BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

GENOMİK VERİ TABANLARINDA METİN TABANLI DENEY GERİ GETİRİMİ

HAZIRLAYAN

SELEN BAŞAK

YÜKSEK LİSANS TEZİ

ANKARA – 2021

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

GENOMİK VERİ TABANLARINDA METİN TABANLI DENEY GERİ GETİRİMİ

HAZIRLAYAN

SELEN BAŞAK

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

DR. ÖĞR. ÜYESİ DUYGU DEDE ŞENER

ANKARA – 2021

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı çerçevesinde Selen Başak tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 10 / 08 / 2021

Tez Adı: Genomik Veri Tabanlarında Metin Tabanlı Deney Geri Getirimi

Tez Jüri Üyeleri **İmza**

Prof. Dr. Hasan OĞUL, Çankaya Üniversitesi

Dr. Öğr. Üyesi, Duygu DEDE ŞENER, Başkent Üniversitesi

Dr. Öğr. Üyesi, Tunç AŞUROĞLU, Başkent Üniversitesi

ONAY

Prof. Dr. Faruk ELALDI
Fen Bilimleri Enstitüsü Müdürü
Tarih: … / … / 2021

BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLER ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 24 / 08 / 2021

Öğrencinin Adı, Soyadı : Selen Başak

Öğrencinin Numarası : 21910537

Anabilim Dalı : Bilgisayar Mühendisliği

Programı : Bilgisayar Mühendisliği Tezli Yüksek Lisans

Danışmanın Unvanı/Adı, Soyadı : Dr. Öğr. Üyesi Duygu DEDE ŞENER

Tez Başlığı : Genomik Veri Tabanlarında Metin Tabanlı Deney Geri Getirimi

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 40 sayfalık kısmına ilişkin, 24/08/2021 tarihinde tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 4'dür. Uygulanan filtrelemeler:

1. Kaynakça hariç

2. Alıntılar hariç

3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:

ONAY

Tarih: …. / 08 / 2021

Dr. Öğr. Üyesi Duygu DEDE ŞENER

TEŞEKKÜR

Çalışmam boyunca bana yol gösteren, bilgisini, ilgisini ve desteğini tüm süreç boyunca benimle paylaşan, öğrencisi olmaktan her zaman gurur duyacağım, tez danışmanlığımı üstlenen kıymetli hocam Dr. Öğretim Üyesi Duygu DEDE ŞENER'e,

Çalışma dönemimde bana güç ve cesaret veren, tüm yaşamımda benden sevgi ve ilgilerini esirgemeyen canım annem Emel BAŞAK ve canım babam Güray BAŞAK'a,

Çalışma sürecimde manevi desteklerini hep hissettiğim değerli arkadaşlarım Sena Büşra YENGEÇ TAŞDEMİR ve Araştırma Görevlisi Begüm ERKAL'a

Sonsuz teşekkürlerimi sunarım…

ÖZET

Selen BAŞAK

GENOMİK VERİ TABANLARINDA METİN TABANLI DENEY GERİ GETİRİMİ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

2021

Genomik veriden biyolojiksel bilginin çıkarımı biyoinformatik alanında oldukça önemlidir. Genomik veri deneysel ölçümler, sekans verileri, network olmak üzere farklı veri formatlarında saklanmaktadır. Bu veri yapılarının GEO (Gene Expression Omnibus), ArrayExpress ve GenBank gibi veri tabanlarında hızla artan miktarları ile birlikte doğru ve ilgili veriye ulaşmak önemli bir konu haline gelmiştir. Çalışmada veri tabanlarının deney benzerliklerini bulma yetersizliğinin giderilmesi amacı ile genomik veri tabanlarından ilgili deneylerin geri getirimi için metin tabanlı bir geri getirim alt yapısı geliştirilmesi hedeflenmiştir. Geliştirilen alt yapıda sözcük tabanlı (lexical) ve anlam (semantic) tabanlı benzerlik bulma yöntemleri kullanılıp, yöntem performansları kıyaslanmıştır. Sözcük tabanlı yöntem olarak Jaccard benzerlik metriği kullanılırken, anlam tabanlı benzerlik yöntemleri olarak Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) ve Latent Dirichlet Allocation (LDA) yöntemleri kullanılmıştır. Bildiğimiz kadarıyla, bu çalışma ilk kez anlamsal benzerlik yöntemlerini zaman serisi mikrodizi deneylerinin geri getiriminde kullanan çalışmadır. Sistem performansı, GEO veri tabanından alınan Arabidopsis Thaliana bitkisine ait 120 farklı zaman serisi deneyinin açıklama metinleri ile test edilmiş ve elde edilen deneysel sonuçların biyolojiksel ve istatiksel anlamlılıkları doğrulanmıştır. Sonuçlara göre anlam tabanlı yöntemlerin ilgili deneyleri tespit etmede sözcük ve içerik-tabanlı geri getirim yöntemlerine göre daha başarılı olduğu gözlemlenmiştir. Geliştirilen sistemin mevcut deney geri getirim alt yapılarına göre daha başarılı performans elde etmesi ile gelecek çalışmalara ışık tutması beklenmektedir. Aynı zamanda geliştirilen alt yapı farklı türdeki genomik verilerin benzerliklerini bulmak için kolaylıkla uyarlanabilecek bir sistemdir.

ANAHTAR KELİMELER: LSA, PLSA, LDA, Jaccard, Bilgi Geri Getirimi

ABSTRACT

Selen BAŞAK
TEXT-BASED EXPERIMENT RETRIEVAL IN GENOMIC DATABASES
Başkent University Institute of Science and Engineering
Department of Computer Engineering
2021

Extraction of biological information from genomic data is very important in the field of bioinformatics. Genomic data is stored in different data formats such as experimental measurements, sequence data and network. With the rapid growth of these data structures in databases such as GEO (Gene Expression Omnibus), ArrayExpress and GenBank, reaching accurate and relevant data has become an important issue. In this study, it is aimed to develop a text-based experiment retrieval framework for the retrieval of relevant experiments from genomic databases. In the proposed framework, Jaccard similarity metric was used as a lexical similarity method, Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) methods were used as semantic similarity methods. To the best of our knowledge, this is the first study to use semantic similarity methods in the retrieval of time-series microarray experiments. The system performance was tested with the textual information of 120 different time series experiments of Arabidopsis Thaliana plant obtained from the GEO database. With the biological and statistical significance tests, obtained results were verified. As a result, it was observed that semantic similarity approaches were more successful than text-based and content-based retrieval methods in identifying relevant experiments. It is expected that the developed system will shed light on future studies by achieving a successful performance in retrieving relevant experiments. Moreover, the framework can be easily adapted to different types of genomic data.

KEYWORDS: LSA, PLSA, LDA, Jaccard, Information Retrieval

ÖNSÖZ

Bu çalışmada günümüzde hızla artan veri miktarı nedeni ile ortaya çıkan büyük miktardaki veriyi organize etme, işleme ve ilgili verinin veri tabanından çıkarılması problemleri ele alınmıştır. Bahsedilen sorunların yaşandığı alanlardan biri biyoinformatik alanıdır. Çalışmada soruna genomik veri tabanı GEO'dan (Gene Expression Omnibus) çekilen deneylerin üst verisine Jaccard, LSA (Latent Semantic Analysis), PLSA(Probabilistic Latent Semantic Analysis) ve LDA (Latent Dirichlet Allocation), yöntemleri uygulanarak ile ilgili deneylerin geri getirimini amaçlayan bir yapı ile çözüm aranmıştır. Bu amaçla 2020 yılında Başkent Üniversitesi Öğretim Üyesi Dr. Duygu DEDE ŞENER danışmanlığında çalışmaya başlanmıştır. Çalışma sonuçları aynı veri seti ile yapılan içerik tabanlı deney getirimi yapan başka bir çalışma ile karşılaştırılmış ve en iyi performans LSA yöntemi ile elde edilmiştir. Bununla beraber anlam tabanlı yöntemlerin tamamı içerik ve sözcük tabanlı yöntemlerden başarılı bulunurken sözcük tabanlı yöntem en düşük performansa sahip olan yöntem olarak değerlendirilmiştir.

