GEO VERİ TABANINA KAYITLI MİKRODİZİ VERİLERİNİN ANALİZİ İÇİN YENİ BİR BİLGİSAYAR PROGRAMININ TASARLANMASI

OSMAN VARIŞLI

TEMMUZ 2022
DİYARBAKIR

DİCLE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GEO VERİ TABANINA KAYITLI MİKRODİZİ VERİLERİNİN ANALİZİ İÇİN YENİ BİR BİLGİSAYAR PROGRAMININ TASARLANMASI

OSMAN VARIŞLI

DİCLE ÜNİVERSİTESİ LİSANSÜSTÜ EĞİTİM-ÖĞRETİM VE SINAV YÖNETMELİĞİNİN BİR PARÇASI OLARAK
ELEKTRİK - ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALINDA
YÜKSEK LİSANS TEZİ
OLARAK HAZIRLANMIŞTIR

TEMMUZ 2022

DİYARBAKIR

GEO VERİ TABANINA KAYITLI MİKRODİZİ VERİLERİNİN ANALİZİ İÇİN YENİ BİR BİLGİSAYAR PROGRAMININ TASARLANMASI

Osman VARIŞLI tarafından Dicle Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin bir parçası olarak hazırlanan bu çalışma, aşağıda bilgileri yazılı jüri üyeleri tarafından değerlendirilerek **Elektrik - Elektronik Mühendisliği Ana Bilim Dalı**'nda **Yüksek Lisans Tezi** olarak kabul edilmiştir.

Prof. Dr. Neslihan Dalkılıç _____
Müdür, **Fen Bilimleri Enstitüsü**

Dr.Öğr. Üyesi Cafer Budak _____
Danışman, **Elektrik Elektronik Mühendisliği,**
Dicle Üniversitesi

Sınav Jürisi:

Dr.Öğr. Üyesi Cafer Budak _____
Elektrik Elektronik Mühendisliği Bölümü, Dicle Üniversitesi

Dr.Öğr. Üyesi Hüseyin Acar _____
Elektrik Elektronik Mühendisliği Bölümü, Dicle Üniversitesi

Doç. Dr. Yılmaz Kaya _____
Bilgisayar Mühendisliği Bölümü, Batman Üniversitesi

ONAY

Savunma Tarihi: 28 / 06 / 2022

Anneme…

Dicle Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırlanan bu tez çalışmasında yer alan tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu beyan ederim. Ayrıca, bahse konu bu kural ve ilkelerin gerektirdiği üzere, bu çalışmada özgün olmayan tüm bilimsel içerikleri kurallara uygun biçimde alıntılayıp kaynak gösterdiğimi beyan ederim. Beyanımla çelişen herhangi bir delil bulunduğu takdirde tüm sorumluluğu üstleneceğimi kabul ederim.

Ad, Soyad: Osman VARIŞLI

İmza: …….……

TEŞEKKÜR

Bu tez çalışmasının her aşamasında yardımını, bilimsel katkılarını, tecrübelerini hiçbir zaman esirgemeyen ve sabırla beni çalışmaya teşvik eden değerli hocam Dr. Öğr. Üyesi Cafer Budak'a teşekkür ederim. Tez yazım aşamasında engin bilgilerini benimle paylaşıp devamlı motive olmamı sağlayan değerli hocam Dr. Öğr. Üyesi M. Ali ARSERİM'e teşekkür ederim.

Manevi desteklerini esirgemeyen ve her zaman yanımda olan sevgili aileme teşekkürlerimi bir borç bilirim.

İÇİNDEKİLER

TEŞEKKÜR v
İÇİNDEKİLER vi
ŞEKİLLER LİSTESİ viii
TABLOLAR LİSTESİ ix
SİMGELER VE KISALTMALAR LİSTESİ x
ÖZET xii
ABSTRACT xiii
1. GİRİŞ 14
2. GENEL BİLGİLER 16
2.1 Hücre 16
2.2 Nükleik Asitler 16
2.3 Gen Expresiyonu 19
2.4 Mikrodiziler 21
2.5 Gene Expression Omnıbus (GEO) 25
2.5.1 GEO'da kayıtlı veriler 25
2.6 Kümeleme Algoritmaları 28
2.7 K-Means KümelemeAlgoritması 28
2.8 PCA (Temel Bileşen Analizi) 31
3. MATERYAL VE YÖNTEM 33
3.1 Programın Oluşturulması 33
3.2 Flask 33
3.3 JQuery 33
3.4 Global Normalizasyon 33
4. BULGULAR VE TARTIŞMA 35
4.1 Programın Kullanılması 35
4.2 Programın Test Edilmesi 40

4.2.1 Veri setinin seçilmesi 40
5. SONUÇ 48
KAYNAKLAR 51
ÖZGEÇMİŞ 54

ŞEKİLLER LİSTESİ

Şekil 2.1 Deoksiribonükleotid ve Ribonükleotid [2] 17
Şekil 2.2 DNA Yapısı [3] 18
Şekil 2.3 Protein Sentezi[4] 20
Şekil 2.4 GeneChip [13] 22
Şekil 2.5 Genlerin kesişim noktaları [13] 23
Şekil 2.6 Affymetrik çipleri ve yüzeyinde bulunan probların yerleşimi [13] 24
Şekil 2.7 Affymetrix Gene Chip Human Genome U133 Plus 2.0 [13] 24
Şekil 2.8 GEO web sitesi 27
Şekil 2.9 K-Means için oluşturulan data ve rastgele oluşturulan merkezler. 29
Şekil 2.10 K-Means merkezinin iyileştirilmesi için hesaplamaların yapılması. 29
Şekil 2.11 K-Means yeni merkezlerin belirlenmesi. 30
Şekil 2.12 K-Means gruplamanın bittiği durum 31
Şekil 2.13 Deneysel boyut indirgenme işlemi 32
Şekil 4.14. GEO accession değerinin ve platform değerinin girilmesi 35
Şekil 4.15 GEOMT Normalizasyon İşlemi 36
Şekil 4.16 GEOMT Gruplama İşlemi 36
Şekil 4.17 Normalize Edilmiş Gruplar 37
Şekil 4.18 Kat Değişim Değerlerinin Gösterilmesi 37
Şekil 4.19 Volcano Grafiği 38
Şekil 4.20 Kat değişim sıralanmış hali ile tabloların oluşturulması 39
Şekil 4.21 Kat değişim için K-Means grafiği 39
Şekil 4.22 Hücrelerdeki geleneksel androjen-AR mekanizması [25] 41

TABLOLAR LİSTESİ

Tablo 3.1 16 Saat androjen uygulamasıyla expresiyonu değişen genlerin listesi. 42
Tablo 3.2 Androjen uyarısıyla ekspresyonunun değiştiği belirlenen genlerin dâhil olabileceği hücresel mekanizmalar 46
Tablo 3.3 Gen Açıklamaları 46
Tablo 5.4 Farklı araçlar tarafından sunulan veri analizi yöntemleri 49

SİMGELER VE KISALTMALAR LİSTESİ

Simge	**Açıklama**
µg	Mikrogram
µm	Mikrometre

Kısaltma	**Açıklama**
GEOMT	GEO Microarray Tool
DNA	Deoksiribo Nükleik Asit
piRNA	Piwi Etkileşimli RNA
RNA	Ribo Nükleik Asit
rRNA	Ribozomal RNA
SAGE	Gen İfadesinin Seri Analizi
siRNA	Küçük İnterferenzci RNA
snoRNA	Küçük Nükleolar RNA
mRNA	Mesajcı RNA
miRNA	Mikro RNA
A	Adenin
G	Guanin
C	Sitozin
T	Timin
Pentoz	5 karbonlu bir şeker
cDNA	Tamamlayıcı DNA
tRNA	Taşıyıcı RNA
GEO	Gene Expression Omnıbus
GSE	Gene Expression Data Series
GPL	Gene Expression Data Platform
GDS	Gene Expression Data DataSet
NCBI	National Centre for Biotecnology Information (Ulusal Biyoteknoloji Bilgi Merkezi)
PCR	Polimeraz Zincir Reaksiyonu

GB	Giga Bayt
PCA	Principal Component Analysis

ÖZET

GEO VERİ TABANINA KAYITLI MİKRODİZİ VERİLERİNİN ANALİZİ İÇİN YENİ BİR BİLGİSAYAR PROGRAMININ TASARLANMASI

Osman VARIŞLI
Yüksek Lisans, Elektrik - Elektronik Mühendisliği Bölümü
Danışman: Dr. Öğr. Üyesi Cafer Budak
Temmuz 2022, 56 sayfa

Gen ekspresyonunun mikrodizi tabanlı analizi, biyomedikal araştırmalar için oldukça önemli deney sistemleridir. Bununla beraber bu deneylerden elde edilen verilerin analizi uygun yazılım araçlarıyla desteklenmediği takdirde ya sonuç elde edilememekte ya da hatalı sonuçlar elde edilmektedir. Bu nedenle, bu verilerin analizini istenilen kalitede yapabilecek ve grafik sistemleri ile bunu görselleştirebilecek programlar bu alanda çalışan araştırmacılar açısından oldukça önemlidir. Güncel olarak kullanılmakta olan yazılımlar incelendiğinde açık kaynak kodlu olmayışları, gerekli raporların indirilmesine izin vermemeleri, yeterli grafik desteği olmayışları, işletim sistemi bağımsız çalışamamaları ya da güncel geliştirilen algoritma ve teknolojileri kullanmamaları gibi bir veya birden fazla eksiklik barındırdıkları görülmektedir. Bu tür eksiklikleri olmayan yeni bir yazılım geliştirilmesi mikrodizi tabanlı gen ekspresyon analizi yapan veya yapmayı planlayan temel bilim araştırmacıları ve klinik hekimler için önemli olacaktır.

Bu tez çalışmasında GEO analizi için kullanımı kolay ve ücretsiz bir web aracı olan, GEOMT(GEO Microarray Tool) tasarlanmıştır. GEOMT, özellikle daha fazla programlama ve algoritma yeterliliği veya biyoinformatik bilgisi olmayan araştırmacılara için değerli bir araç olacaktır.

Anahtar Kelimeler: GEO, Mikrodizi, Affymetrix, Biyoinformatik, WSGI

ABSTRACT

DESIGNING OF A NEW COMPUTER PROGRAM FOR ANALYSIS OF MICROARRAY DATA THAT DEPOSITED IN GEO DATABASE

Varışlı, Osman
Master of Science in Department of Electronics Engineering
Supervisor: Dr.Öğr. Üyesi Cafer Budak
July 2022, 56 pages

Microarray-based analysis of gene expression is very important experimental systems for biomedical research. However, if the analysis of the data obtained from these experiments is not supported with appropriate software tools, either no results can be obtained or erroneous results are obtained. For this reason, programs that can analyze these data with the desired quality and visualize it with graphic systems are very important for researchers working in this field. When the software currently being used is examined, it is seen that they have one or more deficiencies such as not being open source, not allowing the download of required reports, not having sufficient graphic support, not working independently of the operating system, or not using up-to-date algorithms and technologies. The development of new software that does not have such shortcomings will be important for basic science researchers and clinicians who are doing or planning to perform microarray-based gene expression analysis.

In this thesis, an easy-to-use and free web tool, GEOMT (GEO Microarray Tool), was designed for GEO analysis. GEOMT will be a valuable tool, especially for researchers without further programming and algorithm proficiency or bioinformatics knowledge.

Keywords: GEO, Microarray, Affymetrix, Bioinformatics,WSGI

1. GİRİŞ

Hücrelerin temel özelliklerinden biri uyarılara cevap verebilmeleridir. Çeşitli uyaranlara karşı geliştirilen en yaygın hücresel cevap ise genellikle gen ekspresyonu değişimiyle sağlanmaktadır. Gen ekspresyonu, DNA üzerinde bulunan genlerin kalıp olarak kullanılması ile mRNA transkripsiyonunu ifade eder. Daha genel bir tanımlama ile hücreler değişen koşullara bağlı olarak bazı genlerin transkripsiyonunu arttırırken bazı genlerin transkripsiyonunu ise azaltır. Bu şekilde yapılan düzenlemeler ile hücrelerdeki çoğalma, farklılaşma, metabolik faaliyetler v.b. bütün mekanizmalar kontrol edilir.

Gen ekspresyonundaki değişimler patolojik koşullarda da gerçekleşmektedir. Bu durumlardaki değişimler bazen metabolizmanın hastalığa karşı savunmasını aktive ederken bazen de hastalık etkenine bağlı olarak gerçekleşip hastalığın prognozunu etkilemektedir. Bu şekilde bilinen en iyi örnekler kanser oluşumu ve gelişmesinde meydana gelen gen ekspresyon değişimleridir.

Bir genin ekspresyonu onun aktivitesini yani RNA sentezini ifade etmektedir. Bu nedenle değişen koşulların gen aktivitesindeki değişimlerini incelemek için gen ekspresyon analizleri uzun yıllardan beri yapılmaktadır. Özellikle ilk dönem analiz yöntemleri olarak Northern Blotting gibi zaman alıcı ve manipülasyona nispeten daha açık yöntemler kullanılırken günümüzde Real Time PCR gibi daha elverişli sistemler kullanılmaktadır.

Gen ekspresyonu için kullanılan geleneksel yöntemler (northern blotting, real time PCR v.b.) genellikle tek bir genin ekspresyonunu araştırma temelinde çalışmaktadır ve araştırılacak her bir genin ekspresyonu için ayrı ayrı deney seti kurulmaktadır. Bununla beraber özellikle ökaryot genomlarının 10.000'lerce gen içerdiği ve herhangi bir uyaran durumunda bunların aktivitelerinde farklı kombinasyonlarda değişimler meydana geldiği göz önüne alındığında bu yöntemler yetersiz kalmaktadır. Bu nedenle High Throughput analizler ön plana çıkmaktadır. Bu yöntemler ile gerek genom ve gerekse de RNA ve protein seviyesinde yüksek kapasiteli analizler yapılabilmektedir. RNA seviyesinde yapılan high throughput analizler temel olarak üç büyük grupta toplanabilmektedir. Bunlar; SAGE, mikrodizi

teknolojisi ve RNA sekanslama stratejileridir. Bunlardan SAGE nispeten biraz eski bir analiz yöntemidir ve günümüzde pek fazla kullanılmamaktadır. Her ne kadar bu alandaki en güncel yöntem RNA sekanslama stratejisi olsa bile bu yöntem için gerekli deneysel altyapının oldukça pahalı olması, tecrübeli personelin bulunmasının zorluğu ve gerekli yazılım teknolojilerinin kompleksliği araştırmacıları ağırlıklı olarak mikrodizi teknolojisine yönlendirmektedir.

