM
BRPF1	EGR2	ITIH2	PAFAH1B3	SEMA3C	VIP
BRS3	EGR3	ITIH3	PAFAH2	SEMA3D	VIPR1
BRSK2	EGR4	ITIH4	PAGE1	SEMA3E	VIPR2
BRWD1	EHBP1	ITK	PAGE4	SEMA3F	VLDLR
BSCL2	EHD1	ITM2A	PAH	SEMA4D	VNN1
BSG	EHHADH	ITM2B	PAICS	SEMA4F	VNN2
BSN	EHMT2	ITPA	PAIP1	SEMA5A	VPRBP
BST1	EI24	ITPK1	PAK1	SEMA6C	VPS11
BST2	EIF1	ITPKA	PAK2	SEMA7A	VPS13A
BTAF1	EIF1AX	ITPKB	PAK3	SEMG1	VPS13B
BTBD3	EIF1AY	ITPKC	PAK4	SEMG2	VPS13D
BTC	EIF1B	ITPR1	PAK7	SENP3	VPS26A
BTD	EIF2AK1	ITPR2	PALLD	SENP5	VPS39
BTF3	EIF2AK2	ITPR3	PALM	SENP6	VPS41
BTG1	EIF2B1	ITSN1	PAM	SEPHS1	VPS4B
BTG2	EIF2B2	ITSN2	PANX1	SEPHS2	VPS52
BTG3	EIF2B4	IVL	PAPOLA	SEPP1	VPS72
BTK	EIF2B5	IVNS1ABP	PAPPA	SEPW1	VRK1
BTN1A1	EIF2C2	JAG1	PAPPA2	SERINC1	VRK2

BTN2A1	EIF2S1	JAG2	PAPSS1	SERINC3	VSIG4
BTN3A1	EIF2S2	JAK1	PAPSS2	SERP1	VSNL1
BTN3A2	EIF2S3	JAK2	PAQR3	SERPINA1	VTI1B
BTN3A3	EIF4A1	JAK3	PAQR4	SERPINA3	VTN
BTNL3	EIF4A2	JAKMIP2	PARD3	SERPINA4	VWF
BTRC	EIF4E	JARID2	PARD6A	SERPINA5	WARS
BUB1	EIF4E2	JOSD1	PARD6B	SERPINA6	WAS
BUB1B	EIF4EBP1	JRK	PARG	SERPINA7	WASF1
BUB3	EIF4EBP2	JRKL	PARK2	SERPINB1	WASF3
BYSL	EIF4G1	JTB	PARK7	SERPINB13	WASL
BZRAP1	EIF4G2	JUN	PARN	SERPINB3	WBP2
C10orf10	EIF4G3	JUNB	PARP1	SERPINB5	WBP4
C10orf116	EIF5	JUND	PARP2	SERPINB6	WBSCR22
C10orf12	EIF5A	JUP	PARP3	SERPINB7	WDFY3
C10orf137	EIF5B	KAL1	PARP4	SERPINB8	WDHD1
C10orf26	ELAC2	KALRN	PARVB	SERPINB9	WDR1
C10orf28	ELAVL1	KARS	PASK	SERPINC1	WDR13
C10orf72	ELAVL2	KATNA1	PAX1	SERPIND1	WDR18
C11orf51	ELAVL3	KATNB1	PAX2	SERPINE2	WDR37
C11orf9	ELAVL4	KAZALD1	PAX3	SERPINF1	WDR45
C12orf24	ELF1	KBTBD11	PAX4	SERPINF2	WDR45L
C14orf1	ELF2	KBTBD2	PAX5	SERPING1	WDR46
C14orf147	ELF3	KCNA1	PAX6	SERPINH1	WDR47
C14orf2	ELF4	KCNA3	PAX7	SERPINI1	WDR61
C14orf79	ELF5	KCNA4	PAX8	SERPINI2	WDR62
C15orf39	ELK1	KCNA5	PAX9	SERTAD2	WDR67
C16orf3	ELK3	KCNA6	PAXIP1	SET	WDR7
C16orf45	ELK4	KCNAB1	PBX1	SETBP1	WDR73
C16orf7	ELL	KCNAB2	PBX2	SETD1A	WDR77
C17orf75	ELMO1	KCNB1	PBX3	SETD3	WDTC1
C18orf1	ELN	KCNB2	PBXIP1	SETD4	WEE1
C18orf10	ELOVL2	KCNC1	PC	SETDB1	WFDC2
C18orf25	ELOVL5	KCNC3	PCBD1	SETX	WFS1
C19orf10	ELOVL6	KCNC4	PCBP1	SEZ6L	WHSC1