İÇİNDEKİLER

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

AUC	Area Under curve
BOW	Bag Of Words
CTM	Correlated Topic Model
DNA	Deoksiribonükleik Asit
EM	Exception Maximization
FPR	False Positive Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
ML	Machine Learning
NLP	Natural Language Processing
PLSA	Probabilistic Latent Semantic Analysis
ROC	Receiver Operating Characteristic
TF-IDF	Term Frequency-Inverse Document Frequncy
TPR	True Positive Rate
VSM	Vector Space Model

1. GİRİŞ

Veri tabanlarında saklanan genomik verinin miktarı gün geçtikçe hızla artmakta ve artan bu verinin içinden ilgili bilginin çıkarımı zorlaşmaktadır. Organizmalara ait gen verileri GEO (Gene Expression Omnibus) [1], GenBank [2], ArrayExpress [3] ve Arabidopsis Information Resource (TAIR) [4] gibi genomik veri tabanlarında saklanmaktadır. Bu sebeple araştırmacılar için etkili arama ve geri getirim yöntemleri önemli ihtiyaç haline gelmiştir. Bu çalışmada önerilen deney geri getirim alt yapısı sözcüksel (lexical) ve anlamsal (semantic) yöntemler ile deney benzerliklerini bulmaktadır. Bildiğimiz kadarıyla, anlamsal analiz yöntemleri ile deney geri getirim alt yapısının geliştirilmesi ilk kez bu çalışmada ele alınmıştır. Gene Expression Omnibus'tan (GEO) elde edilmiş Arabidopsis mikrodizi deneylerinin metinsel açıklamaları üzerinde deneysel bir çalışma yapılmıştır.

Bu çalışmada, deney benzerliklerini, sadece deney bilgi metinlerinin sözcüksel benzerliklerine göre değil, anlamsal benzerliklerine göre bulabilen bir deney geri getirim modelinin geliştirilmesi planlanmaktadır. Çalışmanın temel hedefi kullanıcıya deneyin bilgi metnine sözcüksel ve anlamsal olarak en yakın deneylerin geri getirimini sağlamaktır. Önerilen modelde, sözcüksel benzerlik olarak Jaccard benzerliği, anlamsal benzerlik yaklaşımları olarak Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) ve Latent Dirichlet Allocation (LDA) yöntemleri deneylerin metinsel tanımları arasındaki benzerliği bulmak için kullanılmıştır.

Çalışmada mevcut bir problem olan veri tabanlarında deney aramaya farklı bir bakış açısı ve daha önce kullanılmayan yöntemler ile çözüm sunulmuştur. Çalışma sonunda sözcüksel ve anlamsal benzerliğe dayalı deney geri getirimi alt yapısının, içerik tabanlı geliştirilen alt yapı ile karşılaştırması yapılarak alandaki gelecek çalışmalara ışık tutması hedeflenmektedir. Deneysel sonuçlar, alt yapının biyolojik olarak ilgili deneyleri etkili bir şekilde geri getirilebileceğini doğrulamıştır.

Çalışma giriş, yöntemler, sonuçlar ve tartışma olmak üzere dört ana bölüm içermektedir. Giriş bölümünde çalışmanın motivasyonu, amacı, çalışmayla ilgili genel bilgiler ve önceki çalışmalar anlatılmaktadır. Yöntemler bölümünde benzerlik ölçütlerinden, çalışmada kullanılan sözcüksel benzerlik yönteminden, anlamsal benzerlik yöntemlerinden ve yöntemlerin karşılaştırılmasından bahsedilmiştir. Sonuçlar bölümünde ise veri setine,

değerlendirme yöntemlerine ve deneysel sonuçlara değinilmiştir. Son bölüm olan tartışma kısmında ise tartışma ve gelecek çalışma planları anlatılmıştır.

1.1 Motivasyon, Çalışmanın Amacı

Genomik veri tabanlarında bir deney aramak için, kullanıcılar genellikle organizma adı, yazar adı, deney açıklaması, laboratuvar tasarımı ve deney düzeneği hakkında kısa bilgileri içeren, ek açıklama olarak da adlandırılan üst-veri bilgilerini kullanır. Bu veriler sadece birer metin belgesi olarak kabul edilir. Ayrıca mevcut veri tabanları, büyük veri koleksiyonları içindeki metinsel açıklamaların sözcüksel eşleşmesine veya benzerliğine dayalı olarak sadece üst-veri tabanlı arama sağlar. Bununla birlikte, bir veri tabanında arama yaparken, niteliklerin değerlerini birleştirmek için AND ve OR gibi mantıksal operatörler kullanılabilir. Bu tür bir arama kolay gerçekleştirilmesine rağmen, kullanıcılar için sınırlamalara sahiptir. Bu tür aramalar kullanıcıların anlamsal ihtiyaçlarını yapılandırırken, sorgu ile uyumlu olmasını engelleyebilir. Öte yandan, bir deneyi aramak için anahtar kelime tabanlı (keyword-based) arama kullanılabilir. Bu arama yöntemi ile metin açıklamaları kullanılarak sorgu ile uygunluk sağlanabilir, ancak bazen metin açıklamaları deney hakkında yeterli bilgi içermez, deney sonuçları ve deney biyolojik bilgileri metin açıklamalarında bulunmayabilir [5]. Anlam tabanlı arama ise, bu sınırlamaların üstesinden gelmek için geliştirilmiş bir tekniktir. Ayrıca, sözcüksel benzerlik, doğal dildeki kelimelerin çok anlamlılığını ve eş anlamlılığını dikkate almazken, anlamsal benzerlik ile cümlelerdeki kelimeler için eş anlamlılık ve çokanlamlılık dikkate alınarak benzerlik bulunabilir. Bu nedenle anlam tabanlı arama mevcut sınırlamaların üstesinden gelmek için oldukça güçlü bir tekniktir. Önerilen bu çözümün yanı sıra, sözcüksel üst veri tabanlı yöntemlerin sınırlamalarını aşmaya yönelik; veri geri getirimi için sorguya göre örnek geri getirimi yapan çalışmalar vardır. Bu çalışmalar bir deneyi temsil etmek için soyut bir içeriğin oluşturulduğu ve daha sonra elde edilen soyut içerikler arasındaki benzerliği kullanarak bilgi geri getirimi yapılan çalışmalardır (content-based retrieval).

Bu çalışmada, mevcut çalışmalardan ve yaklaşımlardan farklı olarak, sözcüksel ve anlamsal benzerlik yaklaşımlarını kullanarak genomik veri tabanları için metin tabanlı bir deney geri getirim sistemi geliştirilmiştir. GEO veri tabanından elde edilmiş mikrodizi deney koleksiyonunun metinsel bilgileri üzerine deneysel bir çalışma yapılmıştır. Bildiğimiz kadarıyla bu çalışma, verilen veri koleksiyonundan ilgili deneyleri geri getirmek için gen

ekspresyonu deneylerinin metinsel verilerine anlamsal analiz yaklaşımlarını uygulamaya yönelik ilk çalışmadır. Deneysel sonuçlar, önerilen sistemin biyolojik olarak ilgili deneyleri geri getirebileceğini doğrulamaktadır ve ayrıca sistem farklı veri formatlarına kolayca uyarlanabilir bir yapıdadır.