Araştırmacılar tarafından yapılan mikrodizi sonuçlarının bilimsel dergilerde yayınlanması için elde edilen ham verilerin belirli veri tabanlarında kayıt edilmesi zorunludur. Bu şekilde mikrodizi ham verilerinin dünyadaki bütün araştırmacıların kullanımına sunulması hedeflenmektedir. Mikrodizi sonuçlarının kayıt edilmesi gereken en büyük ve kapsamlı veri tabanı Gene Expression Omnibus (GEO)'dur. GEO, 2 milyondan fazla gen ekspresyonu verisi içeren devasa bir platformdur ve dünyadaki bütün araştırmacılar kolaylıkla bu verilere girip analiz edebilmektedir. Bununla beraber bu verilerin analizi için iyi bir Biyoinformatik ve bilgisayar bilgisine sahip olmak gerekmektedir. GEO'da bulunan veriler temelde sağlık bilimleri ile ilişkili olduğundan dolayı araştırmacılar bunların analizinde sorunlar yaşamakta ve çoğu zaman bilgisayar kullanım eksikliğinden dolayı ya sonuç elde edememekte ya da yanlış sonuçlar elde etmektedirler. Bu sorunun üstesinden gelmek için hazırlanan birkaç program ise R temelli olarak çalışmakta ve yine iyi bir bilgisayar bilgisi gerektirmektedir. Bu nedenle mevcut eksikliği gidermek için Biyoinformatik bilgisi olmayan bilim insanlarının ve/veya klinikçilerin GEO'da bulunan verileri analiz edebileceği bir aracın varlığı oldukça değerli olacaktır. Bu noktadan yola çıkarak önerilen bu tez çalışmasında Bilgisayar programlama, algoritma tasarımı ve biyoinformatik bilgisi olmayan biyomedikal bilimlerde çalışan laboratuvar araştırmacıları ve klinikçilerin GEO'da kayıtlı bulunan mikrodizi verilerini analiz edebileceği bir bilgisayar programı tasarlanmıştır.

2. GENEL BİLGİLER

2.1 Hücre

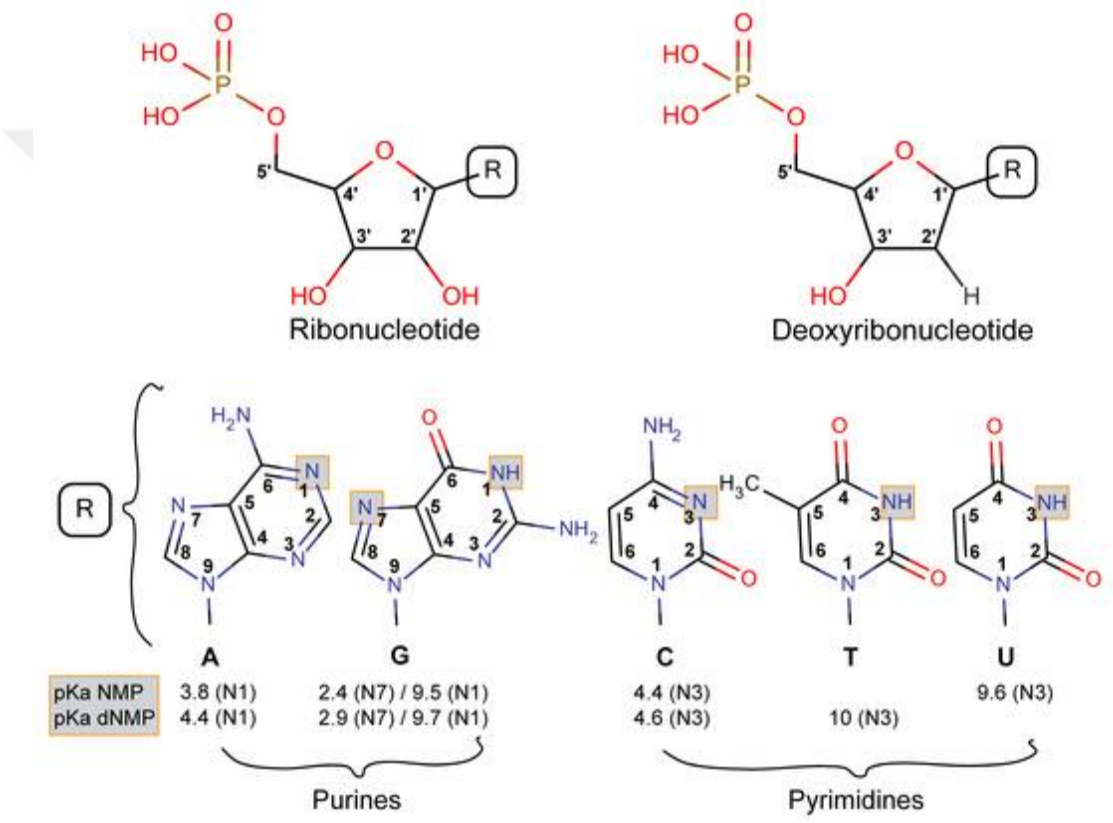
Hücreler, ilk olarak 1600'lü yıllarda Robert Hook tarafından birkaç mercek kullanılarak yapılmış ilkel bir mikroskop kullanılarak ve ölü bir mantar dokusundaki boş gözenekler halinde gözlenmiştir. Ardından gelen yaklaşık olarak 2 yüzyıllık bir süre içinde tek hücreliler ile beraber hayvan ve bitki hücreleri de gözlemlenmeye başlanmış ve bu amaçla bir Hücre Teorisi ortaya atılmıştır. Hücre teorisi, temelde "tüm canlı organizmalar hücrelerden meydana geldiği" ve "bir hücre ancak başka bir hücreden meydana gelebilir" ifadeleri ile özetlenebilir. Her ne kadar değişiklikler ve eklemeler yapılmış olsa bile bu ifadeler bugün de geçerliliğini korumaktadır.

Hücreler yaşamın temel birimleridirler. Bakteri, maya ve amip gibi bazı mikroorganizmalar tek hücre halinde bulunurken hayvan ve bitkiler çoğunlukla doku adı verilen kollektifler halinde organize olan trilyonlarca hücreden oluşurlar. Bu anlamda hücreler çoğalırlar, binlerce metabolik faaliyet yürütürler ve ölürler. Ayrıca özellikle çok hücreli organizmalarda hücreler bulundukları doku ve organa bağlı olarak belirli fonksiyonlar yapmak üzere farklılaşarak özelleşirler. Her ne kadar bu faaliyetler çok farklı mekanizmalar tarafından kontrol edilip düzenlense bile, hücreler temelde taşıdıkları nükleik asitlerin iyi düzenlenmiş bir koordinasyonunu kullanarak bütün bu aktiviteleri gerçekleştirilmektedir.

2.2 Nükleik Asitler

Hücrelerde bulunan nükleik asitler deoksiribonükleik asit (DNA) ve ribonükleik asit (RNA) olmak üzere iki çeşittir. Hem DNA hem de RNA nükleotidlerden meydana gelmektedir. Bir nükleotid, 5 karbonlu bir şeker (pentoz), bir baz ve bir de fosfat molekülünden meydana gelmektedir. Şekil 2.1'de görüldüğü gibi DNA'nın yapısında Adenin (A), Guanin (G), Sitozin (C) ve Timin (T) bazlarını taşıyan deoksiribonükleotidler bulunurken RNA'nın yapısında Adenin (A), Guanin (G), Sitozin (C) ve Urasil (U) bazı bulunan ribonükleotidler bulunur. Nükleotidler olarak bilinen bu dört harfin çeşitli kombinasyonları, insanlar ve tüm canlılar arasındaki çeşitli biyolojik farklılıkları ortaya çıkarır. İnsan genomu 3 milyar'dan fazla nükleotid'den meydana gelen oldukça büyük bir DNA molekülüne sahiptir. Bu anlamda her konumda dört olasılık (A, T, C veya G) ile zilyonlarca kombinasyon

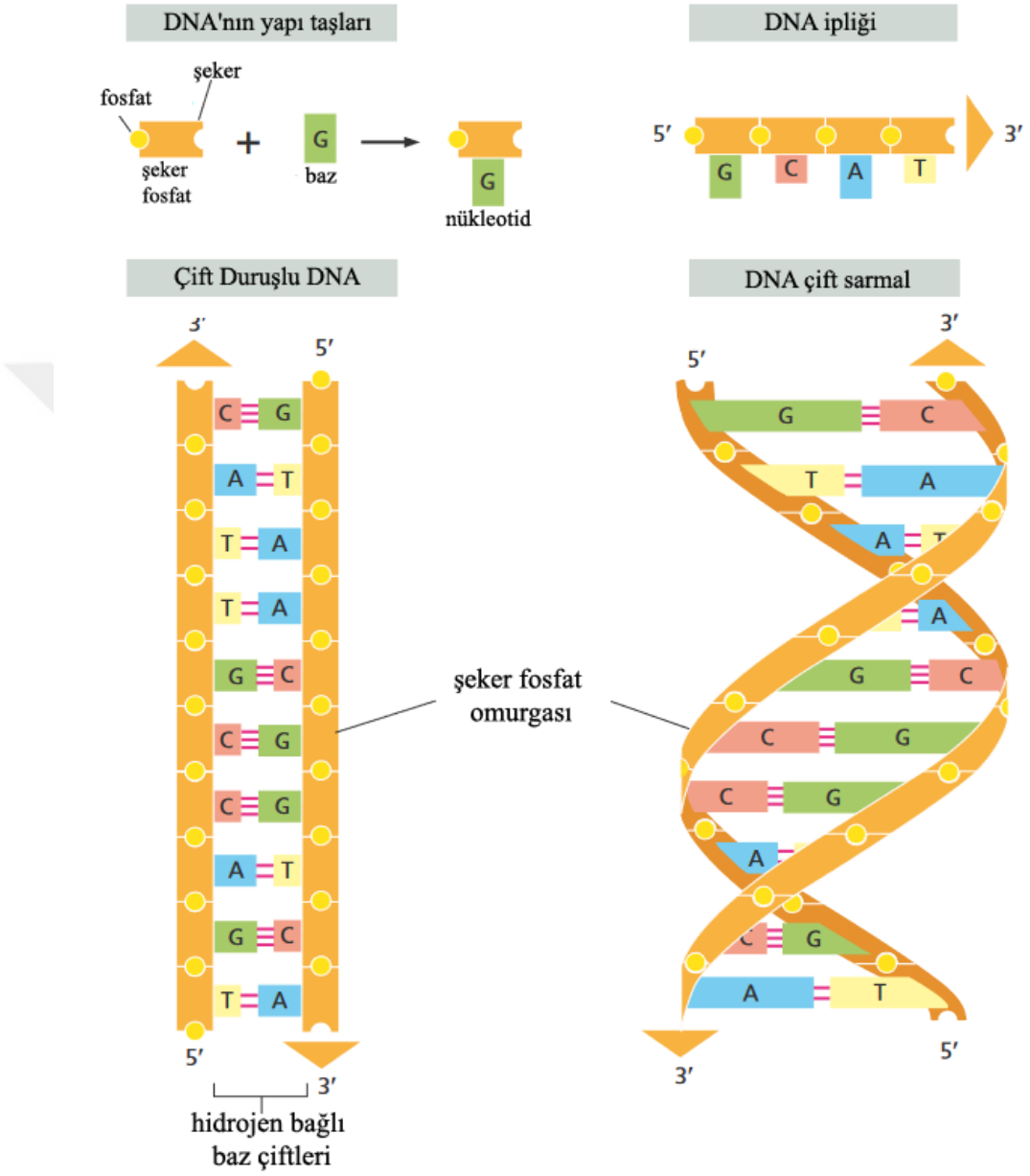
oluşması mümkündür. Aslında, DNA'nın bilgi içeriği, bilgisayarların veri saklama yönetimine benzemektedir. Bilgisayarlarda ikili veri sisteminin (birler ve sıfırlar) farklı kombinasyonları kullanılırken, DNA'da 4 nükleotidin farklı kombinasyonları kullanılmaktadır.[1, pp. 20–21]



Şekil 2.1 Deoksiribonükleotid ve Ribonükleotid [2]

Şekil 2.2'de gösterildiği gibi DNA çift zinciri meydana gelirken her bir zincirdeki nükleotidlerin şeker-fosfat omurgaları dışarıda ve bazları ise birbirine bakacak şekilde içeride bulunurlar. Bu anlamda DNA'nın çift zincir oluşturması, onun monomeri olan nükleotidlerin taşıdığı bazlara göre olmaktadır. DNA'nın yapısında Adenin bazı taşıyan nükleotid karşı zincirde bulunan ve Timin bazı taşıyan nükleotidle 2 hidrojen bağı kurarak eşleşirken Guanin bazı taşıyan nükleotid de karşı

zincirde bulunan Sitozin bazı taşıyan nükleotid ile 3 hidrojen bağı kurarak eşleşir. Bu olaya komplementerlik adı verilir.



Şekil 2.2 DNA Yapısı [3]

Her ne kadar RNA hücre içinde DNA'nın aksine tek zincirli olarak davransa bile her zaman için çift zincir oluşturma eğilimindedir. Bu nedenle hücre içinde RNA çift zincirinin oluşmasının engellenmesi için çeşitli mekanizmalar bulunmaktadır. Ancak hücre dışı deney sistemlerinde bu mekanizmalar bulunmayacağı için RNA

komplementerliği yaygın olarak oluşmakta ve deney sistemlerinde de kullanılmaktadır. Bu anlamda komplementerlik RNA açısından da çok farklı değildir. Sadece DNA'dan farklı olarak RNA'nın yapısında Timin nükleotidi bulunmaz ve onun yerine Urasil nükleotidi geçerek karşı zincirde bulunan Adenin bazları ile 2 hidrojen bağı kurarak eşleşir (kıyaslama sadece baz içeriği üzerinden yapılmıştır). RNA'nın yapısında bulunan Sitozin ve Guanin bazlarında ise DNA'dan farklı bir durum yoktur ve karşı zincirlerde bulunan Sitozin ve Guanin'in arasında tıpkı DNA'da olduğu gibi aralarında 3 hidrojen bağı oluşabilir ve böylece karşı zincirler bir araya gelebilir.

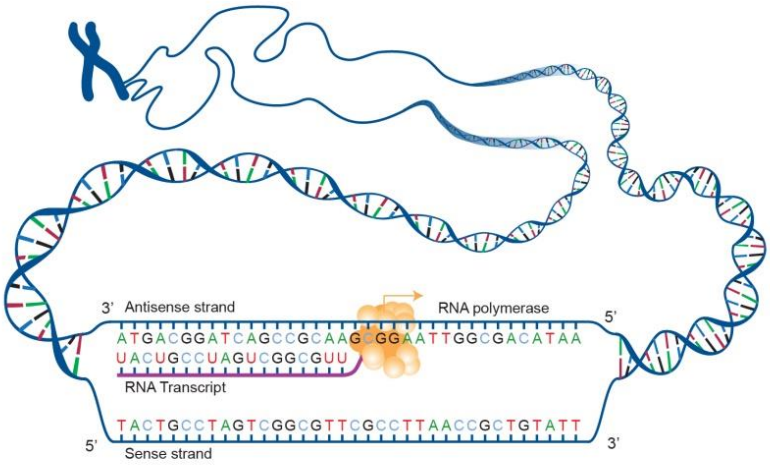
2.3 Gen Expresiyonu

Gen kavramı, moleküler biyolojide muhtemelen açıklaması en çok değişen kavramlardan biridir. Her ne kadar geçmişte bir gen= bir enzim ve bir gen= bir protein gibi teoriler ortaya atılmış olsa bile günümüzde bütün genlerin enzim yapısındaki proteinleri kodlamadığı ya da hiçbir protein kodlamayan genlerin varlığının gösterilmiş olması nedeniyle genler en az bir fonksiyonel ürün (RNA) kodlayan fonksiyonel DNA bölgeleri olarak tanımlamaktadır. Aslında günümüzde kodlanan RNA'ların büyük bir kısmının protein kodlanması için gerekli bilgiyi taşımadığı bilinmektedir. Bu anlamda kodlanan RNA'lar çok farklı şekillerde sınıflandırılabilse bile temelde beş büyük gruba ayrılırlar. Bunlar;

- mRNA
- rRNA
- tRNA
- Küçük kodlamayan RNA'lar (miRNA, piRNA, siRNA, snoRNA, …)
- Uzun kodlamayan RNA'lar (LncRNA)

Buna karşın yine de protein sentezine kalıplık etmelerinden dolayı mRNA'ların ve gen ekspresyonunun düzenlenmesine birden fazla noktada düzenleyici olarak katılmalarından dolayı miRNA'ların özel bir yerleri vardır.