C19orf2	ELP4	KCND1	PCBP2	SF1	WHSC2
C19orf21	EMD	KCND2	PCBP3	SF3A1	WIF1
C19orf26	EMG1	KCND3	PCBP4	SF3A2	WIPI1
C19orf6	EMID1	KCNF1	PCCA	SF3A3	WIPI2
C1D	EMILIN1	KCNG1	PCCB	SF3B1	WISP1
C1QB	EML1	KCNH1	PCDH1	SF3B2	WISP2
C1QBP	EML3	KCNH2	PCDH11X	SF3B3	WISP3
C1QL1	EMP1	KCNJ1	PCDH17	SF3B4	WNK1
C1R	EMP2	KCNJ10	PCDH7	SFI1	WNT1
C1S	EMP3	KCNJ12	PCDH8	SFMBT1	WNT10B
C1orf105	EMR1	KCNJ13	PCDHB11	SFN	WNT11
C1orf114	EMR2	KCNJ15	PCDHGA8	SFPQ	WNT2
C1orf123	EMX1	KCNJ2	PCF11	SFRP1	WNT2B
C1orf144	EMX2	KCNJ3	PCGF1	SFT2D2	WNT4
C1orf21	EN2	KCNJ4	PCGF2	SFTPB	WNT5A
C1orf38	ENC1	KCNJ5	PCGF3	SFTPC	WNT6
C1orf61	ENDOG	KCNJ6	PCK1	SFTPD	WNT7A
C1orf63	ENG	KCNJ8	PCK2	SFXN3	WNT8B
C1orf9	ENO1	KCNJ9	PCM1	SGCA	WRB
C2	ENO2	KCNK1	PCMT1	SGCB	WRN
C20orf111	ENO3	KCNK2	PCMTD2	SGCD	WSB1
C21orf2	ENOSF1	KCNK3	PCNA	SGCE	WSB2
C21orf33	ENPEP	KCNMA1	PCNX	SGCG	WT1
C2orf3	ENPP1	KCNMB1	PCNXL2	SGPL1	WTAP
C3AR1	ENPP2	KCNN1	PCOLCE	SGSH	WWC1
C3orf27	ENPP4	KCNN3	PCP4	SGTA	WWOX
C3orf32	ENSA	KCNN4	PCSK1	SH2B1	WWP1
C3orf37	ENTPD2	KCNQ1	PCSK2	SH2D1A	WWP2
C3orf63	ENTPD3	KCNQ3	PCSK6	SH2D2A	WWTR1
C4BPA	ENTPD4	KCNS1	PCSK7	SH3BGR	XBP1
C4BPB	ENTPD5	KCNS3	PCYT1A	SH3BGRL	XCL1
C4orf6	ENTPD6	KCTD12	PCYT1B	SH3BP1	XDH
C5	EP300	KCTD17	PDAP1	SH3BP2	XK
C5orf13	EP400	KCTD7	PDC	SH3BP5	XPA

C5orf15	EPAS1	KDELR1	PDCD1	SH3GL1	XPC
C6	EPB41	KDELR2	PDCD10	SH3GL2	XPNPEP1
C6orf10	EPB41L1	KDELR3	PDCD2	SH3GL3	XPNPEP2
C6orf105	EPB41L2	KDR	PDCD4	SH3GLB1	XPO1
C6orf106	EPB41L3	KEAP1	PDCD6	SH3PXD2A	XPO7
C6orf108	EPB42	KEL	PDCL	SH3YL1	XRCC1
C6orf130	EPB49	KHDRBS1	PDE10A	SHANK2	XRCC2
C6orf162	EPHA1	KHDRBS2	PDE1A	SHB	XRCC3
C6orf47	EPHA2	KHDRBS3	PDE1B	SHBG	XRCC4
C6orf62	EPHA3	KHK	PDE1C	SHC1	XRCC5
C7	EPHA4	KHSRP	PDE2A	SHC3	XYLT1
C8A	EPHA5	KIAA0020	PDE3A	SHFM1	YAF2
C8B	EPHA7	KIAA0090	PDE3B	SHH	YAP1
C8G	EPHB1	KIAA0100	PDE4A	SHMT1	YARS
C9orf114	EPHB2	KIAA0101	PDE4B	SHMT2	YES1
C9orf125	EPHB3	KIAA0141	PDE4C	SHOX2	YIF1A
C9orf16	EPHB4	KIAA0174	PDE4D	SI	YIPF1
C9orf3	EPHB6	KIAA0182	PDE4DIP	SIAH1	YIPF2
C9orf91	EPHX1	KIAA0195	PDE5A	SIAH2	YIPF3
CA1	EPHX2	KIAA0196	PDE6A	SIGLEC6	YIPF4
CA11	EPM2A	KIAA0232	PDE6B	SIGLEC7	YIPF6
CA12	EPM2AIP1	KIAA0247	PDE6C	SIM1	YKT6
CA2	EPN2	KIAA0284	PDE6D	SIM2	YME1L1
CA3	EPO	KIAA0319	PDE6G	SIN3B	YPEL1
CA4	EPOR	KIAA0355	PDE6H	SIP1	YTHDC1
CA5A	EPRS	KIAA0391	PDE8A	SIPA1	YTHDC2
CA5B	EPS15	KIAA0467	PDE8B	SIPA1L1	YTHDF3
CA6	ERAL1	KIAA0494	PDE9A	SIPA1L3	YWHAB
CA7	ERBB2	KIAA0513	PDGFA	SIRPB1	YWHAE
CA9	ERBB3	KIAA0528	PDGFB	SIT1	YWHAH
CABIN1	ERBB4	KIAA0586	PDGFRA	SIX1	YWHAQ
CABP1	ERCC1	KIAA0649	PDGFRB	SIX3	YWHAZ
CACNA1A	ERCC2	KIAA0664	PDGFRL	SIX6	YY1
CACNA1B	ERCC3	KIAA0753	PDHA1	SKIL	ZAP70