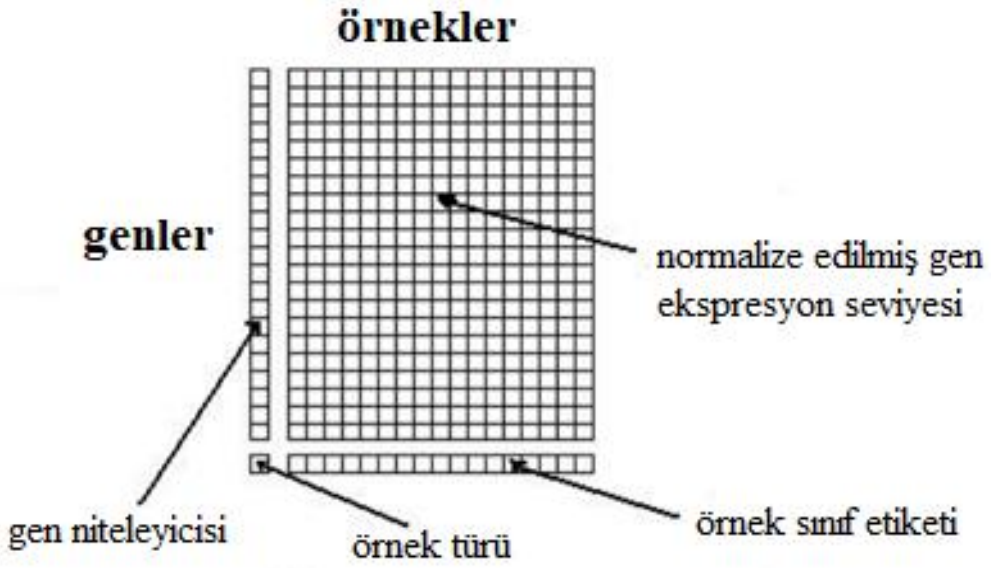
1.2. Genel Bilgiler

Bu kısımda çalışmada kullanılan genomik veri, veri getirimi (information retrieval-IR), deney üst verisi, sorgu deneyi, deney veri tabanı ve deney geri getirimi terimleri anlatılacaktır.

1.2.1. Genomik veri

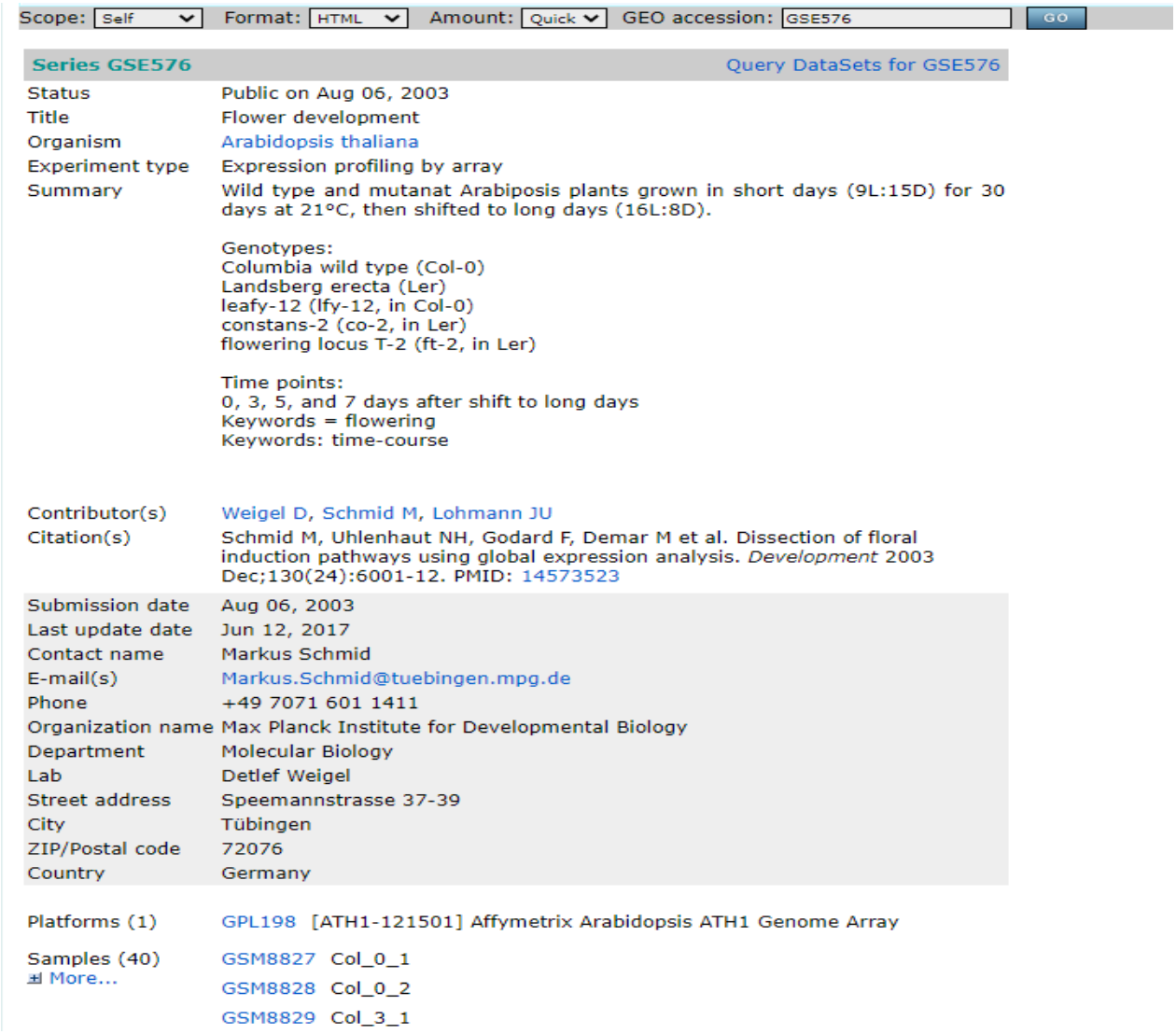
Biyoinformatik yüksek miktarda biyolojik verinin işlenmesi, saklanması, sorgulanması gibi görevlerin gerçekleştirilmesi adına, hesaplamalı biyoloji, istatistik, matematik, moleküler biyoloji ve genetik, bilgisayar bilimlerini de içeren birçok bilimsel alanı birleştiren çok yönlü bir disiplindir [6].

Biyoinformatikte veriler organizmalardan elde edilir. Genom, organizmaların genetik yönergelerinin tutulduğu temel yapı taşlarıdır [7,8]. Genomik ise farklı türlere ait genomların incelenmesine yönelik çalışmalara verilen isimdir [9]. Gen ekspresyonu (gen ifadesi – gene expression) DNA'daki (Deoksiribonükleik Asit) gen bilgilerinin protein vb. bir ürüne çevrilmesi sürecidir [10]. Mikro dizi teknolojisi ile hangi süreçlerde hangi genlerin ifade edildiği anlaşılabilmektedir [11]. Organizmalara ait gen verileri önceden belirtildiği gibi GEO (Gene Expression Omnibus), GenBank, ArrayExpress ve Arabidopsis Information Resource (TAIR) gibi genomik veri tabanlarında saklanmaktadır. Veriler genomik veri tabanlarında diziler, yapılar, ağlar veya deneysel ölçümler gibi çeşitli veri formatlarında olabilmektedir. Şekil.1.1 gen ekspresyonu matrisinin saklanma şeklini temsil etmektedir [12]. Gen ekspresyonu matrisi genin; gen nitelendirici (gene identifier) bilgisini, örneğin türü (sample type), sınıf etiketi (sample class lable) ve normalize edilmiş gen ekspresyonu seviyeleri (normalized gene axpression level) bilgilerini içermektedir.



Şekil 1.1. Gen ekspresyon matrisi [12]

Şekil.1.2 ise deneylerin GEO veri tabanında nasıl tutulduğunu göstermektedir [13]. GEO veri tabanında deney adı ile arama yaparak deneye erişmek mümkündür. Şekilde görüldüğü gibi deneyler GEO veri tabanında kısa deney açıklaması metinleriyle tutulur, deney üst verisi olarak isimlendirilen bu kısa deney metinleri deney adı, organizma adı ve deney özeti gibi açıklamalardan oluşan açıklayıcı kısa yazılardır. Şekilde GSE576 deneyinin; deney durumu, başlığı, deneyde kullanılan organizması, deney özeti ve örneklerini içeren deneye özgü açıklayıcı bilgilerle GEO veri tabanında saklandığı görülmektedir. Kullanıcılar bu bilgiler üzerinden ilgili bilgiye erişmek için aramalarını yapmaktadır.