Çoğu zaman genin aktivitesi ile de özdeş olarak kullanılan gen ekspresyonu kavramı, bir genin bulunduğu bölgedeki DNA'nın kaynak olarak kullanılmasıyla RNA'nın transkripsiyonunu ifade etmektedir. Şekil 2.3'de gösterildiği gibi transkripsiyon gerçekleşirken kalıp olarak kullanılan DNA zinciri üzerindeki bazlara komplementer ribonükleotidler RNA Polimeraz adı verilen bir enzim aracılığı ile enzimatik olarak uç uca fosfodiester bağı ile eklenerek bir zincir oluşturulur.



Şekil 2.3 Protein Sentezi[4]

Yüksek yapılı organizmaların genomunda 10.000'lerce genin olduğu ve bu genlerin bir koordinasyon içinde ekspresyonunun düzenlendiği bilinmektedir. Gen ekspresyonunda bu şekilde yapılan düzenlemeler ile hücrelerdeki çoğalma, farklılaşma, metabolik faaliyetler v.b. bütün mekanizmalar kontrol edilir. Benzer bir şekilde patolojik koşullarda da hücrelerdeki gen ekspresyonunda değişimler gerçekleşmektedir. Bu durumlardaki değişimler bazen metabolizmanın hastalığa karşı savunmasını aktive ederken bazen de hastalık etkenine bağlı olarak gerçekleşip hastalığın prognozunu etkilemektedir. Bu şekilde bilinen en iyi örneklerden biri kanser oluşumu ve gelişmesinde meydana gelen gen ekspresyon değişimleridir [5].

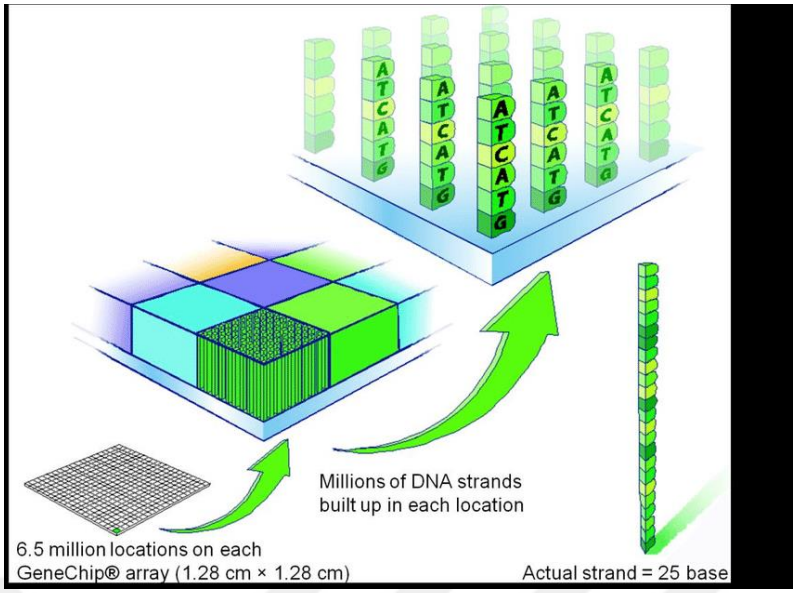
Gen aktivitesi incelemek için gen ekspresyon analizleri uzun yıllardan beri yapılmaktadır. Özellikle ilk dönem analiz yöntemi olarak Northern Blotting gibi zaman alıcı ve manipülasyona nispeten daha açık yöntemler kullanılırken günümüzde Real Time PCR gibi daha elverişli sistemler kullanılmaktadır [6, pp.

376–399]. Bununla beraber, kullanılan geleneksel yöntemler (northern blotting, real time PCR v.b.) genellikle tek bir genin ekspresyonunu araştırma temelinde çalışmaktadır ve araştırılacak her bir genin ekspresyonu için ayrı ayrı deney seti kurulmaktadır. Bununla beraber özellikle ökaryot genomlarının 10.000'lerce gen içerdiği ve herhangi bir uyaran durumunda bunların aktivitelerinde farklı kombinasyonlarda değişimler meydana geldiği göz önüne alındığında bu yöntemler yetersiz kalmaktadır [6, pp. 376–379]. Bu nedenle High Throughput yöntemler adı verilen ve eş zamanlı olarak binlerce genin analizine imkan veren analiz yöntemleri ön plana çıkmaktadır. Bu yöntemler ile gerek genom ve gerekse de RNA ve protein seviyesinde yüksek kapasiteli analizler yapılabilmektedir. RNA seviyesinde yapılan yüksek verimli analizler temel olarak üç büyük grupta toplanabilmektedir. Bunlar; SAGE[7], mikrodizi teknolojisi[5] ve RNA sekanslama [8] stratejileridir. Bunlardan SAGE nispeten biraz eski bir analiz yöntemidir ve günümüzde pek fazla kullanılmamaktadır. Her ne kadar bu alandaki en güncel yöntem RNA sekanslama stratejisi olsa bile bu yöntem için gerekli deneysel altyapının oldukça pahalı olması, tecrübeli personelin bulunmasının zorluğu ve gerekli yazılım teknolojilerinin kompleksliği araştırmacıları ağırlıklı olarak mikrodizi teknolojisine yönlendirmektedir [4, 7, 8].

2.4 Mikrodiziler

Mikrodizi teknolojisi ile, DNA üzerindeki tek nükleotid polimorfizmleri, proteinlerin DNA'ya bağlanma bölgelerinin belirlenmesi gibi pek çok analiz yüksek kapasite ile yapılabilmektedir. Bununla beraber bu teknolojinin en sık kullanım amacı gen ekspresyon değişimlerini araştırmaktır. Mikrodizi teknolojisi ile tek bir deneyle binlerce genin ekspresyon değişimleri eş zamanlı olarak analiz edilebilmektedir. Bu verilerde sağlık bilimleri açısından gerek hastalığın tanısında ve gerekse de tedavi yaklaşımı ve prognoz takibi için kullanılabilmektedir [9,11].

Şekil 2.4'de gösterildiği gibi mikrodizi teknolojisi, temelde prob olarak isimlendirilen binlerce tek zincirli kısa DNA oligonükleotidinin cam, silikon ya da plastikten meydana gelen katı bir yüzeye tutturulması ile hazırlanan çipleri kullanmaktadır.

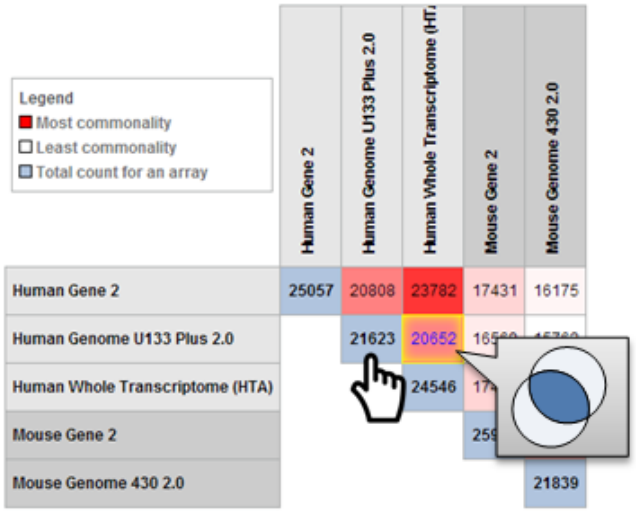


Şekil 2.4 GeneChip [13]

Her ne kadar ilk dönem mikrodizi çipleri manuel olarak yapılsa bile günümüzde farklı organizmalar (insan, fare, sıçan v.b.) için ve farklı amaçlar için kullanılabilecek (mRNA, miRNA, DNA v.b.) çipler ticari olarak satılmaktadır. Çiplerin yüzeyindeki problar ilgili genom tarafından kodlanan RNA'lar referans olarak alınarak ve tek bir RNA'ya özgü olacak şekilde tasarlanmaktadır. Bu şekilde her bir prob sadece bir tek RNA'ya bağlanabilmektedir [11].

İnsan genomunda yaklaşık olarak 32.000 gen bulunmaktadır ve en az bunun iki katı kadar da bu genlerden transkribe edilen RNA olduğu bilinmektedir. Bununla beraber en yeni mikrodizi versiyonlarında bile 25.000 civarı prob bulunmakta ve bu problar ancak 10-12.000 civarı gen ile eşleşmektedir. Yani bir bütün olarak insan genomu tarafından kodlanan bütün RNA'ları belirleyebilecek bir mikrodizi sistemi henüz bulunmamaktadır.

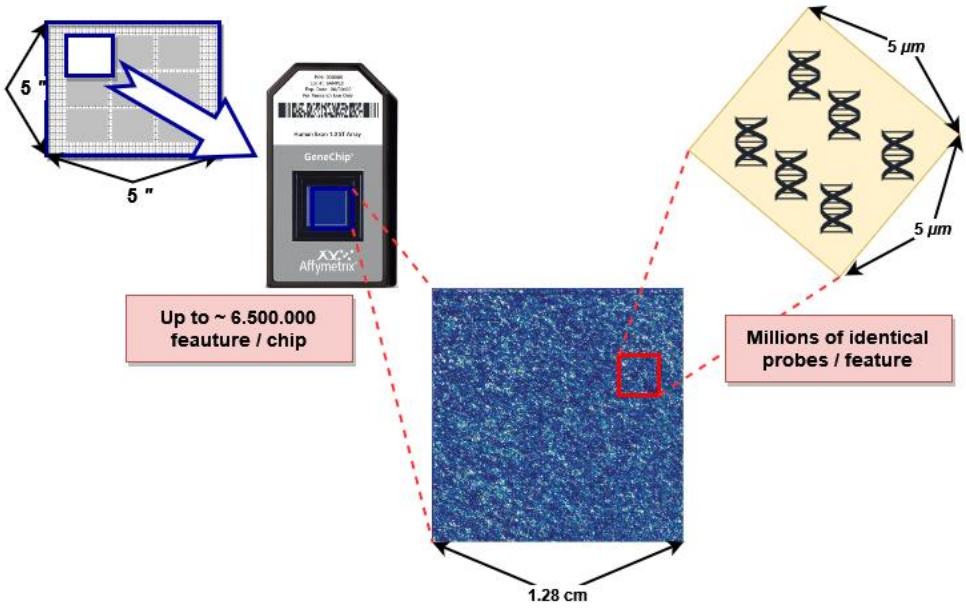
Sonuç olarak, hücreden izole edilen RNA havuzundaki bütün RNA'lar uygun pH, sıcaklık değerleri ve inhibitörlerin yokluğu gibi durumlar yaratılarak çip yüzeyindeki problar ile hibridize edilirse RNA havuzundaki her bir RNA çip yüzeyinde bulunan ve kendisine özgül olarak tasarlanan prob veya problar (çiplerin yüzeyinde genellikle iç kontrol olarak bir RNA'ya bağlanabilecek birden fazla prob yerleştirilir) ile eşleşebileceklerdir [6, pp. 38–46]



Şekil 2.5 Genlerin kesişim noktaları [13]

Çip üzerindeki probların özgül RNA'ları ile bağlanmalarının seviyelerinin sayısal veriye dönüştürülmesi için hibridizasyondan önce RNA havuzundaki bütün RNA'lar önce daha kısa parçalara enzimatik olarak fragmante edilir ve bu RNA fragmentleri Biyotin ile işaretlenir. Bu aşamadan sonra Biyotinlenmiş RNA fragmentleri çipin yüzeyinde bulunan problar ile hibridize edilerek özgül olarak bağ kurmaları için uygun ortam sağlanır. Bir lazer tarayıcı ve okuyucu yardımıyla her bir probun bulunduğu fiziksel alan için ayrı ayrı biyotin yoğunluğu ölçülür ve böylece her bir bağlanma sayısal veriye dönüştürülür [14].

Bir mikrodizi tasarlanırken, çipin yüzeyine tutturulan probun bulunduğu koordinatlar aynı zamanda bir bilgisayar yazılımına kaydederek bulunduğu yer fiziksel olarak tanımlanır. Bu fiziksel lokasyonda bir probun milyonlarca kopyası bulunmaktadır (Şekil 2.6).



Şekil 2.6 Affymetrik çipleri ve yüzeyinde bulunan probların yerleşimi [13]

Her ne kadar mikrodizi teknolojisine sahip olup bunu ticari olarak piyasaya sunan birden fazla firma bulunsa bile bu alandaki lider firma hiç kuşkusuz Affymetrix'tir. Affymetrix firmasının pek çok canlı türü ve/ veya farklı deneysel yaklaşım için piyasada bulunan çok sayıda çipi bulunmaktadır. Şekil 2.7'de Affymetrix firmasının insan için ürettiği en güncel çip Affymetrix Gene Chip Human Genome U133 Plus 2.0 gösterilmiştir. Bu çip üzerinde on binlerce gen'in kodladığı mRNA'ları özgül olarak belirleyecek problar taşımaktadır [15].



Şekil 2.7 Affymetrix Gene Chip Human Genome U133 Plus 2.0 [13]

Mikrodizi deneylerinin sonunda elde edilen sinyaller sayısal verilere dönüştürülerek iki boyutlu ve binlerce satırdan meydana gelen uzun bir tablo haline getirilir. Bu yapı vaka-kontrol çalışmalarının verimli bir şekilde analizine olanak sağlar [15] .

2.5 Gene Expression Omnıbus (GEO)

Araştırmacılar tarafından yapılan mikrodizi sonuçlarının bilimsel dergilerde yayınlanması ve bu verilerin dünyadaki diğer araştırmacıların kullanımına sunulması için bu verilerin belirli veri tabanlarında kayıt edilmesi zorunludur. Bu veri tabanlarından en büyüğü ve en kapsamlı olanı Gene Expression Omnibus (GEO)'dur. GEO, Amerika Ulusal Biyoteknoloji enstitüsünde (National Center for Biotechnology Information= NCBI [16]) merkezli olarak faaliyet gösteren, 2 milyondan fazla gen ekspresyonu verisini barındıran ve dünyadaki bütün araştırmacıların kolaylıkla erişip verilerini analiz edebildikleri devasa bir veri tabanıdır [17].