CACNA1C	ERCC4	KIAA0776	PDHA2	SKIV2L	ZBED1
CACNA1D	ERCC5	KIAA0907	PDHB	SKIV2L2	ZBED4
CACNA1E	ERCC6	KIAA0922	PDHX	SKP2	ZBTB1
CACNA1F	EREG	KIAA1279	PDIA2	SLA	ZBTB11
CACNA1G	ERF	KIAA1539	PDIA3	SLAMF1	ZBTB16
CACNA1H	ERG	KIDINS220	PDIA4	SLBP	ZBTB17
CACNA1S	ERGIC3	KIF11	PDIA5	SLC10A1	ZBTB20
CACNA2D1	ERH	KIF13B	PDIA6	SLC10A2	ZBTB24
CACNA2D2	ERN1	KIF14	PDK1	SLC10A3	ZBTB25
CACNB1	ESD	KIF1A	PDK2	SLC11A1	ZBTB33
CACNB2	ESPL1	KIF1B	PDK3	SLC11A2	ZBTB39
CACNB3	ESR1	KIF1C	PDK4	SLC12A1	ZBTB40
CACNB4	ESR2	KIF21B	PDLIM1	SLC12A2	ZBTB5
CACNG1	ESRRA	KIF23	PDLIM3	SLC12A3	ZBTB7A
CACNG3	ESRRB	KIF25	PDLIM4	SLC12A5	ZBTB7B
CACYBP	ESRRG	KIF2C	PDLIM5	SLC13A2	ZC3H11A
CAD	ETF1	KIF3A	PDLIM7	SLC13A3	ZC3H14
CADPS	ETFA	KIF3B	PDPK1	SLC14A2	ZC3H3
CALB1	ETFB	KIF3C	PDPN	SLC15A1	ZC3H7B
CALB2	ETFDH	KIF5A	PDXK	SLC15A2	ZC3HAV1
CALCA	ETHE1	KIF5B	PDYN	SLC16A1	ZCCHC11
CALCB	ETS1	KIF5C	PDZK1	SLC16A2	ZCCHC14
CALCOCO1	ETS2	KIFAP3	PDZK1IP1	SLC16A3	ZDHHC17
CALCOCO2	ETV1	KIFC1	PDZRN3	SLC16A4	ZDHHC18
CALCR	ETV2	KIFC3	PEA15	SLC16A6	ZDHHC3
CALD1	ETV3	KIN	PEBP1	SLC16A7	ZFP161
CALML3	ETV4	KIR2DL1	PECAM1	SLC17A1	ZFP36
CALR	ETV5	KIR2DL4	PEG10	SLC17A2	ZFP36L1
CALU	ETV6	KIR3DL1	PELP1	SLC17A3	ZFP37
CAMK1	EVI2B	KISS1	PEMT	SLC17A4	ZFPL1
CAMK1G	EVPL	KIT	PENK	SLC17A7	ZFR
CAMK2A	EVX1	KITLG	PEPD	SLC18A1	ZFX
CAMK2B	EWSR1	KL	PER1	SLC18A2	ZFYVE16
CAMK2G	EXO1	KLF1	PER2	SLC19A1	ZFYVE9

CAMK4	EXOSC10	KLF10	PES1	SLC19A2	ZHX2
CAMKK2	EXOSC2	KLF11	PET112L	SLC1A1	ZHX3
CAMP	EXOSC7	KLF12	PEX1	SLC1A2	ZIC1
CAMSAP1	EXOSC8	KLF4	PEX10	SLC1A3	ZIC3
CAMTA1	EXOSC9	KLF5	PEX11A	SLC1A4	ZKSCAN1
CAMTA2	EXPH5	KLF6	PEX11B	SLC1A5	ZMPSTE24
CAND1	EXT1	KLF9	PEX12	SLC1A6	ZMYM3
CANX	EXT2	KLHDC3	PEX13	SLC1A7	ZMYM4
CAP1	EXTL1	KLHL18	PEX14	SLC20A1	ZMYND10
CAP2	EXTL2	KLHL20	PEX19	SLC20A2	ZMYND11
CAPG	EXTL3	KLHL21	PEX3	SLC22A1	ZNF10
CAPN1	EYA1	KLHL23	PEX5	SLC22A13	ZNF117
CAPN11	EYA2	KLHL25	PEX6	SLC22A14	ZNF124
CAPN2	EYA3	KLHL4	PEX7	SLC22A18	ZNF132
CAPN3	EYA4	KLHL9	PF4	SLC22A2	ZNF133
CAPN5	EZH1	KLK1	PF4V1	SLC22A4	ZNF134
CAPN7	EZH2	KLK10	PFDN1	SLC22A5	ZNF135
CAPN9	F10	KLK11	PFDN4	SLC22A6	ZNF136
CAPNS1	F11	KLK13	PFDN5	SLC23A2	ZNF140
CAPZA2	F12	KLK2	PFDN6	SLC24A1	ZNF141
CAPZB	F13A1	KLK3	PFKFB1	SLC25A1	ZNF146
CARD10	F13B	KLK6	PFKFB2	SLC25A11	ZNF148
CARD8	F2R	KLK7	PFKFB3	SLC25A12	ZNF160
CARS	F2RL1	KLK8	PFKFB4	SLC25A13	ZNF165
CASC3	F2RL2	KLKB1	PFKL	SLC25A14	ZNF167
CASK	F3	KLRB1	PFKM	SLC25A16	ZNF174
CASP1	F5	KLRC3	PFKP	SLC25A17	ZNF175
CASP10	F7	KLRC4	PFN1	SLC25A20	ZNF177
CASP2	F8	KLRD1	PFN2	SLC25A24	ZNF184
CASP3	F9	KLRG1	PGAM2	SLC25A3	ZNF187
CASP4	FAAH	KMO	PGAP1	SLC25A36	ZNF189
CASP5	FABP1	KNG1	PGC	SLC25A4	ZNF192
CASP6	FABP2	KNTC1	PGCP	SLC25A5	ZNF193
CASP7	FABP3	KPNA3	PGD	SLC26A10	ZNF195