Scope: Self Format: HTML Amount: Quick GEO accession: GSE576 GO

Series GSE576 Query DataSets for GSE576

Status	Public on Aug 06, 2003
Title	Flower development
Organism	Arabidopsis thaliana
Experiment type	Expression profiling by array
Summary	Wild type and mutanat Arabiposis plants grown in short days (9L:15D) for 30 days at 21°C, then shifted to long days (16L:8D). Genotypes: Columbia wild type (Col-0) Landsberg erecta (Ler) leafy-12 (lfy-12, in Col-0) constans-2 (co-2, in Ler) flowering locus T-2 (ft-2, in Ler) Time points: 0, 3, 5, and 7 days after shift to long days Keywords = flowering Keywords: time-course
Contributor(s)	Weigel D, Schmid M, Lohmann JU
Citation(s)	Schmid M, Uhlenhaut NH, Godard F, Demar M et al. Dissection of floral induction pathways using global expression analysis. *Development* 2003 Dec;130(24):6001-12. PMID: 14573523
Submission date	Aug 06, 2003
Last update date	Jun 12, 2017
Contact name	Markus Schmid
E-mail(s)	Markus.Schmid@tuebingen.mpg.de
Phone	+49 7071 601 1411
Organization name	Max Planck Institute for Developmental Biology
Department	Molecular Biology
Lab	Detlef Weigel
Street address	Speemannstrasse 37-39
City	Tübingen
ZIP/Postal code	72076
Country	Germany
Platforms (1)	GPL198 [ATH1-121501] Affymetrix Arabidopsis ATH1 Genome Array
Samples (40)	GSM8827 Col_0_1 GSM8828 Col_0_2 GSM8829 Col_3_1

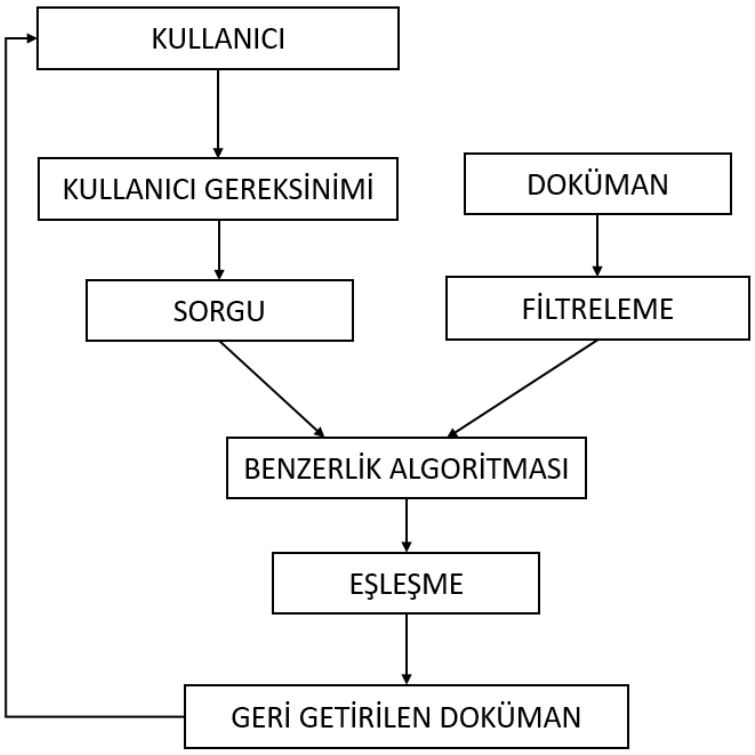
± More...

Şekil 1.2. GEO veri tabanından alınan bir deneyin metin açıklaması görüntüsü [13]

Son yıllarda veri tabanlarında genomik verilerin hızlı artışı ile beraber doğru veriye ulaşma ve bilgi geri getirimi (Information Retrieval-IR) konularında zorluklar yaşanmaktadır. Bu yüzden veri tabanlarında bilgi geri getirimi (IR) konusunda etkin arama yöntemlerine ihtiyaç duyulmaktadır. Araştırmacıların, elde edilen veri kümesinden gerekli bilgileri çıkarması ve hipotezler üretmesi için veri tabanlarında saklanan verilere erişmesi gerekir. Bu sebeple araştırmacılar araştırma yaparken hayatlarını kolaylaştırıp, doğru, ilgili veriye erişimlerini kolaylaştırabilecek daha hızlı ve daha akıllı bilgi geri getirimleri yapabilecek uygulamalara ihtiyaç duymaktadır.

1.2.2. Bilgi geri getirimi (ınformation retrieval - IR)

Bilgi geri getirimi (IR) çeşitli kaynaklardaki işlenmemiş bilgilerden doğru, ilgili olana ulaşmaya yönelik bilgisayar bilimi faaliyetidir [5,14]. Kullanıcının ihtiyaçları doğrultusunda oluşmuş sorgu dokümanının veri geri getirim sistemi tarafından doküman koleksiyonu ile karşılaştırılıp sorgu ile ilgili olan dokümanların belirlenip, kullanıcıya döndürülmesi sürecidir [5,14]. Söz konusu süreç Şekil 1.3'de anlatılmıştır [15]. Şekilde gösterildiği gibi kullanıcı ihtiyaçları doğrultusunda oluşturulan sorgu filtreden geçirilmiş (gerekli ön işlem adımlarından geçirilmiş) bir dokuman ile birlikte benzerlik algoritması tarafından işlenir, eşleşme sonuçlarına göre geri getirilen dokümanlar kullanıcıya döndürülür. Günümüzde verinin hızlı artışı bu süreci oldukça zor bir hale getirmiştir. Bu zorlukla birlikte kullanıcıların hızlı bir şekilde bilgiye ulaşma istek ve gereklilikleri de artmaktadır, bu nedenle IR araştırılmaya ve geliştirilmeye açık bir alandır. IR'da asıl hedeflenen sorgu ile ilgili mümkün olduğunca çok nesne getirirken getirilen nesnelerden mümkün olduğunca azının sorgu ile ilgisiz olmasını sağlamaktır [5,14].



Şekil 1.3. Bilgi geri getirimi [15]

1.2.3. Deney

Çalışmada kullanılan deneyler; GEO veri tabanı üzerinden farklı platformlardan elde edilmiş, Arabidopsis bitkisine ait, zaman serisi (time series) profiline sahip gen mikrodizi (gene mikroarray) deneyleridir. Veri tabanı ve veri setinin detayları ilerleyen bölümlerde anlatılacaktır.

1.2.4. Deneylerin üst verisi

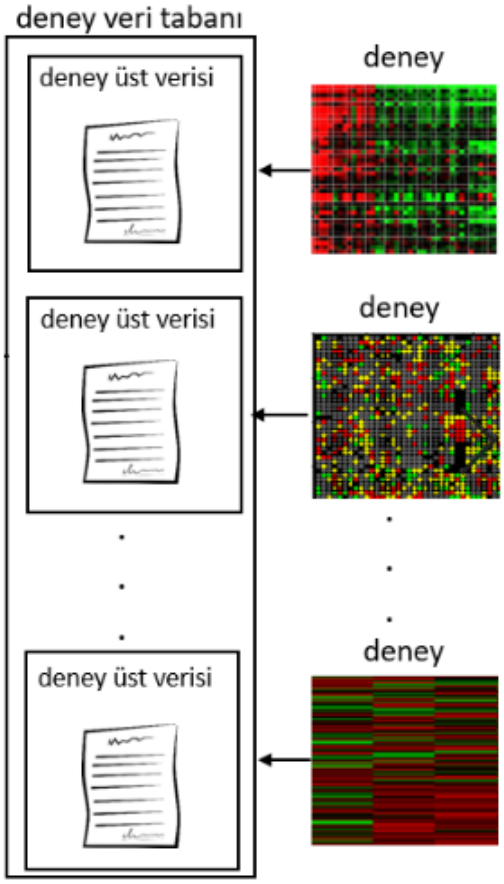
Deney üst verisi (metadatası); deneyde geçen organizma, yazar adı, deney açıklaması laboratuvar tasarımı ve deney düzeneği gibi deneye özgü bilgiler aracılığı ile deneyi açıklamaya yönelik kısa metinsel bilgilerdir.