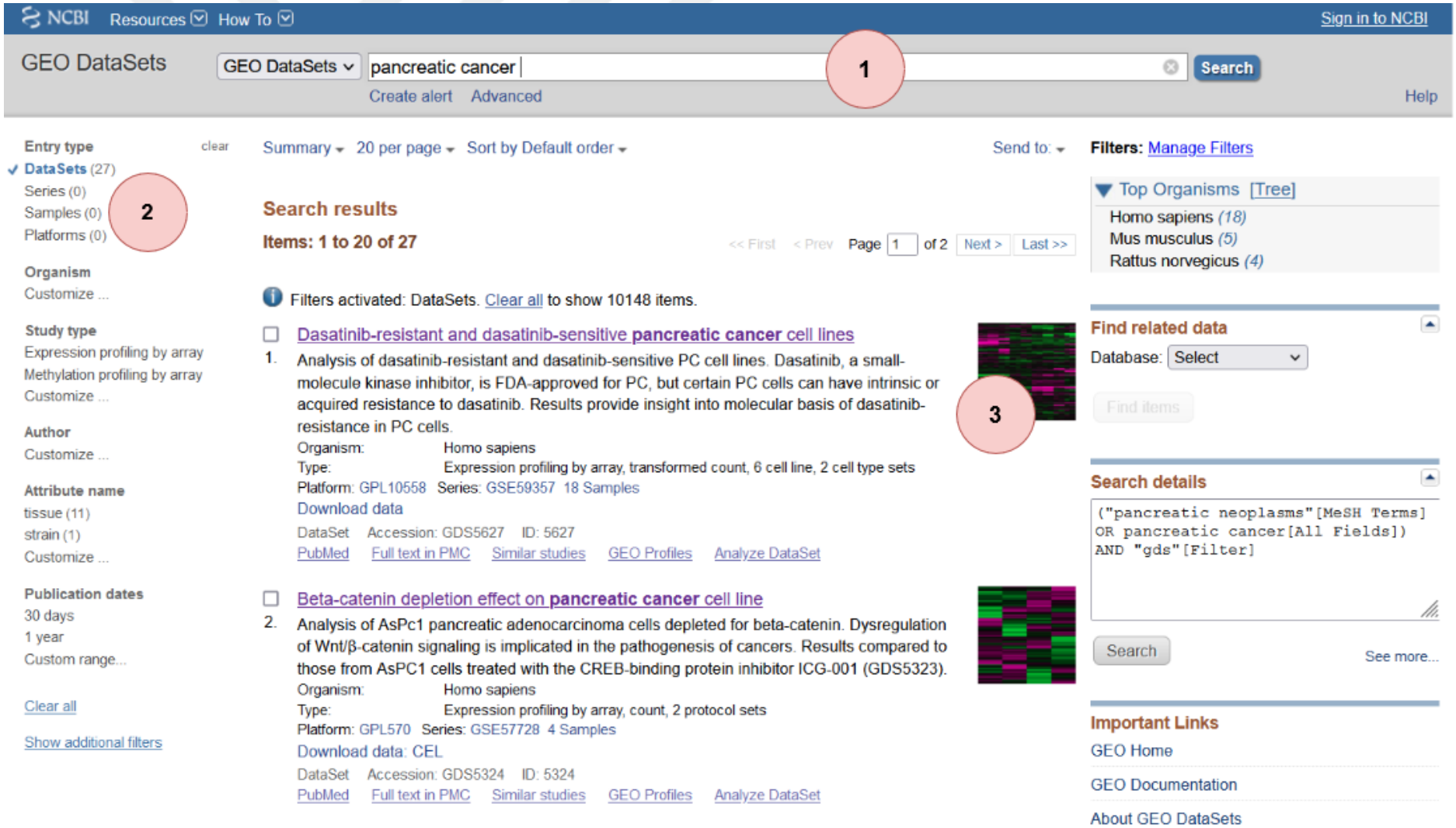
Bununla beraber, GEO'da bulunan verilerin analizi için araştırmacıların çoğu zaman biyoinformatik, bilgisayar, istatistik ve bazen de yapay zekanın makine öğrenmesi alanlarında iyi bir birikime sahip olmaları gerekmektedir. Bu alanlarda yeterli tecrübeye sahip olmayan araştırmacılar GEO verilerinin analizinde çoğu zaman sonuç alamamakta ya da yanlış sonuçlar elde etmektedirler [18]. Bu amaçla çeşitli araştırma grupları tarafından, bu tür bilgi birikimine sahip olmayan ya da az bilgiye sahip olan araştırmacıların GEO'da bulunan bu verileri kolay bir şekilde analiz edebileceği, değişik özeliklere sahip, yazılımlar üretmişlerdir [14].

2.5.1 GEO'da kayıtlı veriler

GEO'ya kayıt edilen her bir veriye bir erişim numarası verilir ve bu erişim numarası "accession number" olarak isimlendirilir. Bu anlamda erişim numaraları da kayıt edilen veriye göre belirli bir düzen içinde değişebilmektedir. GEO'da kayıtlı seriler GSE, Platformlar GPL ve DataSet'ler ise GDS ile başlayan bir erişim numarası ile etiketlenir. Bu anlamda gönderici tarafından sağlanan her veri Platform ve Seri kayıtlarının yanı sıra küratörlü gen ifadesi olan DataSet kayıtlarını da kaydeder. Diğer taraftan GEO'da kayıtlı veriler aynı zamanda anahtar kelimelere göre de

sınıflandırıldığı için ilgilenilen çalışmanın erişim numarası bilinmiyorsa bir anahtar kelime girerek de verilere ulaşmak mümkündür.

GEO verilerine erişim http://www.ncbi.nlm.nih.gov/geo/ adresinden yapılabildiği gibi, NCBI ana sayfasındaki [16] ana arama kutusundan açılan veri tabanları menüsünden de yapılabilir [19]. Erişilen her kayıt, başlık, özet, organizma ve erişimin yanı sıra ilgili verilere olan bağlantıları içerir. Şekil 2.8'de GEO web sitesinin iş akışı ekran görüntüsü gösterilmiştir. Şekil 2.8'in 1 ile işaretlenerek gösterildiği gibi GEO veri kümeleri arama kutusuna bir arama terimi yazdıktan ve Şekil 2.8'in 2 ile işaretlenerek gösterildiği gibi veri kümesi girişlerini kısıtlamak için filtre özelliğini kullandıktan sonra, Şekil 2.8'in 3 ile işaretlenerek gösterildiği gibi kullanıcı ilgili kayıtlarını alır.



Şekil 2.8 GEO web sitesi

2.6 Kümeleme Algoritmaları

Verileri benzer özelliklerine göre gruplara ayırma işlemine kümeleme denir. Genel olarak iki temel kümeleme algoritma tipi mevcuttur. Bunlardan biri "hiyerarşik kümeleme yöntemleri", diğeri ise "hiyerarşik olmayan kümeleme" yöntemleridir. Bu iki yöntem arasındaki en temel fark; kümeleme yapılacak sayı önceden belli değilse hiyerarşik kümeleme yöntemi, önceden belli ise yani verinin kaç kümeye ayrılacağı önceden belirtilmiş ise hiyerarşik olmayan kümeleme yöntemi kullanılır. GEO tarafından elde edilen datanın gerekli normalizasyon ve filtrelerden geçirildikten sonra, anormalite gösteren değerlerin bulunması için kullanılması gereken kümeleme işleminde, önceden kaç kümeye ayrılacağı belirtildiğinden, hiyerarşik olmayan kümeleme yöntemi kullanılmalıdır. Oluşturulan programda hiyerarşik olmayan kümeleme yöntemlerinden olan K-Means kümeleme yöntemi kullanıldı.

2.7 K-Means KümelemeAlgoritması

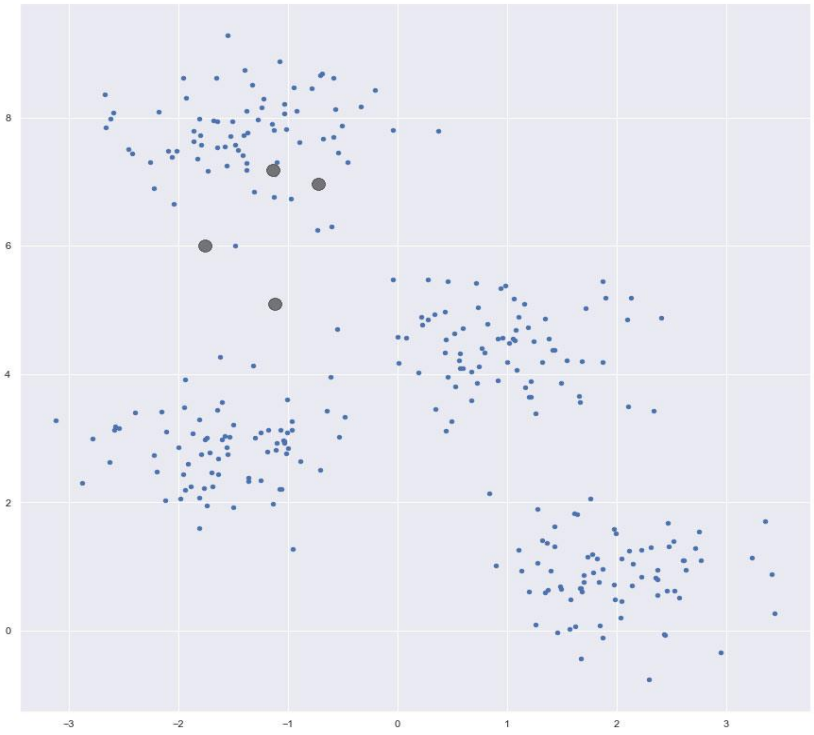
1967 yılında J.B. MacQueen tarafından geliştirilen K-Means Kümeleme Algoritması verileri sınıflandırmak için kullanılan en yaygın kümeleme algoritmalardandır. K-Means' te K değeri önceden belirtilen küme sayısıdır. K değeri önceden verilmesi gerekir, k değerinin seçimi için birkaç analiz yöntemi olsa da en makul olanı algoritmayı farklı k değerlerinde yürütüp işimiz için en çok kullanışlı olanı kullanmaktır. Çünkü farklı sayıda gruplar, farklı sonuçları çıkarabilir [20].

Algoritma K adet özgün küme oluşturup her kümenin merkezi, kümedeki değerlerin ortalamasına olan uzaklığın hesaplandığı için K-Ortalamalar denmektedir.

Algoritma istatistiksel olarak benzer nitelikteki kayıtları kümelenmesini sağlar. Bir elemanın yalnızca bir kümeye ait olmasına izin verilir. Küme merkezi kümeyi temsil eden değerdir [21].

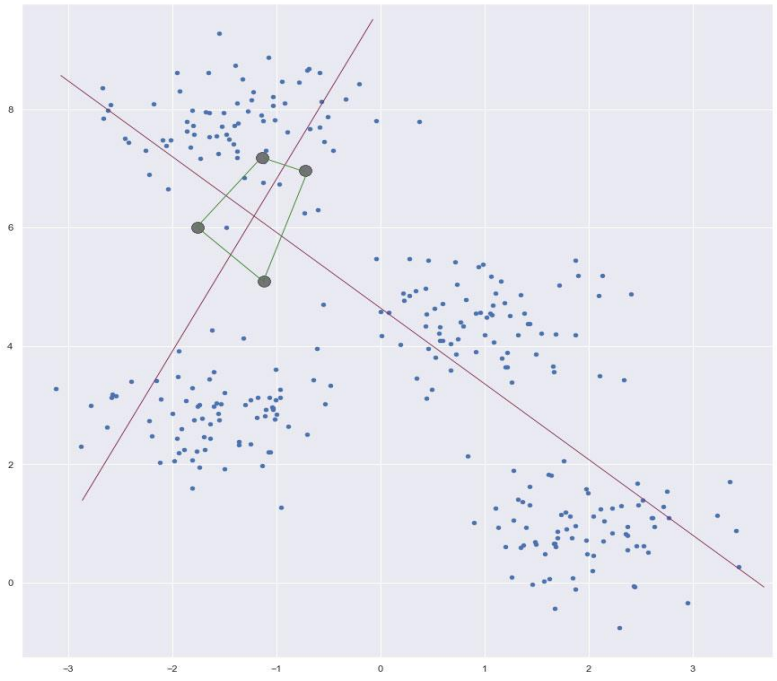
K-Means çalışma prensibi şu şekildedir.

1- Kaç küme belirlenecekse veriler içinde o kadar merkez belirlenir. Genel de merkezler rastgele belirlenir, ama önceden bir bilgi doğrultusunda merkezler önceden seçilebilir. Şekil 2.9'da oluşturulan sistemde merkezlerle ilgili herhangi bir bilgi olmadığı için merkezeler rastgele belirlendi.



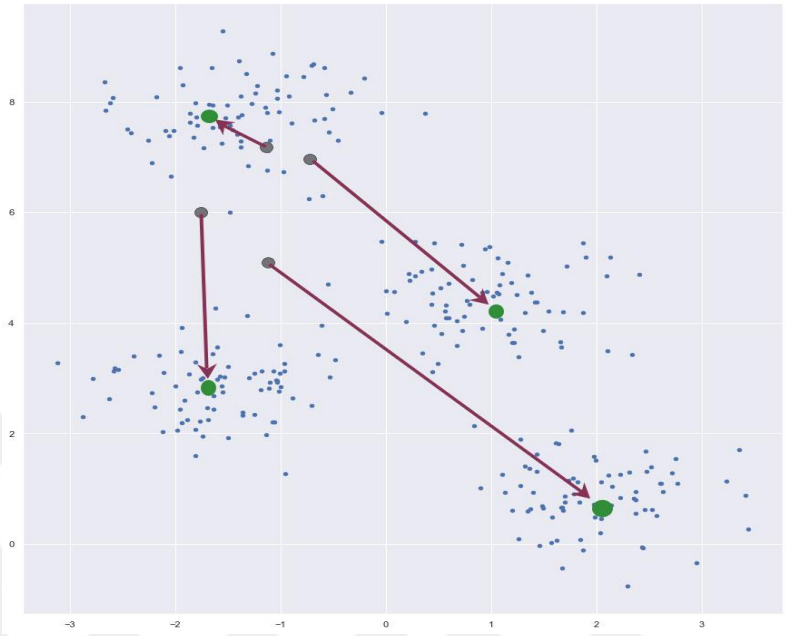
Şekil 2.9 K-Means için oluşturulan data ve rastgele oluşturulan merkezler.

2- Şekil 2.10'da gösterildiği gibi veri merkezleri arasındaki mesafe belirlenir ve ortalama konum sayısı hesaplanır (2.1). Buradaki amaç merkez konumu iyileştirmektir.



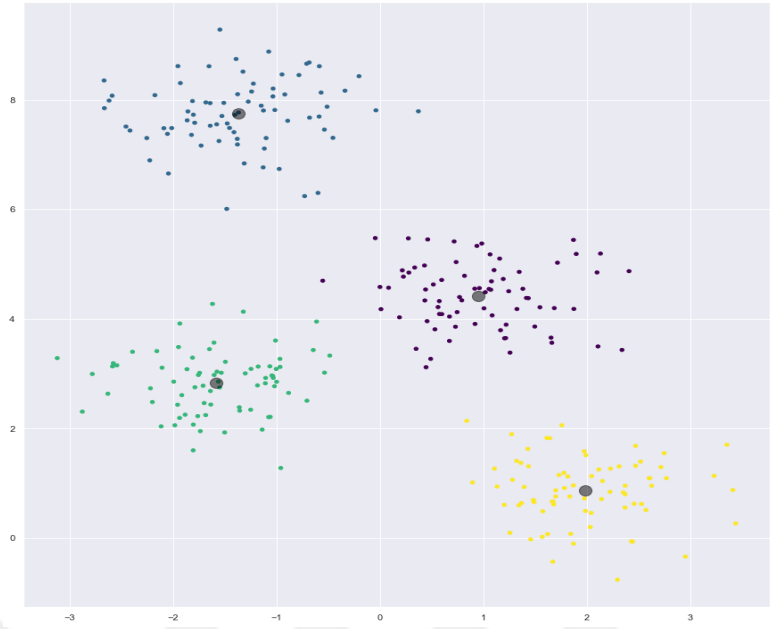
Şekil 2.10 K-Means merkezinin iyileştirilmesi için hesaplamaların yapılması.