CASP9	FABP6	KPNA5	PGF	SLC26A2	ZNF197
CASQ2	FABP7	KPNA6	PGGT1B	SLC26A3	ZNF20
CASR	FADD	KRAS	PGK1	SLC26A4	ZNF200
CAST	FADS1	KRIT1	PGLYRP1	SLC27A2	ZNF202
CAT	FADS2	KRT1	PGM1	SLC28A1	ZNF205
CAV1	FADS3	KRT10	PGM3	SLC29A1	ZNF207
CAV2	FAH	KRT12	PGM5	SLC29A2	ZNF208
CAV3	FAIM2	KRT13	PGR	SLC2A1	ZNF211
CBARA1	FAIM3	KRT14	PGRMC1	SLC2A2	ZNF217
CBFA2T2	FAM102A	KRT15	PGRMC2	SLC2A3	ZNF22
CBFA2T3	FAM105A	KRT16	PHACTR2	SLC2A4	ZNF222
CBFB	FAM107A	KRT17	PHACTR4	SLC2A5	ZNF23
CBL	FAM120A	KRT18	PHB	SLC30A1	ZNF238
CBLB	FAM20B	KRT20	PHB2	SLC30A3	ZNF239
CBLN1	FAM36A	KRT33A	PHC2	SLC30A4	ZNF24
CBR1	FAM3A	KRT4	PHEX	SLC30A9	ZNF248
CBR3	FAM48A	KRT5	PHF1	SLC31A2	ZNF250
CBR4	FAM49A	KRT7	PHF14	SLC33A1	ZNF253
CBS	FAM50A	KRT9	PHF15	SLC34A1	ZNF254
CBX1	FAM50B	KRTAP5-9	PHF16	SLC35A1	ZNF259
CBX4	FAM53B	KSR1	PHF2	SLC35A2	ZNF263
CBX5	FAM5B	KTN1	PHF20	SLC35A3	ZNF264
CBX6	FAM5C	KYNU	PHF21A	SLC35B1	ZNF266
CBX7	FAM76A	L1CAM	PHF3	SLC35D1	ZNF267
CCBL1	FAM89B	LAD1	PHGDH	SLC35D2	ZNF268
CCBP2	FAM8A1	LAG3	PHIP	SLC36A1	ZNF273
CCDC22	FANCA	LAIR1	PHKA1	SLC37A4	ZNF274
CCDC28A	FANCC	LAIR2	PHKA2	SLC38A3	ZNF3
CCDC46	FANCG	LALBA	PHKB	SLC38A6	ZNF318
CCDC6	FANCL	LAMA1	PHKG1	SLC39A14	ZNF32
CCDC64	FAP	LAMA2	PHKG2	SLC39A6	ZNF324
CCDC69	FARP1	LAMA3	PHLDA1	SLC39A8	ZNF330
CCHCR1	FARP2	LAMA4	PHLDA2	SLC43A1	ZNF337
CCIN	FARS2	LAMA5	PHLDB1	SLC4A1	ZNF345

CCK	FAS	LAMB1	PHOX2B	SLC4A2	ZNF35
CCKAR	FASLG	LAMB2	PHTF1	SLC4A3	ZNF354A
CCKBR	FASN	LAMB3	PHTF2	SLC4A4	ZNF365
CCL1	FASTK	LAMC1	PHYH	SLC4A7	ZNF384
CCL11	FASTKD2	LAMC2	PHYHIP	SLC4A8	ZNF41
CCL13	FAT2	LAMP1	PI15	SLC5A1	ZNF410
CCL16	FAU	LAMP2	PI3	SLC5A2	ZNF415
CCL17	FBL	LAMP3	PIAS1	SLC5A3	ZNF423
CCL18	FBLN1	LANCL1	PIAS2	SLC5A4	ZNF43
CCL19	FBLN2	LAPTM4A	PIAS3	SLC5A5	ZNF44
CCL2	FBLN5	LAPTM4B	PICALM	SLC5A6	ZNF443
CCL20	FBN1	LAPTM5	PICK1	SLC6A1	ZNF45
CCL21	FBN2	LARGE	PIGA	SLC6A11	ZNF451
CCL22	FBP1	LARP1	PIGB	SLC6A12	ZNF467
CCL23	FBP2	LARP4	PIGC	SLC6A13	ZNF473
CCL25	FBXL14	LARP7	PIGF	SLC6A15	ZNF507
CCL27	FBXL2	LARS2	PIGH	SLC6A2	ZNF510
CCL5	FBXL4	LAS1L	PIGK	SLC6A3	ZNF529
CCL7	FBXL5	LASP1	PIGL	SLC6A4	ZNF536
CCL8	FBXL7	LBP	PIGO	SLC6A5	ZNF549
CCNA1	FBXO21	LBR	PIGQ	SLC6A6	ZNF592
CCNA2	FBXO28	LBX1	PIGR	SLC6A7	ZNF593
CCNB1	FBXO46	LCAT	PIK3C2A	SLC6A8	ZNF643
CCNB2	FBXO7	LCE2B	PIK3C2B	SLC6A9	ZNF646
CCNC	FBXO9	LCK	PIK3C2G	SLC7A1	ZNF652
CCND1	FBXW11	LCMT2	PIK3C3	SLC7A11	ZNF688
CCND2	FCAR	LCN1	PIK3CA	SLC7A2	ZNF7
CCND3	FCER1G	LCN2	PIK3CB	SLC7A4	ZNF710
CCNE1	FCER2	LCP1	PIK3CD	SLC7A5	ZNF76
CCNE2	FCGR2B	LCP2	PIK3CG	SLC7A6	ZNF79
CCNF	FCGRT	LCT	PIK3R1	SLC7A7	ZNF8
CCNG1	FCHSD2	LDB1	PIK3R2	SLC7A8	ZNF80
CCNG2	FCN1	LDB2	PIK3R3	SLC8A1	ZNF85
CCNH	FCN2	LDHA	PIK3R4	SLC9A1	ZNF91