1.2.5. Sorgu deneyi

Sorgu deneyi için; GEO veri tabanı üzerinden çekilen deneylerin üst verisi elde edilir ve diğer deneylerin üst verileri ile karşılaştırılmak üzere sorgu deneyi olarak belirlenir. Çalışmada tüm deneyler sıra ile sorgu deneyi olarak kullanılmıştır, bu şekilde olası tüm deney eşleşmeleri için benzerlik bulunabilmiştir. Deney üst verisi ise önceden bahsedildiği gibi deneylerin adı, özetleri, tasarımı, varsa uygulanan işlemi vb. deneye özel bilgilerin içerildiği küçük metin parçalarının birleştirilmesi ile oluşturulmuş kısa metin parçalarıdır.

1.2.6. Deney veri tabanı

Deney veri tabanı; çalışmada kullanılmak üzere GEO veri tabanından seçilmiş deneylerin üst verilerinin tutulduğu veri tabanıdır. Şekil 1.4 farklı deneylerin üst verilerinin birleşimi ile oluşturulmuş veri tabanını göstermektedir.



Şekil 1.4. Deney üst verilerinin birleşimi ile deney veri tabanı oluşumu

1.2.7. Deney geri getirimi

Deney geri getirimi sorgu deneyinin deney veri setindeki diğer deneylerle karşılaştırılıp sorgu deneyine en benzer deneyin getirilmesi sürecidir. Sorgu deneyi önceden belirlenmiş veri setindeki her bir deneyle karşılaştırılacak ve karşılaştırma sonucu benzerlik skorları hesaplanarak kullanıcıya döndürülecektir.

1.3. Önceki çalışmalar

Bu kısımda bu çalışmaya temel olmuş iki çalışma, deney geri getirimi ile ilgili yapılmış önceki çalışmalar ve LSA, PLSA ve LDA yöntemleri ile ilgili önceden yapılmış çalışmalar incelenecektir.

Şimdiye kadar, ilgili deneyi geri getirme problemi için pek çok alanda farklı çalışmalar yapılmıştır. Örneğin cDNA mikrodizileri alanında 2009 yılında Caldas et al. [16] tarafından ve 2010 yılında ise Engreitz et al. [17], tarafından çalışmalar gerçekleştirilmiştir, zaman

serisi mikrodizileri alanında Hayran et al. [18] 2014 yılında çalışmalar yapmıştır ve metagenom dizileme örnekleri alanı için ise Seth et al. [19] bir çalışma gerçekleştirmiştir.

Bu çalışmaya temel olan D.D Şener ve H. Oğul [20] tarafından 2015 yılında gerçekleştirilmiş ilk çalışmada, içerik tabanlı arama için GEO veri tabanından 108 zaman serisi deney kullanılmıştır. Çalışmada, her deney için deneyleri temsil eden deney imzaları (fingerprint) çıkarılmıştır. Deney imzaları çıkarılırken gen ifadelerinin zamana bağlı geçişleri kullanan bir metot izlenmiştir. Metot üç farklı tür geçiş içermektedir bunlar: genin ekspresyon seviyesinin yüksekten düşüğe ya da düşükten yükseğe geçtiği birinci tip tek adımlı geçişler, gen ekspresyon değerinin düşükten yükseğe değişebildiği veya tam tersi değişimi gösterip aynı değere dönebildiği iki aşamalı ikinci tip geçişler ve ifade seviyesinin sabit olduğu üçüncü tip geçişlerdir. Genin ait olduğu kategori ise yöntem tarafından hesaplanan bir p değeri ve global bir yanlış keşif oranı (global false discovery rate) ile bulunmuştur. İmza benzerliğini belirlemek için overlap skor kullanılmıştır. Çalışmada temel gerçek bilgisi tüm deneylere Gene Set Enrichment Analysis uygulanması ile elde edilmiştir. Deneylerin gen seti benzerliğini belirlemek için ise Jaccard katsayısı kullanılmıştır. Çalışmada imza çıkarımı için tüm zaman serilerinin kullanılmasının etkisi araştırılmıştır. Geri kazanım performansı, ROC eğrilerinden elde edilen AUC değerleri ile ölçülmüştür. Sonuç olarak tüm zaman serilerinin kullanımının imza çıkarımı için faydalı olduğu gözlemlenmiştir.

D.D. Şener ve H. Oğul [21] tarafından 2016 yılında yapılan ve bu çalışmaya temel olan çalışmaların ikincisinde ise potansiyel olarak faydalı deneylerin nasıl bulunacağı problemini çözmek için zaman serisi deneylerini sorgu olarak alan ve ilgili deneyleri bulan içerik tabanlı bir hesaplama alt yapısı önerilmiştir. Model 120 adet Arabidopsis bitkisine ait zaman serisi özellikli deney kullanılarak test edilmiştir. Çalışmada ilk çalışmada olduğu gibi deney imzaları çıkarılırken gen ifadelerinin zamana bağlı geçişleri kullanan bir metot izlenmiştir. Ancak bu çalışmadaki metot beş farklı geçiş türü içermektedir bunlar: genin ifade seviyesinin sabit olduğu sıfırıncı tip geçişler, genin ekspresyon seviyesinin yüksekten düşüğe geçtiği birinci tip tek adımlı geçişler, genin ekspresyon seviyesinin düşükten yükseğe geçtiği ikinci tip tek adımlı geçişler, gen ekspresyon değerinin yüksekten düşüğe değişip sonrasında aynı değere dönebildiği iki aşamalı üçüncü tip geçişler ve son olarak gen ekspresyon değerinin düşükten yükseğe değişip sonrasında aynı değere dönebildiği iki aşamalı dördüncü tip geçişlerdir. Genin ait olduğu kategorinin bulunması için çalışmada bir adaptif regresyon algoritması kullanılmıştır. Çalışmada imza benzerliği overlap skor ile

ölçülmüştür ve deneylerin gen kümesi tabanlı benzerliği Jaccard katsayısı ile bulunmuştur. Çalışmanın temel gerçek bilgisi tüm deneylere Gene Set Enrichment Analysis uygulanarak elde edilmiştir. Modelin performansı, ROC eğrilerine dayalı AUC verileri ile değerlendirilmiştir. Sonuç olarak içerik benzerliğine göre ilgili deneylerin başarılı bir şekilde bulunabileceği gösterilmiştir. Ayrıca kullanılan imza çıkarma yöntemi ve karşılaştırma yöntemi de başarılı bulunmuştur.

K. Açıcı et al. [22] çalışmalarında mikro-dizi (microarray) verisi tutan veri tabanlarında microRNA deneylerini geri getirmeyi incelemişlerdir. Çalışmada diferansiyel olarak eksprese edilen mikroRNA'ları tanımlamak için normal-düzgün dağılımlı bir karışım modeli tanımlanmış, ardından sıra tabanlı eşik değer kullanılarak ikili gerçek değerli deney imzaları elde edilmiştir. Ayrıca, deneylerden elde edilen deney imzaları arasındaki benzerliği bulmak için etkili bir benzerlik metriği geliştirilmiştir. Çalışma GEO veri tabanından seçilen 135 deneyden oluşan bir veri seti ile gerçekleştirilmiştir. Yöntem performansları AUC skorları ile değerlendirilmiştir. Paired t-test ve Wilcoxon Signed Rank testi ile yapılan istatistiksel doğrulamalar sonucu önerilen her iki yöntemde başarılı bulunmuştur.