3- Şekil 2.11'de gösterildiği gibi yapılan matematiksel işlemlerle sınıflandırma merkezleri için yeni yerler belirlenir, eski yerler silinir, guruplar büyüyüp küçülebilir.



Şekil 2.11 K-Means yeni merkezlerin belirlenmesi.

4- İşlem adımına göre (istenirse belirlenir) 2. ve 3. adımların tekrarlanması. Şekil 2.12'de görüldüğü gibi işlem bittiğinde merkezler ve guruplar netleşmiş olur.



Şekil 2.12 K-Means gruplamanın bittiği durum

$$d(A,B) = \sqrt[2]{(b1-a1)^2 + (b1-a1)^2} \tag{2.1}$$

$$W(P^q) = \frac{1}{q}\sum_{k=1}^{q}\frac{1}{n_k}\sum_{i=1}^{n_k} d(x_i, c_k) \tag{2.2}$$

GEO'dan alınacak olan verinin birden fazla boyutlu olacağından kümeleme ile mantıklı analiz edilmesi nerdeyse imkânsızdır. Bundan dolayı temel amacı boyut indirgeme olan PCA (Temel Bileşen Analizi) yöntemi kullanılmalıdır.

2.8 PCA (Temel Bileşen Analizi)

PCA'in temel mantığı çok boyutlu bir veri setinde, verilerindeki temel özellikleri inceleyip verileri daha az sayıda değişkenle göstermektir.

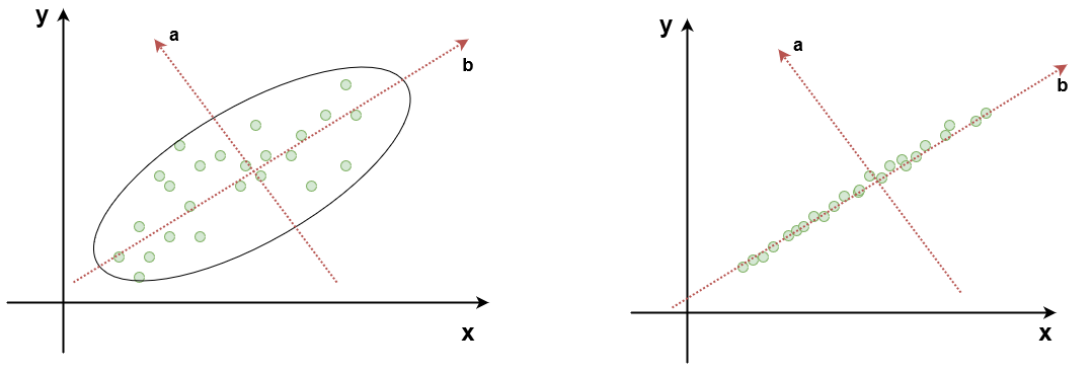
Varyans, bir veri setinin ortalamasına bulan uzaklığını ifade etmektedir, varyans hesaplanması denklem (2.3)'de verilmiştir. Varyans değeri küçük çıkarsa, veri setinin birbiri ile ilişkili olduğu anlamına gelmektedir. Varyans büyükse verilerin bir biri ile ilişkilerinin olmadığı anlamına gelmektedir.

$$S^2 = \frac{\sum(x-\overline{x})^2}{(n-1)} \tag{2.3}$$

Kovaryans, iki değişkenin bir biri ile ilişkisinin anlaşılmasını sağlar. Kovaryans hesaplanması denklem (2.4)'de verilmiştir. Kovaryans sıfırdan büyük çıkarsa iki değişkenin bir biri ile ilişkili olduğu anlamına gelir, kovaryans sıfırdan küçük çıkarsa iki değişkenin birbiriyle ilişkili olmadığı anlamına gelmektedir.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \tag{2.4}$$

Şekil 2.13'te X-Y koordinat sisteminde ölçülen iki değişkenli bir veri setini temsil ettiği ve bu veri setinin tek boyuta indirgemek istenildiği farz edilsin. Verilerin orta noktalarından a ve b eksenleri çizilir. Her (x; y) koordinatını karşılık gelen (a; b) değerine dönüştürülür, veriler kovaryans olur, yani a ve b değişkenleri arasındaki ortak varyans sıfır olur [22].



Şekil 2.13 Deneysel boyut indirgenme işlemi

3. MATERYAL VE YÖNTEM

3.1 Programın Oluşturulması

GEOMT (GEO Microarray Tool), platform bağımsız olması ve erişe bilirliği kolay olması için web ara yüzlü olarak tasarlandı. Hızlı ve kullanım esnekliğinden dolayı yazılım dili python seçildi. Web ara yüzü içinde modüler yapısından dolayı python ile oluşturulmuş bir çerçeve olan Flask kullanıldı.

3.2 Flask

Python ile yazılmış bir web uygulama çerçevesidir. Pocco adlı uluslararası Python açık kaynak kodlu çalışmalar yapan bir ekip tarafından geliştirildi. Flask ; WSGI araç takımına ve Jinja2 şablon motoruna dayanmaktadır. Her ikisi de Pocco projeleridir. WSGI (Web Sunucusu Ağ Geçidi Arayüzü) ; Python web uygulaması geliştirme için bir standart olarak kullanılmıştır. WSGI, web sunucusu ile ve web uygulamaları arasında bir köprü görevi görür. Jinja, Python için popüler bir şablon motorudur. Bir web şablon sistemi, dinamik bir web sayfası oluşturmak için bir şablonu belirli bir veri kaynağıyla birleştirir. Jinja, Python değişkenlerini HTML şablonlarına geçirmenizi sağlar:

Flask ile birlikte, dinamik bir yapı oluşturmak için javascript kütüphanesinden jQuery kullanıldı.

3.3 JQuery

Uzun JavaScript kodlarını daha kısa halde yazma kolaylığı sağlamaktadır. JQuery birçok özellik barındırmış olsa bile temel kullanılma amacı POST ve GET metotlarında ajaxtan faydalanmak istenmesindendir.

3.4 Global Normalizasyon

Elde edilmiş olan gen ekspresyon değerlerini etkilemiş olması muhtemel varyasyon kaynaklarının ortadan kaldırılması işlemidir. Her ne kadar birden fazla normalizasyon yöntemi bulunsa bile bütün yöntemler genel olarak birbirine yakın sonuçlar vermektedir. GEOMT'te analiz edilecek verilerin normalizasyonu için "global normalizasyon" yaklaşımı kullanıldı. Denklem (3.1)'de Global Normalizasyonun hesaplanma yöntemi gösterilmiştir.

$$N_{toplam} = \sum_{n=1}^{Narray} Rn \qquad (3.1)$$

"N_{toplam}" değeri her bir sütun için ayrı ayrı toplanmış değerler demektir ve dizi yapısındadır.

$$\text{Ortalama} = \frac{\sum_{n=1}^{Ss} N_{toplam}}{ST} \tag{3.2}$$

Denklem (3.2)'de "N_{toplam}" değerlerinin toplamının dizin sayısına yani "ST" değerine bölümüyle ortalama değerini elde edilmektedir.

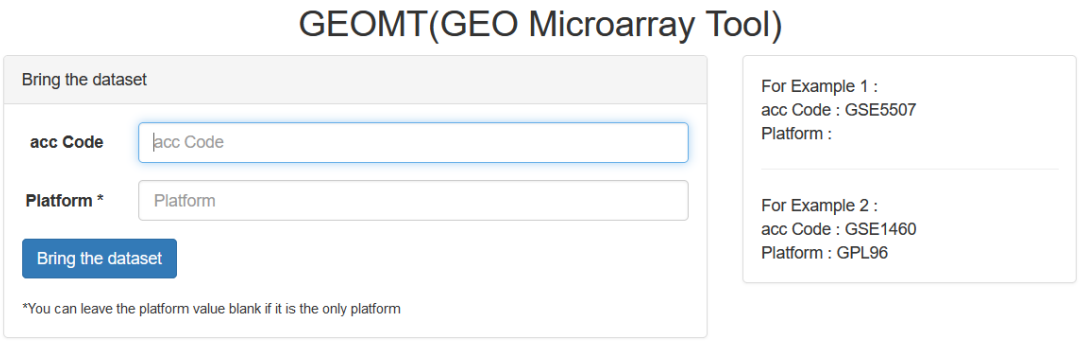
$$Nd = No * \frac{Ortalama}{N_{toplam}} \tag{3.3}$$

Denklem (3.3)'de "No" ham veri ve "Nd" ise normalize edilmiş veri anlamına gelmektedir ve her bir veri için bu formüller uygulanarak verileri normalize edilmiş değerler elde edilmiştir.

4. BULGULAR VE TARTIŞMA

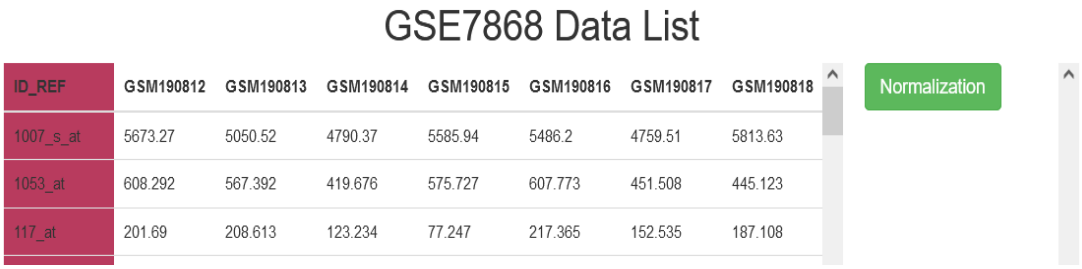
4.1 Programın Kullanılması

Şekil 4.14'de görüldüğü gibi programın ara yüzü bilgisayar ya da biyoinformatik bilgisi sınırlı araştırmacıların kullanabilmesi için oldukça basit bir şekilde tasarlanmıştır. Program çalıştırıldığında kullanıcıdan GEO accession (acc code) ve Platform numaralarının girilmesi istenmektedir. Analiz yapılacak GEO data setinin tek bir platform değerine sahip olması durumunda platform numarası girilmeyebilir ancak ilgili set birden fazla platform numarasına sahipse hangi platformdaki verinin işleneceğinin seçilmesi gerekmektedir. Gerekli veriler girildikten sonra program arka planda GEO sunucusundan veri setini bulup sıkıştırılmış halini programın kurulu olduğu sunucuya indirmekte ve sonrasında da sıkıştırılmış olan dosyayı açıp içindeki verileri kullanıcıya ikinci sayfada göstermektedir.



Şekil 4.14. GEO accession değerinin ve platform değerinin girilmesi

Şekil 4.15'de verilerin gösterildiği ikinci sayfada her bir gene karşılık gelen probların ID_REF numaraları ilk sütundaki satırlar boyunca gösterilmekte ve her bir probun farklı deney setlerindeki değerleri aynı satır boyunca listelenmektedir. Bu sayfada aynı zamanda "Normalization" butonu ile verilerin normalizasyon işlemi gerçekleştirilmekte ve ardından 3. Sayfaya geçerek burada normalize edilmiş veriler gösterilmektedir .



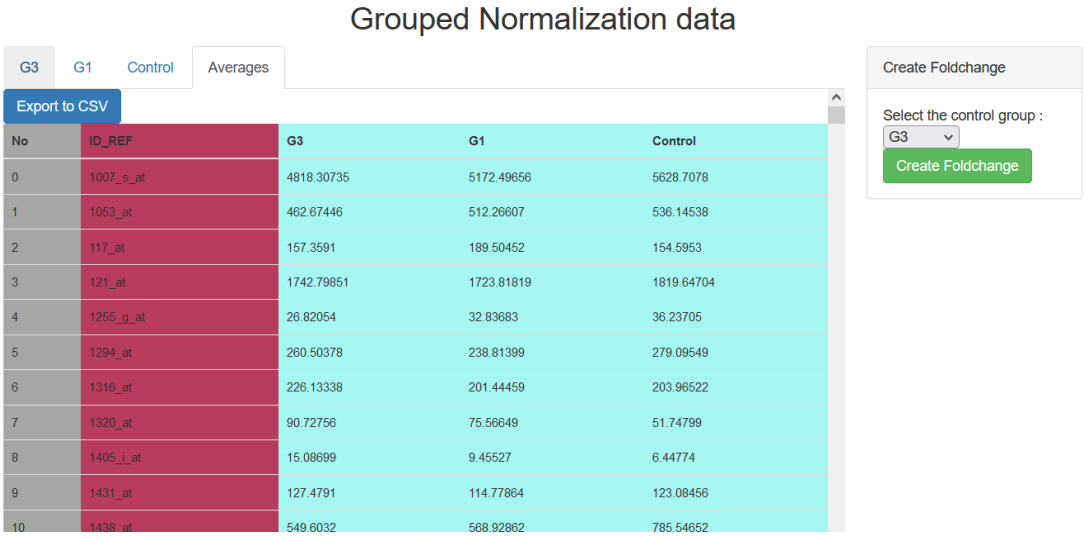
Şekil 4.15 GEOMT Normalizasyon İşlemi

Deneysel çalışmalar, elde edilecek sonuçların güvenirliği açısından genellikle birden fazla tekrarla yapıldığından dolayı analizi yapılacak veri setinde de tekrarlı grupların olması ihtimaline karşı bir opsiyone eklenmiştir. Bu nedenle analiz edilecek verilerde tekrarlı gruplar varsa 3. sayfada tanımlanması gerekmemektedir. Bu amaçla tekrar olan veriler seçilerek grup ismi girilerek gruplandırılır. Bunun dışında, bir gruba yeni bir tekrar veri ekleneceği zaman ilgili verinin üzerine gidilerek işaretlenir ve bu veri "Add" butonu ile gruba eklenebilir. Bütün tekrar verileri gruplandırıldıktan sonra "Define Grup" butonu ile her bir gruba bir isim verilerek tanımlanmaktadır (Şekil 4.16).



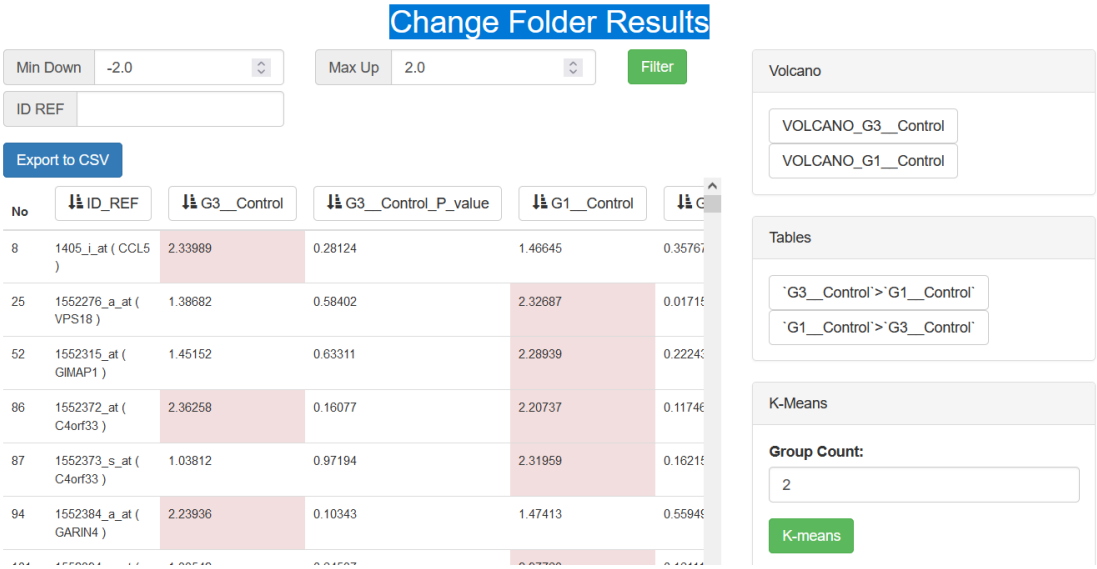
Şekil 4.16 GEOMT Gruplama İşlemi

Şekil 4.17'de yeni açılan sayfada normalize edilmiş her bir tekrar veri sekmeler halinde gruplanmış bir durumda gösterilirken aynı zamanda bu gruptaki normalize verilerin ortalamaları da "Avarage" sekmesi halinde gösterilmektedir. Bu sayfada kontrol grubu Fold change (kat değişim) butonu ile seçilerek her bir test grubundaki verinin kontrol grubuna göre kat değişimi hesaplaması yapılmaktadır.