CCNI	FCN3	LDHB	PIM1	SLC9A3	ZNHIT1
CCNT1	FDFT1	LDHC	PIM2	SLC9A3R1	ZNHIT3
CCNT2	FDX1	LDLR	PIN1	SLC9A3R2	ZP2
CCPG1	FDXR	LDOC1	PINK1	SLC9A6	ZPBP
CCR1	FECH	LECT1	PIP	SLC9A7	ZW10
CCR3	FEM1B	LECT2	PIP5K1A	SLC9A8	ZWINT
CCR4	FEM1C	LEF1	PIP5K1B	SLCO1A2	ZYX
CCR7	FEN1	LEFTY1	PIP5K1C	SLCO2A1	ZZEF1
CCR8	FER	LEFTY2	PIR	SLCO2B1	ZZZ3
CCR9	FES	LEP	PISD	SLIT1	
CCRL2	FETUB	LEPR	PITPNA	SLIT2	
CCS	FEV	LEPREL2	PITPNB	SLITRK3	
CCT2	FEZ1	LEPROT	PITPNM1	SLN	
CCT3	FEZ2	LEPROTL1	PITRM1	SLPI	
CCT4	FGA	LETMD1	PITX1	SLURP1	
CCT5	FGB	LFNG	PITX2	SMAD1	
CCT6A	FGD1	LGALS1	PITX3	SMAD2	
CCT6B	FGF1	LGALS2	PIWIL1	SMAD3	
CCT7	FGF12	LGALS3	PJA2	SMAD4	
CCT8	FGF13	LGALS3BP	PKD1	SMAD5	
CD14	FGF18	LGALS4	PKD2	SMAD6	
CD151	FGF2	LGALS8	PKIA	SMAD7	
CD160	FGF3	LGALS9	PKIG	SMAD9	
CD163	FGF4	LGI1	PKLR	SMAP1	
CD164	FGF5	LGMN	PKM2	SMARCA1	
CD180	FGF6	LGR5	PKMYT1	SMARCA2	
CD19	FGF7	LHB	PKN1	SMARCA4	
CD1A	FGF8	LHCGR	PKN2	SMARCA5	
CD1B	FGF9	LHFPL2	PKNOX1	SMARCB1	
CD1C	FGFBP1	LHX1	PKP1	SMARCC1	
CD1D	FGFR1	LHX2	PKP2	SMARCC2	
CD1E	FGFR1OP	LIAS	PKP3	SMARCD1	
CD2	FGFR2	LIF	PKP4	SMARCD2	
CD200	FGFR3	LIFR	PLA2G10	SMARCD3	

CD22	FGFR4	LIG1	PLA2G1B	SMARCE1	
CD226	FGG	LIG3	PLA2G2A	SMC2	
CD24	FGL1	LIG4	PLA2G4A	SMC3	
CD247	FGL2	LILRA1	PLA2G4C	SMC4	
CD2AP	FGR	LILRA2	PLA2G5	SMCP	
CD2BP2	FH	LILRA3	PLA2G6	SMG1	
CD300A	FHIT	LILRA4	PLA2G7	SMG5	
CD300C	FHL1	LILRB1	PLA2R1	SMG6	
CD302	FHL2	LILRB2	PLAA	SMG7	
CD33	FIBP	LILRB3	PLAG1	SMNDC1	
CD34	FIGF	LILRB4	PLAGL1	SMO	
CD36	FKBP1A	LILRB5	PLAGL2	SMOX	
CD37	FKBP2	LIMK1	PLAT	SMPD1	
CD38	FKBP4	LIMK2	PLAU	SMPD2	
CD3D	FKBP5	LIMS1	PLAUR	SMPDL3A	
CD3E	FKBP6	LIN7A	PLCB1	SMPDL3B	
CD3G	FKBP8	LIPA	PLCB2	SMR3A	
CD4	FLAD1	LIPC	PLCB3	SMR3B	
CD40	FLI1	LIPE	PLCB4	SMTN	
CD40LG	FLII	LIPF	PLCD1	SMURF2	
CD44	FLNA	LIPT1	PLCE1	SMYD2	
CD46	FLNB	LITAF	PLCG1	SMYD5	
CD47	FLNC	LLGL1	PLCG2	SNAI2	
CD48	FLOT1	LLGL2	PLCH1	SNAP23	
CD5	FLOT2	LMAN1	PLCL1	SNAP25	
CD52	FLRT1	LMAN2	PLCL2	SNAP91	
CD53	FLRT2	LMNA	PLD1	SNAPC1	
CD55	FLT3	LMNB1	PLD2	SNAPC2	
CD58	FLT3LG	LMNB2	PLD3	SNAPC3	
CD59	FLT4	LMO1	PLEK	SNAPC4	
CD5L	FMNL1	LMO2	PLEKHA6	SNAPC5	
CD6	FMO1	LMO3	PLEKHB1	SNCA	
CD63	FMO2	LMO4	PLEKHB2	SNCB	
CD69	FMO3	LMO7	PLEKHG3	SNCG	