Başka bir çalışmada Şener et al. [23], tüm metagenom dizilimi örneği geri getirimi için bir geri getirim sistemi geliştirmişlerdir. Çalışmada, dizi (sequence) örneklerinin sahip olduğu içeriklerinin temsili deney imzalarını elde etmek için farklı imza çıkarımı yaklaşımları kullanılmıştır. Ayrıca önerilen sistemin hesaplama karmaşıklığını azaltmak için öznitelik çıkarma (feature extraction) ve öznitelik seçim (feature selection) yöntemleri uygulanmıştır.

Bunların yanı sıra içerik tabanlı erişim için kullanılabilecek bazı yazılım araçları vardır. CellMontage [24], büyük veri tabanları içinde bir sorgu deneyi aramak için kullanılan ilk yazılım uygulamasıdır. Uygulamada karşılaştırılan deneylerin diferansiyel olarak ifade edilen profilleri arasındaki benzerliği bulmak için Spearman's rank korelasyon katsayısı kullanılmıştır. Bir başka yazılım ProfileChaser [25] ise, Engreitz et al. tarafından önerilmiş ve deneylerin imzalarını oluşturmak için diferansiyel ifadeye dayalı yaygın olarak kullanılan araçlardan biridir. Ayrıca ProfileChaser ile genlerin alt kümesini seçerek boyut küçültme uygulanmıştır. SPIEDw [26] ise, kendi koleksiyonundan ifade değerleriyle birlikte bir gen listesinin sorgu olarak alındığı ve benzer ifade değerlerine sahip ilgili deneylerin geri getirildiği başka bir arama motorudur.

H.S. Le et al. [27] kendi çalışmalarında ilaçların farklı türler üzerinde test edildikten sonra insanlar tarafından kullanıma açılmasından yola çıkarak farklı türler arası (insan - fare) yapılan deneylerin benzerliğine bakmışlardır. Türlerin paylaştıkları ortak genler üzerinden deneylerin benzerliklerini bulmuşlardır. Spearman's rank correlation ile deneylerin benzerliklerini belirlemişlerdir. Yöntem testi için GEO veri tabanından yüksek miktarda insanlar ve farelere ait deneyler kullanılmış ve benzer genleri aktive eden deneyler bulunabilmiştir.

J. Caldas et al. [28] çalışmalarında aynı biyolojik süreçlerin aktive edildiği deneyleri bulmayı hedeflemişlerdir. Bu amaç için eski ve yeni olasılıksal makine öğrenimi (Machine Learning - ML) yöntemlerini her bir deneyde aktive olan biyolojik süreçleri belirleyip aynı biyolojik süreçlerin aktive olduğu deneyleri bulmak ve bulunan sonuçları görselleştirmek için kullanmışlardır.

Bu çalışmaların dışında, çalışmada kullanılan anlamsal benzerlik yöntemleri LSA, PLSA ve LDA farklı alanlardaki farklı çalışmalar için kullanılmıştır. S. Lee et al. [29] çalışmalarında vektör uzayı modeli (vector space model, VSM) üzerine inşa edilmiş LSA, PLSA, LDA ve CTM (Correlated Topic Model) yöntemlerini tartışmışlardır. VSM, vektör uzayındaki metin verilerini temsil eder ve dil vektörlerinin mesafesini ölçer. Çalışmada yöntemlerin deneysel karşılaştırması için; konu algılama ve spam filtreleme örnekleri ele alınmıştır. Konu algılama ile ilgili deney için, Amazon web sitesinden özellikle bir kamera (Canon PowerShot A590IS 8MP Dijital Kamera) hakkında 258 kullanıcı yorumu kullanılmıştır. Deneysel sonuçlara göre; LSA, her konudaki benzersizliğe odaklanırken, PLSA, LDA ve CTM gibi üretken modeller, belgelerdeki genel temaları vurgular. Spam filtreleme deneyleri için 2.893 mailin manuel olarak sınıflandırıldığı, Ling-Spam veri setinden lemm derlemi kullanılmıştır. LSA ile karşılaştırıldığında PLSA daha iyi sonuçlar vermiştir. LDA ise az örneğin alanı bilindiğinden direk uygulanamamış Multi-corpus LDA uygulanmıştır ve PLSA ile çok yakın sonuçlara ulaşılmıştır. CTM LDA'ya benzer şekilde uygulanabilmiş ve LDA ile benzer sonuçlar üretmiştir.

D. Gautam et al. [30] çalışmalarında, sanal stajlar olarak adlandırılan ileri öğrenme sistemlerinde öğrenci tarafından oluşturulan içeriğin değerlendirilmesinde derlem büyüklüğünün ve vektör boyutluluğunun etkisini araştırmışlardır. Yöntemin testinde, sınıflandırıcıları değerlendirmek için kullanılan veri seti sanal stajlardaki öğrenci yanıtlarından oluşmaktadır. Çalışmada veri seti olarak 100 staj defteri girdisi 550 cümleye bölünmüş ve filtrelenmiş, ayrıca TASA veri seti de kullanılmıştır. Sonuçlar alana özel bir

derlemden üretilen LSA alanlarının, not defteri değerlendirme görevini daha büyük bir derlemle üretilen alana kıyasla daha başarılı bir şekilde gerçekleştirebildiğini göstermiştir.

Z. Tong ve H. Zhang [31] çalışmalarında metin verilerinde etkili arama, yönetim ve keşif yapabilmeyi amaçlamışlardır. Çalışmada iki deney tasarlanmıştır; ilk deneyde Wikipedia'dan 200.000 makaleyi aşan bir metin verisine LDA uygulanmış ve başarılı bir şekilde makaleler konulara ayrılabilmiştir. İkinci deneyde ise Twitter'daki tweet metinlerine LDA uygulanarak kullanıcılar üzerinden kapsamlı bir araştırma yapılmıştır ve sonuç olarak twitler başarılı bir şekilde konulara ayrılabilmiştir.

Y. Kalepelli et al. [32] çalışmalarında BBC haber veri setine öncelikle ön işleme adımları uygulanmış ardından LSA ve LDA yöntemlerini uygulanmıştır. Elde edilen sonuçlar geleneksel makine öğrenimi yöntemleri olan KNN ve Naive ile karşılaştırılmış ve hem LDA hem de LSA diğer yöntemlerden başarılı bulunmuştur.

V. Rus et al. [33] çalışmalarında metinler arası anlamsal benzerlik problemini ele almışlardır. Çalışmada Microsoft Research Paraphrase derlemi (MSRP) veri seti olarak kullanılmıştır. Veri setine LSA ve LDA yöntemleri uygulanmış ve sonuçları karşılaştırılmıştır. Sonuçlar elde edilirken önce standart LSA sonra açgözlü (greedy) ve en uygun (optimal) yaklaşımları uygulanmış LSA'da uygulanmıştır. LDA yöntemine ise information Radius (IR), hellinher, manhattan, açgözlü ve en uygun yaklaşımları LDA yöntemine uygulanmıştır. Elde edilen tüm sonuçlar eğitim veri setindeki tüm sınıfların çoğunluk sınıfı ile etiketlendiği bir temel ile elde edilen sonuçlarla karşılaştırılmış ve LDA en iyi precision ve kappa sonuçlarını verdiğinden başarılı bulunmuştur. Ayrıca konu sayısı değişiminin LDA üzerine etkisini araştırmak için yöntem 6, 12, 40 ve 300 konu ile denenmiş ve iyi sonuç 40 konu ile elde edilmiştir.