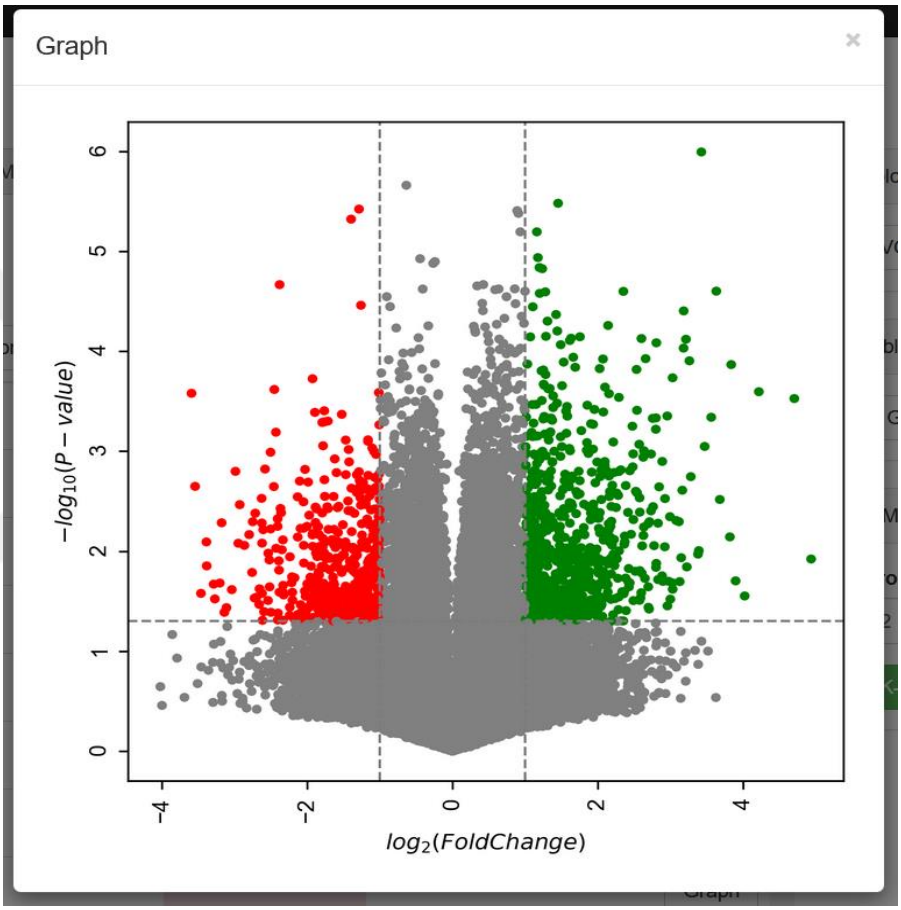
Şekil 4.17 Normalize Edilmiş Gruplar

Şekil 4.18'de bir taraftan her bir prob için kontrol grubuna göre hesaplanan kat değişimi değerleri gösterilirken diğer taraftan da yapılan analizin istatistiksel değerlendirmesi p value değeri olarak listelenmektedir. Bunun dışında, bu sayfada listelenen ve hesaplaması yapılan her bir probun grafiği ayrı ayrı gösterilebildiği gibi Min Down - Max Up değerleri aracılığı ile kat değişim oranları filtrelenerek liste de daraltılabilmektedir.



Şekil 4.18 Kat Değişim Değerlerinin Gösterilmesi

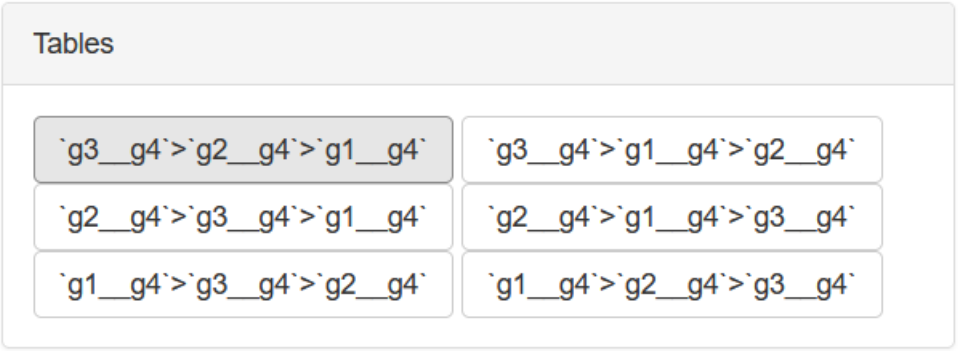
Ayrıca bu sayfadaki listenin ve dolayısıyla verilerin daha kolay incelenebilmesi için her bir kat değişim sütununun başına değerleri büyükten küçüğe veya küçükten büyüğe sıralayabilecek butonlar eklenmiştir. Şekil 4.19'da Veri setlerinin analizinin bir bütün olarak gözlenmesine imkân vermesi için her bir kat değişim analizi için log2(Fold Change) ve log10(P-Value) değerlerini kullanan bir Volcano plot grafiği gösterilmiştir.



Şekil 4.19 Volcano Grafiği

Şekil 4.20'de gösterildiği gibi kat değişim değerlerinin farklı bir yolla analiz edilebilmesi için kat değişim değerlerinin bir birlerinden büyüklüklerini gösterecek şekilde tablolar tasarlandı ve her ihtimalin oluşturula bilinmesi için ihtimallerinin permütasyonu alınarak tablolar oluşturuldu. Örneğin üç farklı deneyden elde edilen kat değişimlerinin karşılaştırılmak istenildiği bir data seti incelediğimizi düşünelim; birinci deneyde en yüksek olan ve üçüncü deneyde en düşük olan kat değişimleri

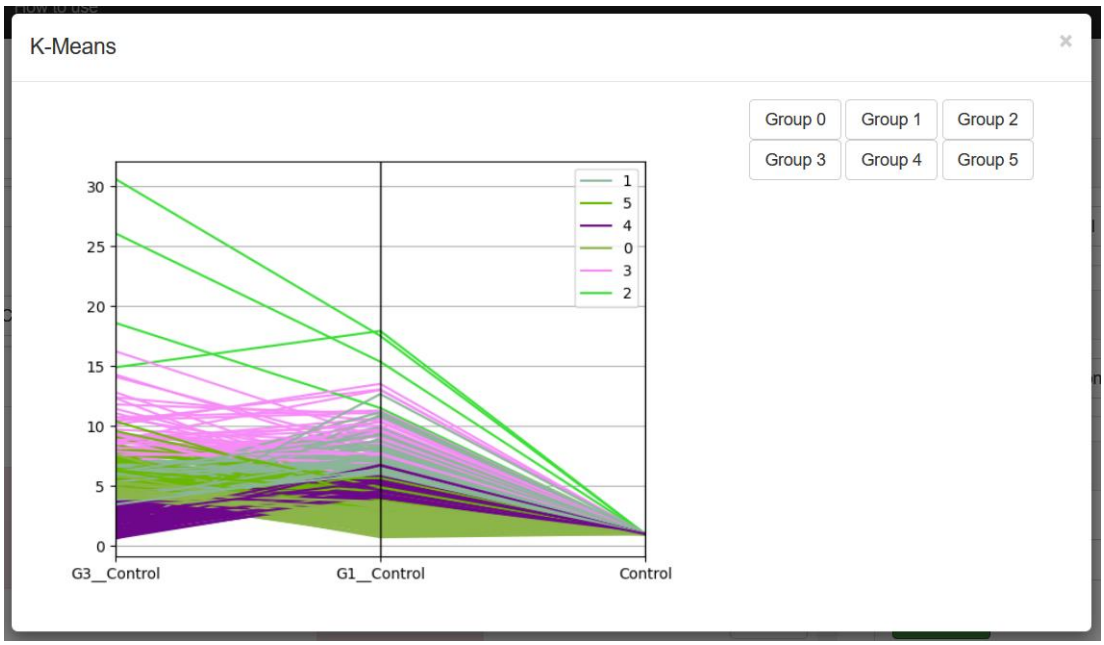
görmek istediğimizde yada birinci deneyde en düşük, ikinci deneyse en yüksek olan kat değişimleri görmek istediğimizde bu modülü kullanırız.



Şekil 4.20 Kat değişim sıralanmış hali ile tabloların oluşturulması

Şekil 4.20'de gösterildiği gibi bir bütün veri setinde benzer şekilde ekspresyon değişimleri gösteren genlerin (probların) gruplandırılması için K-Means algoritması kullanılmış ve bu temelli grafiklerin oluşturulabilmesi sağlanmıştır. Bu grafikler ve alt tablolar oluşturulurken gruplandırma sayılarının da kullanıcının tercihine göre yapılması sağlanmıştır.

.



Şekil 4.21 Kat değişim için K-Means grafiği

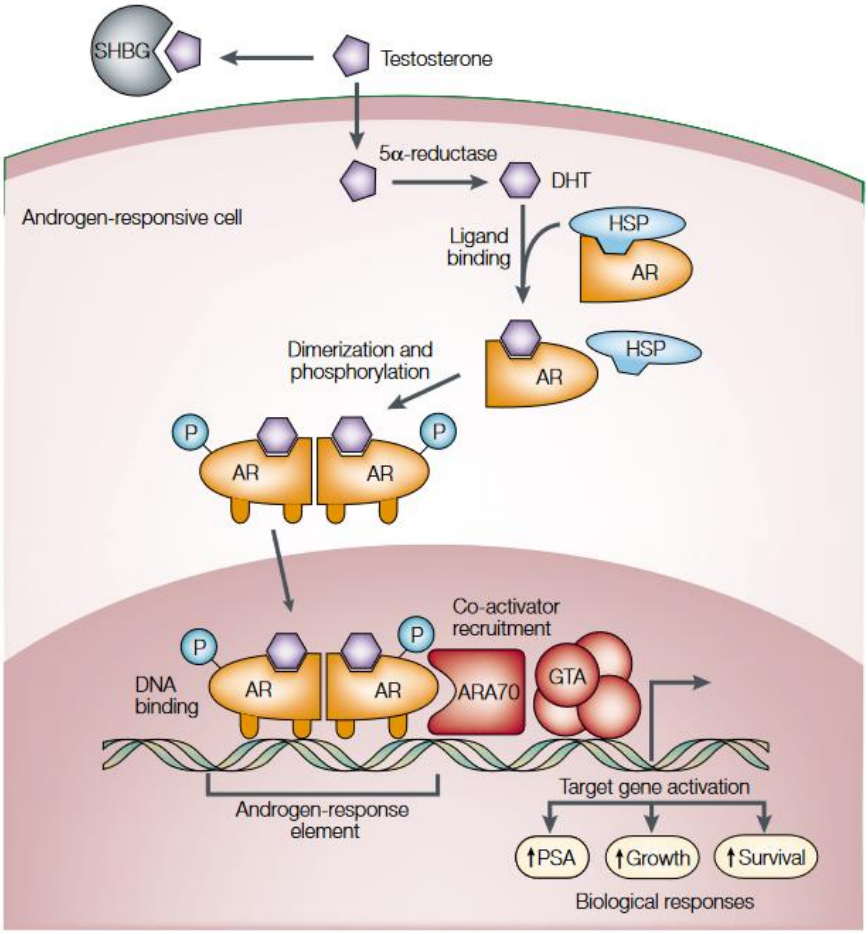
4.2 Programın Test Edilmesi

4.2.1 Veri setinin seçilmesi

Wang Q. ve Li W. tarafından 1 Eylül 2007 tarihinde GEO'da yayınlanıp [23], en son 25 Mart 2019 tarihinde güncellenen GSE7868 [24] serisi programın test işlemi için kullanılmıştır. GSE7868 serisi 9 deney grubundan oluşmaktadır. Bunların 3 tanesi (GSM190812, GSM190815, GSM190818) herhangi bir madde uygulanmamış kontrol grubundan, 3 tanesi (GSM190813, GSM190816, GSM190819) 4 saat androjen verilmiş ve diğer 3 tanesi de (GSM190814, GSM190817, GSM190820) 16 saat androjen verilmiş deney gruplarından oluşmaktadır.

Program analiz yaparken min down ve max up değerleri sırası ile -2 ve 2 olarak belirlendi. Bunun dışında min down ve max up kriterlerini yerine getirse bile her bir prob için ekspresyonal değişim istatistiksel p değeri 0.005'te küçük olanlar seçildi.

Deney setinin açıklaması: Kontrol deney seti olarak androjen sinyalinin seçilmesinin nedeni nispeten iyi karakterize edilmiş olmasıdır. Bu sinyal yolağı geleneksel yolda Androjen Reseptörünü (AR) aktive ederek çalışır. AR'ın gerek normal prostat bezinin gelişimi ve gerekse de prostat kanserinin başlaması ve ilerlemesinde çok önemli görevleri bulunmaktadır. Şekil 4.22'de gösterildiği gibi hücrelerdeki geleneksel androjen-AR mekanizmasında hücre içine giren androjen, sitoplazmada çeşitli proteinler ile bağlanarak inaktif halde tutulan AR'a bağlanarak onu uyarır. Sonuç olarak AR konformasyonel değişim geçirerek onu inaktif halde tutan proteinlerden kurtulur ve çekirdeğe geçerek hedef genlerin promoter bölgeleri üzerindeki özgül DNA motiflerine bağlanarak onların ekspresyonunu kontrol eder.



Şekil 4.22 Hücrelerdeki geleneksel androjen-AR mekanizması [25]

Bu çalışmamızda kullandığımız kontrol setini oluşturan Wang Q. ve Li W. insan prostat kanseri hücrelerinde 21. ve 22. kromozomlarında AR'ın bağlanabileceği 90 adet muhtemel DNA dizisi motifli tanımladılar [23]. DNA dizisi motif verilerini insan prostat kanserlerinden alınan gen ekspresyon profilleri ile entegre ederek, bu motiflerin her birini tanıyan transkripsiyon faktörlerini belirlediler. Zamana bağlı olarak prostat kanserinin büyümesini gen değişimleriyle ilişkilerini inceleye bilinecek data setini oluşturdular.

Programın güvenirliğini test etmek için programın analiz sonrası sonuç olarak verdiği genlerin hangi mekanizma ile ilişkili olduğunun belirlenmesi gereklidir. Bu amaçla gen listesi Enrichr kullanılarak MSigDB Hallmark 2020 veri tabanına karşı taranmıştır. Tarama esnasında ölçüt olarak güvenilirliği maksimuma çıkarmak için p değeri<0.0001 olarak ayarlanmıştır.