CD7	FMO4	LMOD1	PLEKHM1	SND1	
CD72	FMO5	LMTK2	PLG	SNN	
CD74	FMOD	LMX1B	PLK1	SNPH	
CD79A	FMR1	LNPEP	PLK2	SNRK	
CD79B	FN1	LOC81691	PLK3	SNRPA	
CD80	FNBP1L	LOR	PLK4	SNRPA1	
CD81	FNBP4	LOX	PLN	SNRPB	
CD83	FNDC3A	LOXL1	PLOD1	SNRPB2	
CD84	FNTA	LPA	PLOD2	SNRPC	
CD86	FNTB	LPGAT1	PLOD3	SNRPD2	
CD8A	FOLH1	LPHN1	PLP1	SNRPD3	
CD8B	FOLR1	LPHN2	PLP2	SNRPE	
CD9	FOLR2	LPHN3	PLS1	SNRPF	
CD96	FOLR3	LPIN1	PLS3	SNRPN	
CD97	FOS	LPIN2	PLSCR1	SNTA1	
CD99	FOSB	LPL	PLTP	SNTB1	
CDA	FOSL1	LPO	PLXDC1	SNUPN	
CDC14A	FOSL2	LPP	PLXNA2	SNW1	
CDC14B	FOXA1	LPPR4	PLXNA3	SNX1	
CDC16	FOXA2	LPXN	PLXNB1	SNX13	
CDC20	FOXC1	LRBA	PLXNB3	SNX15	
CDC23	FOXC2	LRCH1	PLXNC1	SNX17	
CDC25A	FOXD1	LRIG1	PLXND1	SNX19	
CDC25B	FOXD2	LRIG2	PMAIP1	SNX2	
CDC25C	FOXE1	LRMP	PML	SNX27	
CDC27	FOXF1	LRP10	PMM1	SNX3	
CDC34	FOXF2	LRP2	PMP2	SNX4	
CDC37	FOXH1	LRP4	PMP22	SNX7	
CDC40	FOXI1	LRP5	PMPCA	SOAT2	
CDC42	FOXJ1	LRP6	PMPCB	SOCS1	
CDC42BPA	FOXJ2	LRP8	PMS1	SOCS2	
CDC42EP1	FOXJ3	LRPAP1	PMVK	SOCS3	
CDC42EP2	FOXK2	LRPPRC	PNLIP	SOCS5	
CDC42EP3	FOXM1	LRRC14	PNLIPRP1	SOCS6	

CDC42EP4	FOXN1	LRRC15	PNLIPRP2	SOD1	
CDC5L	FPGS	LRRC17	PNMA2	SOD2	
CDC6	FPGT	LRRC23	PNMT	SOD3	
CDC7	FPR1	LRRC32	PNN	SOLH	
CDH1	FRAT2	LRRC41	PNOC	SON	
CDH11	FRG1	LRRC42	PNPLA2	SORBS2	
CDH12	FRK	LRRC48	PNPLA4	SORBS3	
CDH13	FRMPD1	LRRC6	PNRC1	SORCS3	
CDH15	FRY	LRRFIP1	PODXL	SORL1	
CDH16	FRZB	LRRN3	POFUT1	SORT1	
CDH17	FSCN1	LRRTM2	POFUT2	SOS1	
CDH18	FSCN2	LSAMP	POGZ	SOSTDC1	
CDH19	FSHB	LSM1	POLA2	SOX1	
CDH2	FSHR	LSM14A	POLB	SOX10	
CDH22	FST	LSM2	POLD1	SOX11	
CDH3	FSTL1	LSM3	POLD4	SOX12	
CDH4	FSTL3	LSM4	POLDIP3	SOX13	
CDH5	FSTL4	LSM5	POLE	SOX15	
CDH6	FTL	LSM6	POLE2	SOX2	
CDH8	FTSJ1	LSM7	POLE3	SOX21	
CDK10	FUBP1	LSP1	POLG	SOX3	
CDK2	FUCA1	LSR	POLG2	SOX30	
CDK2AP2	FURIN	LSS	POLR1C	SOX4	
CDK3	FUT1	LST1	POLR2A	SOX5	
CDK4	FUT2	LTA	POLR2B	SOX9	
CDK5	FUT3	LTA4H	POLR2C	SP1	
CDK5R1	FUT4	LTB	POLR2D	SP100	
CDK5R2	FUT5	LTB4R	POLR2F	SP110	
CDK6	FUT6	LTBP1	POLR2G	SP140	
CDK7	FUT7	LTBP4	POLR2H	SP2	
CDK8	FUT8	LTBR	POLR2I	SP4	
CDK9	FUT9	LTC4S	POLR2K	SPAG1	
CDKL1	FXN	LTF	POLR2L	SPAG5	
CDKL2	FXR1	LTK	POLR3C	SPAG6	

CDKL5	FXR2	LUM	POLR3D	SPAG7	
CDKN1A	FXYD1	LUZP1	POLR3F	SPAG8	
CDKN1B	FXYD2	LY6D	POLR3G	SPAG9	
CDKN1C	FXYD3	LY6E	POLRMT	SPAM1	
CDKN2A	FYB	LY6G6C	POMC	SPARC	
CDKN2B	FYN	LY6H	PON1	SPARCL1	
CDKN2C	FZD1	LY75	PON2	SPAST	
CDKN2D	FZD2	LY86	PON3	SPATA2	
CDKN3	FZD5	LY9	POP4	SPDEF	
CDO1	FZD6	LY96	POP5	SPEN	
CDON	FZD7	LYL1	POP7	SPG20	
CDR1	FZD9	LYN	POR	SPG7	
CDR2	FZR1	LYPD1	POSTN	SPHK2	
CDR2L	G0S2	LYPD3	POT1	SPI1	
CDRT1	G3BP2	LYPLA1	POU1F1	SPIB	
CDS1	G6PC	LYPLA2	POU2AF1	SPINK1	
CDS2	GAA	LYST	POU2F1	SPINK2	
CDSN	GAB1	LYZ	POU2F2	SPINK4	
CDT1	GAB2	LYZL6	POU2F3	SPINK5	
CDV3	GABARAP	LZTR1	POU3F1	SPINLW1	
CDX1	GABARAPL1	M6PR	POU3F2	SPINT1	
CDX2	GABARAPL2	MAB21L1	POU3F4	SPINT2	
CDYL	GABBR1	MACF1	POU4F2	SPN	
CEACAM1	GABBR2	MAD1L1	POU6F1	SPOCK2	
CEACAM3	GABPA	MAD2L1	PPAP2A	SPON1	
CEACAM4	GABRA1	MAD2L1BP	PPAP2B	SPOP	
CEACAM5	GABRA2	MADCAM1	PPAP2C	SPP1	
CEACAM6	GABRA3	MADD	PPARA	SPP2	
CEACAM7	GABRA5	MAEA	PPARD	SPPL2B	
CEACAM8	GABRA6	MAF	PPARG	SPR	
CEBPA	GABRB1	MAFF	PPAT	SPRR1A	
CEBPB	GABRB2	MAFK	PPBP	SPRR1B	
CEBPD	GABRB3	MAG	PPEF1	SPRY1	
CEBPE	GABRD	MAGEA1	PPEF2	SPRY2	