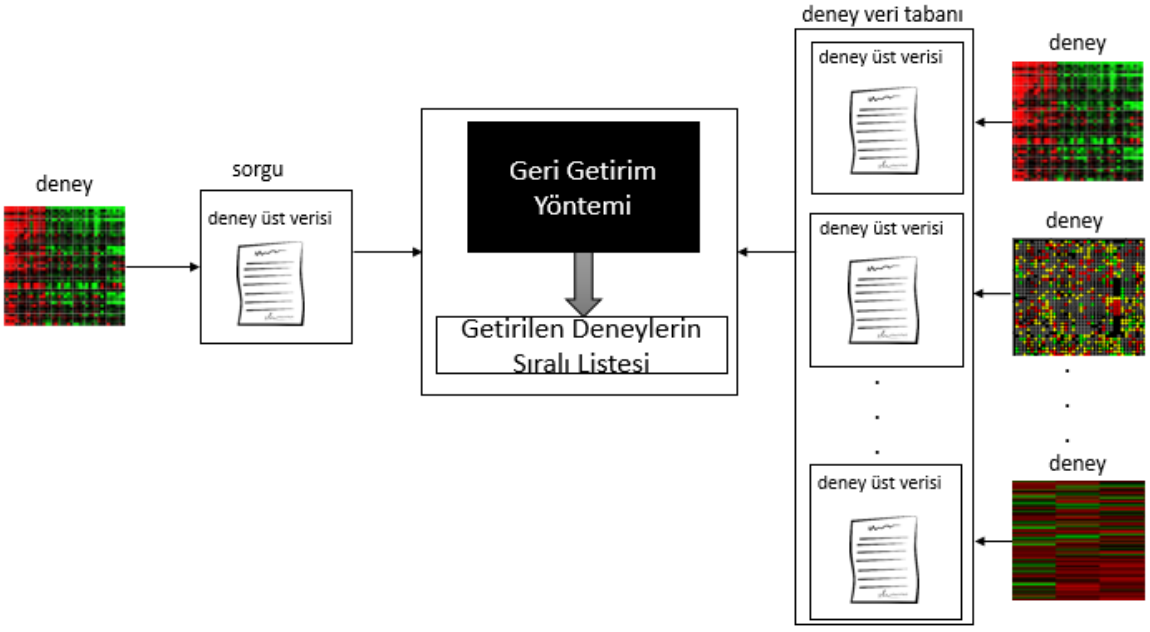
L. Liu et al. [34] tarafından yapılan bir çalışmada hızla artan biyolojik veri ve bu verideki gizli bilgileri çıkarma konusunda yaşanan problemleri biyolojik verilere konu modelleme yöntemlerini uygulayarak çözüm getiren çalışmaları incelemişlerdir. Yapılan incelemeler sonucunda PLSA, LDA ve bunlardan türetilen HLDA (Hierarchical Latent Dirichlet Allocation), PAM (Pachinko Allocation Model), CTM (Correlated Topic Model) gibi diğer konu modellerinin biyolojik veriyi kümeleyebildiği, sınıflandırabildiği, anlaşılır sonuçlar üreterek yorumlamaya yardımcı olduğu ve biyoinformatik için oldukça faydalı olduğu sonucuna varmışlardır.

2. YÖNTEMLER

Bu kısımda çalışma kapsamında geliştirilen alt yapıdan, benzerlik ölçütlerinden çalışmada kullanılan sözcük tabanlı benzerlik yönteminden ve anlam tabanlı benzerlik yöntemlerinden ve son olarak yöntemlerin karşılaştırılmasından bahsedilmektedir.

2.1. Geliştirilen Alt Yapı

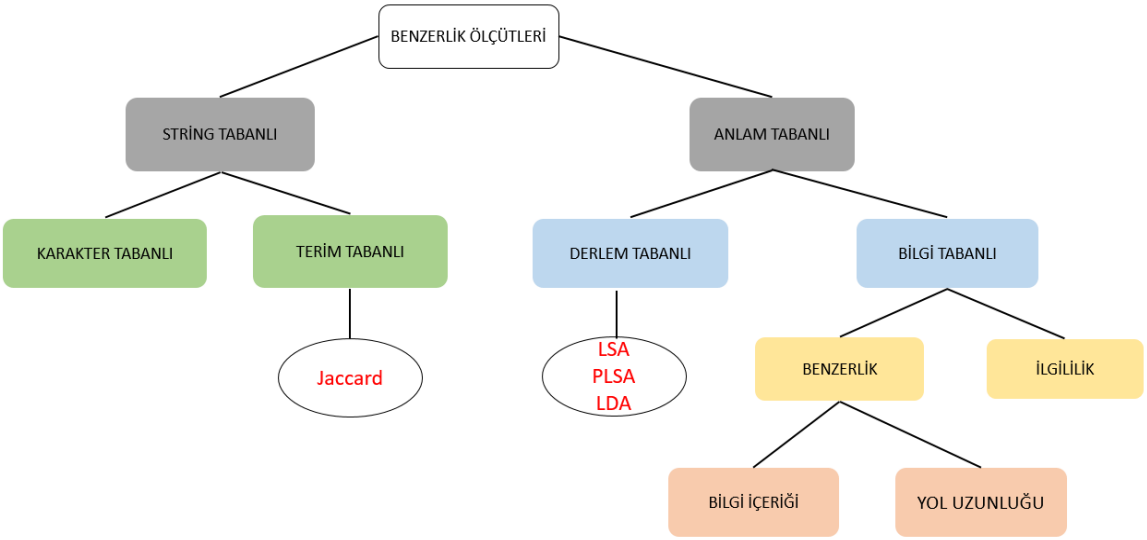
Çalışmada sorgu deneyinden elde edilen sorgu deneyi üst verisi diğer deneylerden elde edilmiş deney üst verileri ile karşılaştırılarak deneyler arası benzerlik oranları bulunmuş ve deney benzerlik oranlarını bulmada kullanılan Jaccard, LSA, PLSA ve LDA yöntemlerinin performansları değerlendirilmiştir. Şekil 2.1 bu süreci yansıtmak üzere geliştirilen geri getirim alt yapısını göstermektedir. Şekilde deneyden elde edilen sorgu ile tüm deneylerin üst verileri ile oluşturmuş veri tabanı geri getirim yönteminin girdileri iken geri getirilen deneylerin sıralı listesi sistemin çıktısıdır.



Şekil 2.1. Geliştirilen geri getirim alt yapısının genel görünümü

2.2. Benzerlik Ölçütleri

Benzerlik ölçütleri metinler arasındaki benzerlikleri bulan yöntemlerdir. Temel olarak benzerlik ölçütleri sözcük tabanlı (string-based) ve anlam tabanlı (semantic-based) olmak üzere iki ana kategoriye ayrılırlar. Sözcük tabanlı yöntemler metin içinde kelimelere ve karakterlerin dizilimine bağlı olarak benzerliği bulurken, anlam tabanlı yöntemler metnin içindeki anlamsal bilgiye dayalı benzerlik bulur [35,36]. Sözcük tabanlı yöntemler kendi içlerinde karakter tabanlı (character-based) ve terim tabanlı (term-based) olarak iki kategoriye ayrılır. Karakter tabanlı yöntemler harflerin dizilimine göre benzerlik bakarken terim tabanlı yöntemler kelimelerin dizilimine bakarak benzerlik bulur. Anlam tabanlı yöntemler ise derlem tabanlı (corpus-based) ve bilgi tabanlı (knowledge-based) olmak üzere ikiye ayrılır. Derlem tabanlı yöntemler büyük metin koleksiyonu aracılığı ile elde edilen bilgilere dayanarak benzerliği bulan yöntemlerdir. Bilgi tabanlı yöntemler ise anlamsal ağlardan elde edilen bilgilere dayanarak benzerlik bulur. Bilgi tabanlı yöntemler kendi içlerinde benzerlik (similarity) ve ilgililik (relatedness) olmak üzere iki kategoriye ayrılır. Bilgi tabanlı yöntemlerin benzerliğine dayanan yöntemleri de bilgi içeriğine veya yol uzunluğuna bağlı olacak şekilde iki kategoriye ayrılır [35,36]. Şekil 2.2 benzerlik ölçütlerini ve alt sınıflarını göstermektedir.