Programın güvenilir bir şekilde çalışıp çalışmadığının test edilmesi için GEO veri tabanında bulunan GSE7868 numaralı veri seti kullanılmıştır. Bu veri seti, androjene cevap veren prostat kanseri hücrelerinde zaman bağımlı androjen uyarısı ile ekspresyonu değişen genlerin belirlenmesine izin vermektedir. Literatürde androjen ile ekspresyonu değişen çok sayıda gen hâlihazırda bilindiğinden dolayı programın belirleyeceği genlerin literatürdeki bu genler ile eşleşmesi durumunda programın doğru çalıştığı sonucuna varılacaktır.

Buradaki amacımız expresossion değişen genlerin gerçekten androjen uyarısı ile expressionlarının değişip değişmediğini belirlenmesidir. Bu amaçla 16 saat androjen uygulanmış hücrelerde expressionun değişen genler program aracılığı ile analiz edilmiş ve 164 genin ekspresyonunun değiştiği belirlenmiştir (Tablo 3.1).

Tablo 3.1 16 Saat androjen uygulamasıyla expresiyonu değişen genlerin listesi.

No	Resmi Gen Sembolü	Gen Adı
1	ABCC1	ATP binding cassette subfamily C member 1(ABCC1)
2	ABCC4	ATP binding cassette subfamily C member 4(ABCC4)
3	ABHD2	abhydrolase domain containing 2, acylglycerol lipase(ABHD2)
4	ACSL3	acyl-CoA synthetase long chain family member 3(ACSL3)
5	ADTRP	androgen dependent TFPI regulating protein(ADTRP)
6	AFF3	AF4/FMR2 family member 3(AFF3)
7	AJAP1	adherens junctions associated protein 1(AJAP1)
8	AKNAD1	AKNA domain containing 1(AKNAD1)
9	ALDH1A3	aldehyde dehydrogenase 1 family member A3(ALDH1A3)
10	ANKH	ANKH inorganic pyrophosphate transport regulator(ANKH)
11	ARG2	arginase 2(ARG2)
12	ARID5B	AT-rich interaction domain 5B(ARID5B)
13	ARRDC1	arrestin domain containing 1(ARRDC1)
14	ATAD2	ATPase family AAA domain containing 2(ATAD2)
15	B3GNT7	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 7(B3GNT7)
16	BEND4	BEN domain containing 4(BEND4)
17	BMPR1B	bone morphogenetic protein receptor type 1B(BMPR1B)
18	BRINP2	BMP/retinoic acid inducible neural specific 2(BRINP2)
19	C11orf40	chromosome 11 putative open reading frame 40(C11orf40)
20	C19orf48	chromosome 19 open reading frame 48(C19orf48)
21	C1orf116	chromosome 1 open reading frame 116(C1orf116)
22	C1orf21	chromosome 1 open reading frame 21(C1orf21)
23	CAMKK2	calcium/calmodulin dependent protein kinase kinase 2(CAMKK2)

24	CDC14B	cell division cycle 14B(CDC14B)
25	CDK5R1	cyclin dependent kinase 5 regulatory subunit 1(CDK5R1)
26	CENPN	centromere protein N(CENPN)
27	CHRNA2	cholinergic receptor nicotinic alpha 2 subunit(CHRNA2)
28	CLEC7A	C-type lectin domain containing 7A(CLEC7A)
29	COL11A1	collagen type XI alpha 1 chain(COL11A1)
30	COLEC10	collectin subfamily member 10(COLEC10)
31	CROT	carnitine O-octanoyltransferase(CROT)
32	CSGALNACT1	chondroitin sulfate N-acetylgalactosaminyltransferase 1(CSGALNACT1)
33	CTBP1	C-terminal binding protein 1(CTBP1)
34	CTC1	CST telomere replication complex component 1(CTC1)
35	CYP3A43	cytochrome P450 family 3 subfamily A member 43(CYP3A43)
36	DHCR24	24-dehydrocholesterol reductase(DHCR24)
37	DOCK4-AS1	DOCK4 antisense RNA 1(DOCK4-AS1)
38	DST	dystonin(DST)
39	DUXAP10	double homeobox A pseudogene 10(DUXAP10)
40	EFCAB12	EF-hand calcium binding domain 12(EFCAB12)
41	EFNA5	ephrin A5(EFNA5)
42	ELL2	elongation factor for RNA polymerase II 2(ELL2)
43	ELOVL7	ELOVL fatty acid elongase 7(ELOVL7)
44	ENDOD1	endonuclease domain containing 1(ENDOD1)
45	ENPEP	glutamyl aminopeptidase(ENPEP)
46	ERRFI1	ERBB receptor feedback inhibitor 1(ERRFI1)
47	F5	coagulation factor V(F5)
48	FAM174B	family with sequence similarity 174 member B(FAM174B)
49	FAM66C	family with sequence similarity 66 member C(FAM66C)
50	FAR2P2	fatty acyl-CoA reductase 2 pseudogene 2(FAR2P2)
51	FAR2P3	fatty acyl-CoA reductase 2 pseudogene 3(FAR2P3)
52	FSTL1	follistatin like 1(FSTL1)
53	FZD5	frizzled class receptor 5(FZD5)
54	GLMP	glycosylated lysosomal membrane protein(GLMP)
55	GNMT	glycine N-methyltransferase(GNMT)
56	GPR158	G protein-coupled receptor 158(GPR158)
57	GPR18	G protein-coupled receptor 18(GPR18)
58	GREB1	growth regulating estrogen receptor binding 1(GREB1)
59	GYPA	glycophorin A (MNS blood group)(GYPA)
60	H6PD	hexose-6-phosphate dehydrogenase/glucose 1-dehydrogenase(H6PD)
61	HIF3A	hypoxia inducible factor 3 subunit alpha(HIF3A)
62	HOMER2	homer scaffold protein 2(HOMER2)
63	HPGD	15-hydroxyprostaglandin dehydrogenase(HPGD)
64	HSPB9	heat shock protein family B (small) member 9(HSPB9)
65	HTR6	5-hydroxytryptamine receptor 6(HTR6)
66	IGF1R	insulin like growth factor 1 receptor(IGF1R)
67	INPP4B	inositol polyphosphate-4-phosphatase type II B(INPP4B)
68	JAG1	jagged canonical Notch ligand 1(JAG1)
69	KCNMA1	potassium calcium-activated channel subfamily M alpha

		1(KCNMA1)
70	KCNN2	potassium calcium-activated channel subfamily N member 2(KCNN2)
71	KLF4	Kruppel like factor 4(KLF4)
72	KLHL25	kelch like family member 25(KLHL25)
73	KLHL6	kelch like family member 6(KLHL6)
74	KLK2	kallikrein related peptidase 2(KLK2)
75	KLK3	kallikrein related peptidase 3(KLK3)
76	KLK4	kallikrein related peptidase 4(KLK4)
77	KRT14	keratin 14(KRT14)
78	LAT2	linker for activation of T cells family member 2(LAT2)
79	LIFR	LIF receptor subunit alpha(LIFR)
80	LOC100132078	uncharacterized LOC100132078(LOC100132078)
81	LOC100508046	uncharacterized LOC100508046(LOC100508046)
82	LOC101929572	uncharacterized LOC101929572(LOC101929572)
83	LPAR3	lysophosphatidic acid receptor 3(LPAR3)
84	MBOAT2	membrane bound O-acyltransferase domain containing 2(MBOAT2)
85	MCCC2	methylcrotonyl-CoA carboxylase subunit 2(MCCC2)
86	METAP1D	methionyl aminopeptidase type 1D, mitochondrial(METAP1D)
87	MICAL1	microtubule associated monooxygenase, calponin and LIM domain containing 1(MICAL1)
88	MIDN	midnolin(MIDN)
89	MPC2	mitochondrial pyruvate carrier 2(MPC2)
90	MPHOSPH9	M-phase phosphoprotein 9(MPHOSPH9)
91	MSMB	microseminoprotein beta(MSMB)
92	MTHFR	methylenetetrahydrofolate reductase(MTHFR)
93	NEDD4L	NEDD4 like E3 ubiquitin protein ligase(NEDD4L)
94	NKX1-1	NK1 homeobox 1(NKX1-1)
95	NKX2-1	NK2 homeobox 1(NKX2-1)
96	NKX3-1	NK3 homeobox 1(NKX3-1)
97	NNMT	nicotinamide N-methyltransferase(NNMT)
98	NOX5	NADPH oxidase 5(NOX5)
99	NPPC	natriuretic peptide C(NPPC)
100	OLA1	Obg like ATPase 1(OLA1)
101	OR2B3	olfactory receptor family 2 subfamily B member 3(OR2B3)
102	OR5K1	olfactory receptor family 5 subfamily K member 1(OR5K1)
103	PAAF1	proteasomal ATPase associated factor 1(PAAF1)
104	PALS2	protein associated with LIN7 2, MAGUK p55 family member(PALS2)
105	PCGEM1	PCGEM1 prostate-specific transcript(PCGEM1)
106	PCM1	pericentriolar material 1(PCM1)
107	PLPP1	phospholipid phosphatase 1(PLPP1)
108	PMEPA1	prostate transmembrane protein, androgen induced 1(PMEPA1)
109	PNCK	pregnancy up-regulated nonubiquitous CaM kinase(PNCK)
110	PNLIP	pancreatic lipase(PNLIP)
111	POTEH-AS1	POTEH antisense RNA 1(POTEH-AS1)
112	POTEM	POTE ankyrin domain family member M(POTEM)

113	PPP1R17	protein phosphatase 1 regulatory subunit 17(PPP1R17)
114	PRAG1	PEAK1 related, kinase-activating pseudokinase 1(PRAG1)
115	PRKCA	protein kinase C alpha(PRKCA)
116	PRR5L	proline rich 5 like(PRR5L)
117	PSEN1	presenilin 1(PSEN1)
118	PTGER4	prostaglandin E receptor 4(PTGER4)
119	QSOX2	quiescin sulfhydryl oxidase 2(QSOX2)
120	RAB27A	RAB27A, member RAS oncogene family(RAB27A)
121	RAB3B	RAB3B, member RAS oncogene family(RAB3B)
122	RARS2	arginyl-tRNA synthetase 2, mitochondrial(RARS2)
123	RASEF	RAS and EF-hand domain containing(RASEF)
124	RBAKDN	RBAK downstream neighbor(RBAKDN)
125	RHOU	ras homolog family member U(RHOU)
126	RXFP1	relaxin family peptide receptor 1(RXFP1)
127	SGCD	sarcoglycan delta(SGCD)
128	SLC25A37	solute carrier family 25 member 37(SLC25A37)
129	SLC25A53	solute carrier family 25 member 53(SLC25A53)
130	SLC2A12	solute carrier family 2 member 12(SLC2A12)
131	SLC39A10	solute carrier family 39 member 10(SLC39A10)
132	SLC41A1	solute carrier family 41 member 1(SLC41A1)
133	SLC45A3	solute carrier family 45 member 3(SLC45A3)
134	SLC4A4	solute carrier family 4 member 4(SLC4A4)
135	SNAP25	synaptosome associated protein 25(SNAP25)
136	SORD	sorbitol dehydrogenase(SORD)
137	ST6GALNAC1	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 1(ST6GALNAC1)
138	STK39	serine/threonine kinase 39(STK39)
139	SYCE1L	synaptonemal complex central element protein 1 like(SYCE1L)
140	TARP	TCR gamma alternate reading frame protein(TARP)
141	TEDDM1	transmembrane epididymal protein 1(TEDDM1)
142	TENM2	teneurin transmembrane protein 2(TENM2)
143	THRAP3	thyroid hormone receptor associated protein 3(THRAP3)
144	TMEM79	transmembrane protein 79(TMEM79)
145	TMPRSS2	transmembrane serine protease 2(TMPRSS2)
146	TNFAIP8	TNF alpha induced protein 8(TNFAIP8)
147	TNXB	tenascin XB(TNXB)
148	TPD52	tumor protein D52(TPD52)
149	TRGC2	T cell receptor gamma constant 2(TRGC2)
150	TRGV9	T cell receptor gamma variable 9(TRGV9)
151	TRPM8	transient receptor potential cation channel subfamily M member 8(TRPM8)
152	TTLL10	tubulin tyrosine ligase like 10(TTLL10)
153	VCAN	versican(VCAN)
154	WDR59	WD repeat domain 59(WDR59)
155	WWTR1	WW domain containing transcription regulator 1(WWTR1)
156	YBX2	Y-box binding protein 2(YBX2)
157	ZBTB10	zinc finger and BTB domain containing 10(ZBTB10)
158	ZBTB16	zinc finger and BTB domain containing 16(ZBTB16)

159	ZC3H14	zinc finger CCCH-type containing 14(ZC3H14)
160	ZFP36L1	ZFP36 ring finger protein like 1(ZFP36L1)
161	ZMIZ1	zinc finger MIZ-type containing 1(ZMIZ1)
162	ZNF815P	zinc finger protein 815, pseudogene(ZNF815P)
163	ZNF890P	zinc finger protein 890, pseudogene(ZNF890P)
164	ZP1	zona pellucida glycoprotein 1(ZP1)

2. aşamada ise program tarafından androjen uyarısı ile ekspresyonunun değiştiği belirlenen genlerin dâhil olabileceği mekanizmalar araştırılmıştır. Bu amaçla gen listesi Enrichr [26] kullanılarak MSigDB Hallmark 2020 veri tabanına karşı taranmıştır. Bu araştırma esnasında güvenilirliği maksimuma çıkarmak için p değeri<0.0001 olarak ayarlanmıştır. Enrichr ile yapılan analizler ile elde edilen sonuçlar 28 genin oldukça yüksek bir istatistiksel anlamlılık ile (1.056e-35) androjen cevabı ile ilişkili olduğu gösterilmiştir (Tablo 3.2).