CEBPG	GABRE	MAGEA10	PPFIA1	SPTA1	
CEBPZ	GABRG2	MAGEA11	PPFIA2	SPTAN1	
CEL	GABRG3	MAGEA12	PPFIBP1	SPTB	
CELSR1	GABRR1	MAGEA4	PPFIBP2	SPTBN1	
CELSR2	GABRR2	MAGEA5	PPIB	SPTBN2	
CENPA	GAD1	MAGEA8	PPIC	SPTLC1	
CENPB	GAD2	MAGEB1	PPID	SPTLC2	
CENPC1	GADD45A	MAGEB2	PPIE	SQLE	
CENPE	GADD45B	MAGEB3	PPIF	SQSTM1	
CENPF	GADD45G	MAGEB4	PPIG	SRC	
CENPI	GAK	MAGEC1	PPIH	SRCAP	
CEP135	GAL	MAGED1	PPIL2	SRD5A1	
CEP164	GAL3ST1	MAGED2	PPIL6	SRD5A2	
CEP170	GALC	MAGI1	PPL	SREBF1	
CEP250	GALE	MAGI2	PPM1A	SREBF2	
CEP290	GALK1	MAGOH	PPM1B	SRF	
CEP57	GALK2	MAL	PPM1D	SRGAP2	
CEP68	GALNS	MALT1	PPM1E	SRGAP3	
CES1	GALNT1	MAML1	PPM1F	SRI	
CES2	GALNT10	MAML3	PPM1G	SRM	
CETN1	GALNT2	MAN1A1	PPOX	SRP19	
CETN2	GALNT3	MAN1A2	PPP1CA	SRP54	
CETN3	GALR2	MAN1C1	PPP1CB	SRP72	
CETP	GALR3	MAN2A1	PPP1CC	SRPK1	
CFD	GALT	MAN2A2	PPP1R10	SRPK2	
CFDP1	GAMT	MAN2B1	PPP1R11	SRPR	
CFH	GANAB	MAN2B2	PPP1R12A	SRPX	
CFHR2	GAP43	MAN2C1	PPP1R12B	SRPX2	
CFI	GAPDH	MANBA	PPP1R13B	SRR	
CFL1	GAPDHS	MAOA	PPP1R15A	SRRM1	
CFLAR	GARS	MAOB	PPP1R16B	SRRM2	
CFTR	GART	MAP1A	PPP1R2	SS18	
CGA	GAS1	MAP1B	PPP1R3A	SS18L1	
CGGBP1	GAS2	MAP1LC3B	PPP1R3C	SSB	

CGREF1	GAS2L1	MAP2	PPP1R3D	SSBP1	
CGRRF1	GAS7	MAP2K1	PPP1R7	SSBP2	
CH25H	GAS8	MAP2K3	PPP1R8	SSFA2	
CHAD	GAST	MAP2K4	PPP2CA	SSH1	
CHAF1A	GATA1	MAP2K5	PPP2CB	SSNA1	
CHAF1B	GATA2	MAP2K6	PPP2R1A	SSPN	
CHD1	GATA3	MAP2K7	PPP2R1B	SSR1	
CHD1L	GATA4	MAP3K10	PPP2R2A	SSR2	
CHD2	GATA6	MAP3K11	PPP2R2B	SSR4	
CHD3	GATAD1	MAP3K12	PPP2R3A	SSRP1	
CHD4	GATM	MAP3K13	PPP2R4	SST	
CHD5	GBAS	MAP3K14	PPP2R5A	SSTR2	
CHD8	GBE1	MAP3K3	PPP2R5B	SSTR3	
CHD9	GBF1	MAP3K4	PPP2R5C	SSTR4	
CHEK1	GBP1	MAP3K7	PPP2R5D	SSTR5	
CHEK2	GBP2	MAP3K8	PPP2R5E	SSX1	
CHERP	GBX2	MAP3K9	PPP3CA	SSX2IP	
CHGA	GC	MAP4	PPP3CB	SSX5	
CHGB	GCA	MAP4K1	PPP3CC	ST13	
CHI3L1	GCAT	MAP4K2	PPP4R1	ST14	
CHI3L2	GCC2	MAP4K4	PPP5C	ST18	
CHKA	GCDH	MAP4K5	PPP6C	ST3GAL1	
CHL1	GCG	MAP7	PPRC1	ST3GAL2	
CHM	GCGR	MAPK1	PPT1	ST3GAL4	
CHML	GCH1	MAPK10	PPWD1	ST3GAL5	
CHMP2A	GCHFR	MAPK11	PPY	ST3GAL6	
CHMP2B	GCK	MAPK12	PPYR1	ST5	
CHMP7	GCKR	MAPK13	PQBP1	ST6GAL1	
CHN1	GCLC	MAPK14	PRAF2	ST6GALNAC2	