Tablo 3.2 Androjen uyarısıyla ekspresyonunun değiştiği belirlenen genlerin dâhil olabileceği hücresel mekanizmalar

No	Mekanizma Adı	P-value	Genler
1	Androgen Response	1.056e-35	TNFAIP8;HPGD;ABHD2;STK39;TMPRSS2;KLK3; KLK2;CDC14B;CAMKK2;ANKH;ZMIZ1; PMEPA1;NKX3-1;TPD52;ABCC4;HOMER2; ZBTB10;SORD;ARID5B;LIFR;DHCR24;ACSL3;ELL2;INPP4B;ALDH1A3; CENPN;BMPR1B;PLPP1

Tablo 3.3 Gen Açıklamaları

No	Resmi Gen Sembolü	Gen Adı
1	ABCC4	ATP binding cassette subfamily C member 4(ABCC4)
2	ABHD2	abhydrolase domain containing 2, acylglycerol lipase(ABHD2)
3	ACSL3	acyl-CoA synthetase long chain family member 3(ACSL3)
4	ALDH1A3	aldehyde dehydrogenase 1 family member A3(ALDH1A3)
5	ANKH	ANKH inorganic pyrophosphate transport regulator(ANKH)
6	ARID5B	AT-rich interaction domain 5B(ARID5B)
7	BMPR1B	bone morphogenetic protein receptor type 1B(BMPR1B)
8	CAMKK2	calcium/calmodulin dependent protein kinase kinase 2(CAMKK2)
9	CDC14B	cell division cycle 14B(CDC14B)
10	CENPN	centromere protein N(CENPN)
11	DHCR24	24-dehydrocholesterol reductase(DHCR24)
12	ELL2	elongation factor for RNA polymerase II 2(ELL2)
13	HOMER2	homer scaffold protein 2(HOMER2)
14	HPGD	15-hydroxyprostaglandin dehydrogenase(HPGD)

15	INPP4B	inositol polyphosphate-4-phosphatase type II B(INPP4B)
16	KLK2	kallikrein related peptidase 2(KLK2)
17	KLK3	kallikrein related peptidase 3(KLK3)
18	LIFR	LIF receptor subunit alpha(LIFR)
19	NKX3-1	NK3 homeobox 1(NKX3-1)
20	PLPP1	phospholipid phosphatase 1(PLPP1)
21	PMEPA1	prostate transmembrane protein, androgen induced 1(PMEPA1)
22	SORD	sorbitol dehydrogenase(SORD)
23	STK39	serine/threonine kinase 39(STK39)
24	TMPRSS2	transmembrane serine protease 2(TMPRSS2)
25	TNFAIP8	TNF alpha induced protein 8(TNFAIP8)
26	TPD52	tumor protein D52(TPD52)
27	ZBTB10	zinc finger and BTB domain containing 10(ZBTB10)
28	ZMIZ1	zinc finger MIZ-type containing 1(ZMIZ1)

Diğer taraftan, bu tez konusu olan programın belirlediği sonuçlar androjen uyarısının aynı zamanda metastazın en önemli aşamalarından biri olan epitelden mezenşime geçişi de uyardığını göstermektedir. Gerçekten de daha önceki çalışmalar androjen sinyalinin epitelden mezenşime geçişi uyarabildiğini gösterdiğinden dolayı bu sonuç ta programın doğru çalıştığını göstermektedir.

5. SONUÇ

Literatürde GEO'da kayıtlı mikrodizi verilerin analizini yapmak üzere tasarlanan 35'ten fazla yazılım bulunmaktadır [12, 10, 21, 22]. Bunlar dışında çeşitli ticari veya henüz literatür ile desteklenmemiş yazılımların da olduğu göz önüne alındığında, bu verilerin analizini yapabilecek yeni araçlara olan ihtiyaç ortaya çıkmaktadır. Diğer taraftan bu kadar çok yazılımın olması aynı yazılımların tekrardan yazılmış olma şüphesini uyandırmaktadır. Bu şüphe, çoğu yazılımın 5 senedir herhangi bir güncelleme yayımlamamış olması nedeniyle daha da artmaktadır.

Yapılan araştırma da, kurulum gerektiren yazılımların kurulumlarının genel olarak zor olduğunu gözlemlenmiştir. Bunun en büyük sebebi de kurulum için R ya da Python gibi yazılım dillerinin kurulması gerekliliğidir. Maalesef açık kaynak kodlu olup ve kurulum gerektiren tek program MICA'nın da çalışması için R yazılım dillinin kurulması gereklilidir [29].

Diğer taraftan, sunucu tabanlı çalışan yazılımlarda işletim sistemine bağımlılığın ortadan kaldırıldığını gözlemledik. Bu tür yazılımların ön yüzleri evrensel web dilleri olarak kabul edilen HTML, CSS ve JavaScript ile kodlandığı için Windows, Linux ve IOS işletim sistemlerinde çalışabilmekte, verinin işlenmesi ise arka planda çalışan sunucular tarafından yapılmaktadır. Sunucu tabanlı yazılımlar içinde farklı yaklaşım gösteren tek programın GEO2Enrichr olduğunu gözlemledik. GEO2Enrichr, diğer yazılımların aksine tarayıcıya yüklenen eklenti ile çalışmaktadır. Bu özellik, yazılımın performansı veya kullanışlılığı açısından bir avantaj sağlamadığı gibi ileriki zamanlarda güncellenecek olan tarayıcılarla uyumluluk sorunu yaşayabileceğini de düşündürmektedir.

Kriterlerimize uyan yazılımlar içerisinde shinyGEO ve GEO2Enrichr dışındaki tüm yazılımların analiz sonuçlarını görselleştirmede başarılı olduğu gözlemlendi. Bununla beraber Volcano plot grafiğinin sadece BART ve GEO2R yazılımları tarafından oluşturulabildiği görüldü.

Tablo 5.4 Farklı araçlar tarafından sunulan veri analizi yöntemleri

Özellik	MiCA	BART	GEO2R	GEO2Enrichr	shinyGEO
İşletim Sistemi					
Windows	✓	✓	✓	✓	✓
Linux	✗	✓	✓	✓	✓
macOS	✗	✓	✓	✓	✓
Yazılım Dili	R ve C#	R / Python	R	Python	R
Grafikler					
Heatmap	✓	✓	✓	✓	✗
Histogram	✓	✓	✓	✗	✗
Volcano plot	✗	✓	✓	✗	✗
Profile plot	✓	✓	✓	✓	✓
PCA	✓	✓	✓	✓	✓
Veri Girişi					
Dosya Yükleme	✓	✓	✓	✓	✓
GEO dan direk çekme	✓	✓	✓	✗	✗

Yapılan karşılaştırmada en başarılı yazılımların BART ve GEO2R olduğu tablo 5.4'te görülmektedir. Diğer taraftan normalize edilmiş verilerin indirilememesi, diferansiyel değişimlerinde bütün datanın gösterilmemesi ve bu data sorgulamalarının kısıtlı olmasının bütün yazılımların ortak eksikliğidir.

İncelenen bütün programların olumlu ve olumsuz yönleri bir araya getirildiğinde bu alanda yeni yazılımlara ihtiyaç duyulduğu ortadadır. Bunun dışında mikrodizi verilerinin sürekli olarak artması ve mevcut yazılımların bunların analizi noktasındaki eksiklikleri yeni yazılımların tasarlanması ve kullanılması ihtiyacını daha da net olarak ortaya koymaktadır.

GEOMT web tabanlı olarak tasarlandığı için çalışması gerektirecek herhangi bir programın önceden kurulmasına gerek yoktur. Windows, Linux ve IOS işletim

sistemlerinde çalışabilmektedir. Affymetrx çiplerinin sonuçlarının analizini destekleyebilmektedir, açık kaynak kodludur ve oluşturabildiği grafik çeşidi sayısı ve her raporunun indirilebilmesi kendisini diğer yazılımlardan ayırmaktadır.

KAYNAKLAR

[1] J. M. Butler, *Fundamentals of Forensic DNA Typing*. 2010.

[2] U. Pedreira-Segade *et al.*, "How do Nucleotides Adsorb Onto Clays?," doi: 10.3390/life8040059.

[3] et al Alberts B, Johnson A, Lewis J, *Molecular Biology of the Cell. 4th edition.* 2002.

[4] "Transcription." https://www.genome.gov/genetics-glossary/Transcription (accessed Jul. 07, 2022).

[5] Y. Özkan and Ç. S. Erol, *Biyoenformatik DNA Mikrodizi Veri Madenciliği*. 2017.

[6] et al. Lodish H, Berk A, Zipursky SL, *Molecular Cell Biology 5th Ed*. 2000.

[7] S. M. Wang, "Understanding SAGE data," *Trends Genet.*, vol. 23, no. 1, pp. 42–50, 2007, doi: 10.1016/j.tig.2006.11.001.

[8] F. Ozsolak and P. M. Milos, "RNA sequencing: Advances, challenges and opportunities," *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 87–98, 2011, doi: 10.1038/nrg2934.

[9] R. Fajriyah, "Paper review: An overview on microarray technologies," *Bull. Appl. Math. Math. Educ.*, vol. 1, no. 1, p. 21, 2021, doi: 10.12928/bamme.v1i1.3854.

[10] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling," *Am. J. Obstet. Gynecol.*, vol. 195, no. 2, pp. 373–388, 2006, doi: 10.1016/j.ajog.2006.07.001.

[11] G. Agapito, "Computer Tools to Analyze Microarray Data," *Methods Mol. Biol.*, vol. 1986, pp. 267–282, 2019, doi: 10.1007/978-1-4939-9442-7_13.

[12] P. O. Brown and D. Botstein, "Exploring the new world of the genome with dna microarrays," *Nat. Genet.*, vol. 21, no. 1S, p. 37, 1999, doi: 10.1038/4462.

[13] "Affymetrix GeneChip® Microarrays: Student Manual." https://www.affymetrix.com/about_affymetrix/outreach/educator/microarray_curricula.affx#1_1 (accessed Jul. 07, 2022).

[14] A. Koschmieder, K. Zimmermann, S. Trißl, T. Stoltmann, and U. Leser, "Tools for managing and analyzing microarray data," *Brief. Bioinform.*, vol. 13, no. 1, pp. 46–60, 2012, doi: 10.1093/bib/bbr010.

[15] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada, "The Affymetrix GeneChip® Platform: An Overview," *Methods Enzymol.*, vol. 410, no. 06, pp. 3–28, 2006, doi: 10.1016/S0076-6879(06)10001-4.

[16] "National Center for Biotechnology Information." https://www.ncbi.nlm.nih.gov/ (accessed Jul. 07, 2022).

[17] P. Ahluwalia, R. Kolhe, and G. K. Gahlay, "The clinical relevance of gene expression based prognostic signatures in colorectal cancer," *Biochim. Biophys. Acta - Rev. Cancer*, vol. 1875, no. 2, p. 188513, 2021, doi: 10.1016/j.bbcan.2021.188513.

[18] A. Hammad, M. Elshaer, and X. Tang, "Identification of potential biomarkers with colorectal cancer based on bioinformatics analysis and machine learning," *Math. Biosci. Eng.*, vol. 18, no. 6, pp. 8997–9015, 2021, doi: 10.3934/mbe.2021443.

[19] C. Reilly, *Statistical Genomics*. 2020.

[20] D. Xia, F. Ning, and W. He, "Research on Parallel Adaptive Canopy-K-Means Clustering Algorithm for Big Data Mining Based on Cloud Platform," *J. Grid Comput.*, vol. 18, no. 2, pp. 263–273, 2020, doi: 10.1007/s10723-019-09504-z.

[21] W. Yang, H. Long, L. Ma, and H. Sun, "Research on clustering method based on weighted distance density and k-means," *Procedia Comput. Sci.*, vol. 166, pp. 507–511, 2020, doi: 10.1016/j.procs.2020.02.056.

[22] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Comput. Geosci.*, vol. 19, no. 3, pp. 303–342, Mar. 1993, doi: 10.1016/0098-3004(93)90090-R.

[23] E. K. Keeton *et al.*, "Article A Hierarchical Network of Transcription Factors Governs Androgen Receptor-Dependent Prostate Cancer Growth," vol. 1, pp. 380–392, 2007, doi: 10.1016/j.molcel.2007.05.041.

[24] "GEO Accession viewer." https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7868 (accessed Jul. 07, 2022).

[25] B. J. Feldman and D. Feldman, "The development of androgen-independent prostate cancer," *Nat. Rev. Cancer 2001 11*, vol. 1, no. 1, pp. 34–45, 2001, doi: 10.1038/35094009.

[26] M. V. Kuleshov *et al.*, “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,” *Nucleic Acids Res.*, vol. 44, no. 1, pp. W90–W97, Jul. 2016, doi: 10.1093/NAR/GKW377.

[27] W. a Valdivia-granda and C. Dwan, *Chapter 6 MICROARRAY DATA MANAGEMENT*. 2006.

[28] H.-H. Do, T. Kirsten, and E. Rahm, “Comparative Evaluation of Microarray-based Gene Expression Databases,” *Proc. 10. Fachtagung Datenbanksysteme für Business, Technol. und Web (BTW 2003)*, pp. 1–20, 2002, [Online]. Available: papers2://publication/uuid/25112FB0-1551-438D-99CF-7C98A93EF636.

[29] I. Sarfraz, M. Asif, and K. Hijazi, “MiCA: An extended tool for microarray gene expression analysis,” *Comput. Biol. Med.*, vol. 116, p. 103561, 2020, doi: 10.1016/j.compbiomed.2019.103561.

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyad, Ad	Osman VARIŞLI
Web sayfası (Research Gate, Academia, vs.)	

Eğitim Bilgileri

Derece	Kurum	Mezuniyet Yılı
Yüksek Lisans	Dicle Üniversitesi	2022
Lisans	Mersin Üniversitesi	2005
Lise	Derik Lisesi	2000

İş Deneyimi

Dönem (Yıl)	Şirket, Kurum	Görev
2014-	T.C. Sağlık Bakanlığı	Bilgisayar Mühendisi
2012-2014	Türkiye Cumhuriyeti Devlet Demiryolları	Bilgisayar Mühendisi
2010-2012	Sümer Gıda	Bilgisayar Mühendisi
2007-2008	Dmssoft Yazılım A.Ş.	Bilgisayar Mühendisi
2005-2007	ACDC Elektronik Sistemler San. ve Tic. Ltd. Şti.	Bilgisayar Mühendisi

Yabancı Dil

İngilizce

Yayınlar

1.

2.

3.

4.

Özel İlgiler

DİCLE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
TEZ BENZERLİK BİLDİRİMİ FORMU

Öğrencinin Adı, Soyadı	Osman VARIŞLI		
Öğrenci No	19805202		
Ana Bilim Dalı	Elektrik-Elektronik Mühendisliği		
Program Türü	Proje ☐	Yüksek Lisans ☒	Doktora ☐
Tez Danışmanı (Ünvanı, Adı, Soyadı)	Dr.Öğr. Üyesi Cafer Budak		
(Varsa) II. Tez Danışmanı (Ünvanı, Adı, Soyadı)			
Tez Başlığı	Geo Veri Tabanına Kayıtlı Mikrodizi Verilerinin Analizi İçin Yeni Bir Bilgisayar Programının Tasarlanması		
RAPOR BİLGİLERİ			
Raporlama Aşaması	Tez Savunma Sınavı Sonrası		
Sayfa Sayısı	56		
Raporlama Tarihi	18.07.2022		
Benzerlik Oranı (%)	14		

Yukarıda bilgileri verilen tez çalışmamın toplam 56 sayfalık kısmına ilişkin, 18/17/2022 tarihinde şahsım/tez danışmanım tarafından TURNİTİN isimli intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan intihal raporuna göre, tezimin benzerlik oranı %14 olarak tespit edilmiştir.

Uygulanan filtrelemeler:

☐Başlangıç Bölümleri (Kabul ve Onay sayfası, Teşekkür sayfası, Özet/Abstract) hariç

☐Kaynaklar hariç

☐Alıntılar hariç/dâhil

☐Diğer (Açıklayınız)

Tezimin benzerlik oranı, Dicle Üniversitesi Fen Bilimleri Enstitüsü İntihal Raporu Uygulama Esaslarında belirtilen üst sınır benzerlik oranını aşmamaktadır. Benzerlik oranım üst sınır benzerlik oranının altında olsa dahi aksinin tespit edilmesi durumunda her türlü yasal sorumluğu kabul ettiğimi ve hukuki sonuçlarına razı olduğumu bildirir,
gereğini arz ederim.

Öğrencinin Adı, Soyadı: Osman VARIŞLI

Tarih:19/07/2022

İmza:

Danışman Adı, Soyadı: Dr.Öğr. Üyesi Cafer Budak İmza:

Tarih:19/07/2022

Ana Bilim Dalı Başkanı Adı, Soyadı: Doç.Dr. Bilal GÜMÜŞ İmza:

Tarih:19/07/2022