Ek 3 İstatistiksel Önem Testi Sonuçları

Anova1

Funfem ve Mclust ROC değerleri için Anova1 sonuçları:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Mclust	1	0.16328	0.16328	164.7	1.4e-15	***
Residuals	39	0.03866	0.00099			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Funfem ve DE(Nudge) ROC değerleri için Anova1 sonuçları:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Nudge	1	0.18463	0.18463	416.1	<2e-16	***
Residuals	39	0.01731	0.00044			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Mclust ve DE(Nudge) ROC değerleri için Anova1 sonuçları:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Nudge	1	0.11418	0.11418	774.9	<2e-16	***
Residuals	39	0.00575	0.00015			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Kmeans ve DE(Nudge) ROC değerleri için Anova1 sonuçları:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Nudge	1	0.13069	0.13069	335.2	<2e-16	***
Residuals	39	0.01521	0.00039			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Funfem ve Kmeans ROC değerleri için Anova1 sonuçları:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Kmeans	1	0.19599	0.19599	1284	<2e-16	***
Residuals	39	0.00595	0.00015			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Mclust ve Kmeans ROC değerleri için Anova1 sonuçları:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Mclust	1	0.11444	0.11444	141.8	1.45e-14	***
Residuals	39	0.03147	0.00081			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Mclust ve Funfem AP değerleri için Anova1 sonuçları:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Mclust	1	0.27113	0.27113	224.9	<2e-16	***
Residuals	39	0.04702	0.00121			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Funfem ve DE(nudge) AP değerleri için Anova1 sonuçları:

Df Sum Sq Mean Sq F value Pr(>F)

Nudge 1 0.26673 0.26673 202.3 <2e-16 ***

Residuals 39 0.05142 0.00132

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Funfem ve Kmeans AP değerleri için Anova1 sonuçları:

Df Sum Sq Mean Sq F value Pr(>F)

Kmeans 1 0.30133 0.30133 698.6 <2e-16 ***

Residuals 39 0.01682 0.00043

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Kmeans ve DE(nudge) AP değerleri için Anova1 sonuçları:

Df Sum Sq Mean Sq F value Pr(>F)

Nudge 1 0.13176 0.13176 363.6 <2e-16 ***

Residuals 39 0.01413 0.00036

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Kmeans ve Mclust AP değerleri için Anova1 sonuçları:

Df Sum Sq Mean Sq F value Pr(>F)

Mclust 1 0.12449 0.12449 226.7 <2e-16 ***

Residuals 39 0.02141 0.00055

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Kmeans ve Funfem AP değerleri için Anova1 sonuçları:

Df Sum Sq Mean Sq F value Pr(>F)

Funfem 1 0.13818 0.1382 698.6 <2e-16 ***

Residuals 39 0.00771 0.0002

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

T-test

Funfem ve Mclust ROC değerleri için t-test sonuçları:

t = 13.5989, df = 75.122, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.1626045 0.2184191

sample estimates:

mean of x mean of y

0.7588565 0.5683448

Funfem ve DE(Nudge) ROC değerleri için t-test sonuçları:

t = 5.4634, df = 68.832, p-value = 6.98e-07

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.07043236 0.15145930

sample estimates:

mean of x mean of y

0.7588565 0.6479107

Mclust ve DE(Nudge) ROC değerleri için t-test sonuçları:

t = -4.1797, df = 59.01, p-value = 9.791e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.1176569 -0.0414750

sample estimates:

mean of x mean of y

0.5683448 0.6479107

Funfem ve Kmeans ROC değerleri için t-test sonuçları:

t = 18.0977, df = 79.549, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.2437632 0.3040019

sample estimates:

mean of x mean of y

0.7588565 0.4849740

Mclust ve Kmeans ROC değerleri için t-test sonuçları:

t = 6.2312, df = 77.407, p-value = 2.242e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.0567307 0.1100108

sample estimates:

mean of x mean of y

0.5683448 0.4849740

DE(Nudge) ve Kmeans ROC değerleri için t-test sonuçları:

t = -8.1968, df = 65.826, p-value = 1.198e-11

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.2026264 -0.1232470

sample estimates:

mean of x mean of y

0.4849740 0.6479107

Funfem ve Mclust AP değerleri için t-test sonuçları:

t = 4.2698, df = 80, p-value = 5.341e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.04495219 0.12343318

sample estimates:

mean of x mean of y

0.6943366 0.6101439

Funfem ve DE(Nudge) AP değerleri için t-test sonuçları:

t = 3.3802, df = 75.782, p-value = 0.001147

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.03128438 0.12104245

sample estimates:

mean of x mean of y

0.6943366 0.6181732

Mclust ve DE(Nudge) AP değerleri için t-test sonuçları:

t = -0.356, df = 75.854, p-value = 0.7228

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.05294648 0.03688794

sample estimates:

mean of x mean of y

0.6101439 0.6181732

Funfem ve Kmeans AP değerleri için t-test sonuçları:

t = 12.173, df = 70.311, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.1712051 0.2382924

sample estimates:

mean of x mean of y

0.6943366 0.4895878

Mclust ve Kmeans AP değerleri için t-test sonuçları:

t = 7.1563, df = 70.222, p-value = 6.38e-10

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.08695957 0.15415263

sample estimates:

mean of x mean of y

0.6101439 0.4895878

DE(Nudge) ve Kmeans AP değerleri için t-test sonuçları:

t = -6.4078, df = 60.993, p-value = 2.389e-08

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.16871198 -0.08845875

sample estimates:

mean of x mean of y

0.4895878 0.6181732