Şekil 2.2. Benzerlik ölçütleri

Şekilden anlaşıldığı gibi tez çalışmasında kullanılan yöntemlerden Jaccard benzerliği sözcük tabanlı yöntemlerden olup terim tabanlı bir yöntemken LSA, PLSA ve LDA yöntemleri

derlem tabanlı yöntemlerdir. Kullanılan yöntemlerden PLSA ve LDA konu modelleme yöntemlerindendir. Konu modelleme mevcut belgelerdeki saklı konuları otomatik olarak bulmayı hedefleyen denetimsiz (unsupervised) bir metin analizi algoritmasıdır. Metnin içinde birlikte ortaya çıkan ilişkili kelimeler kelime gruplarını, kelime gruplarının her biri metnin konularını oluşturur [37]. Konu modelleme bahsedildiği gibi her alanda artan veri miktarı sebebiyle doğru veriye ulaşmanın ve ulaşılan verinin işlenmesinin zorlaşması problemine çözüm bulmak amacıyla ortaya çıkmıştır. 90'lı yıllarda Vector Space Model (VSM) ile başlayan konu modelleme çalışmaları LSA, PLSA, LDA yöntemleri ile devam etmekte olup [38] günümüzde hala geliştirilmeye açık bir alandır.

2.3. Sözcüksel (Lexical) Benzerlik Yöntemi

Çalışmada sözcüksel benzerlik yöntemi olarak kullanılan Jaccard benzerliği temelde benzerliği bulunacak iki metinde geçen ortak kelime sayısının (kesişim) toplam kelime sayısına (birleşime) bölümü ile hesaplanan terim tabanlı bir metinsel benzerlik bulma yöntemidir [39]. A ve B olarak verilen dokümanlar arası Jaccard benzerliği Formül (2.1) ile matematiksel olarak gösterilmiştir. Yöntemde benzerlik puanı 0 ile 1 arasında bir değer alırken, 0 eşleşme olmadığını temsil eder 1, karşılaştırılan dokümanlar arasında mükemmel eşleşme olduğunu gösterir.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \qquad (2.1)$$

2.4. Anlamsal (Semantik) Benzerlik Yöntemleri

Bu bölümde çalışmada kullanılan anlam tabanlı yöntemler olan LSA, PLSA ve LDA yöntemlerinin detayları anlatılacaktır.

2.4.1. Gizli Anlam Analizi (Latent Semantic Analysis - LSA)

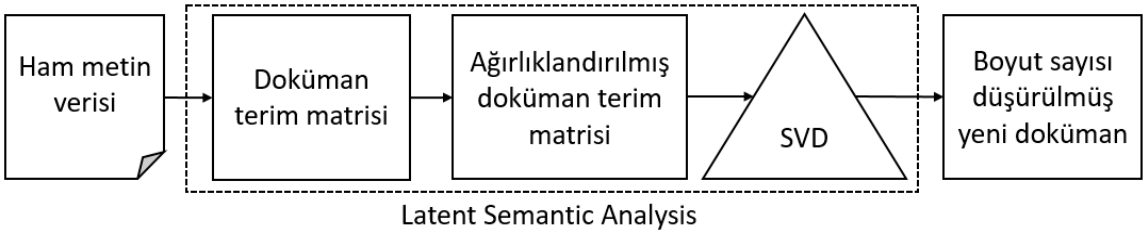
LSA Landauer et al. [40] tarafından tanıtılan derlem tabanlı bir yöntem olup makine öğrenimi alanında eş seslilik ve eş anlamlılık sorunlarının çözümü için bilinen yöntemlerdendir. LSA yöntemi tam otomatik matematik ve istatistiğe dayanan kelimelerin anlamsal ilişkilerini bulmayı hedefleyen bir yöntem olarak tanımlanır ve metinlerin vektör temelli temsilleri aracılığı ile anlamsal bir alan oluşturur [40].

LSA yöntemi girdi olarak işlenmemiş metin verilerini alır bu metin verisi cümle paragraf ya da daha büyük boyutlu olabilir, metinde geçen kelimelerinin sırasının bir önemi olmadığından girdi dokümanı kelime torbası (Bag Of Words - BOW) olarak düşünülebilinir [40]. LSA yönteminde Şekil 2.3.'de görülebileceği gibi girdi metninden bir doküman-terim matrisi (document-term matrix) oluşturulur [41]. Söz konusu matris dokümanda hangi kelimenin kaç defa geçtiğini gösteren bir frekans matrisi olup aynı zamanda girdi verisinin vektör uzay modeli (vector space model - VSM) gösterimidir. Ardından bu frekans matrisi ağırlıklandırılır ve tekil değer ayrışımı (Singular Value Decomposition - SVD) ile matrisin boyutsallığı düşürülür bu şekilde dokümanın düşük boyutlu vektörlerle temsili sağlanmış olur [42]. Ağırlıklandırma yapılırken öncelikle tüm kelimelerin frekanslarının log değerleri bulunur ardından her bir kelime için bilgi teorik ölçütü olan entropy değerleri satırdaki tüm veriler üzerinden ayrı ayrı hesaplanır ve sonuç frekans değerine bölünür. Bu şekilde formül 2.2'de ki gibi metindeki her bir kelimenin ağırlığı belirlenmiş olur [40].

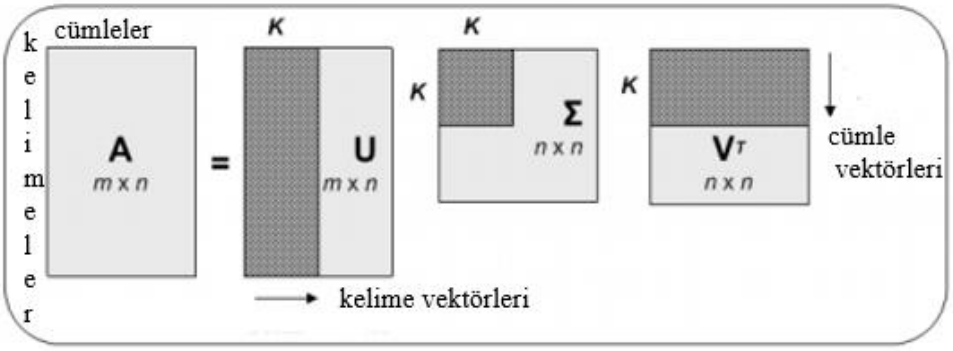
$$W = \frac{f}{-\Sigma\, p\, log\, p} \qquad (2.2)$$

SVD eigen değer ve eigen vektörlere dayanan tamamen matematiksel bir faktör analizi işlemidir. SVD sonucu matris denklem 2.3 ve Şekil.2.4'de gösterildiği gibi üç ayrı matrise ayrılır [42]. Formüldeki üç matrisin ($U * \Sigma * V^T$) çarpımı ile başlangıçtaki matris A tekrar elde edilir.

$$A = U * \Sigma * V^T \qquad (2.3)$$



Şekil 2.3. LSA yöntemi genel yapısı [41]



Şekil.2.4. SVD [42]

Denklem 2.3.'de U matrisinin sütunları sol tekil vektörlerdir, V matrisi orthonormal matristir sütunları sağ tekil vektörlerdir ve Σ matrisi ise tekil değerleri azalan değerlere göre tutan matristir [42]. Şekil.2.4'de görüldüğü gibi A matrisi kelimelerin cümleler içinde geçme sıklığını gösteren frekans matrisiyken SVD ile 3 ayrı matrise ayrılır; U matrisi kelime vektörlerini tutan matris, Σ tekil değerleri tutan matrisken V matrisi cümle vektörlerini göstermektedir [42]. SVD sonucu K adet tekil değer gizli uzayda dokümanların koordinatlarını belirlemek için kullanılır, söz konusu K değeri aynı zamanda matrisin düşürülmüş boyut değeridir. Benzerlik ise bu vektörlerin arasındaki açının kosinüs değeri ile hesaplanabilir [40].

2.4.2. PLSA ve LDA yöntemleri için bazı kavramlar

PLSA ve LDA yöntemlerinde geçen kavramların anlaşılması yöntemlerin anlaşılması için oldukça önemlidir bu sebeple bu kısımda bu kavramlar açıklanacaktır.

- Derlem (Corpus): Dokümanların oluşturduğu koleksiyondur. N dokümanın birleşimi ile oluşur,
- Doküman (Belge - D): Kelimelerin sıralanması ile oluşan yapıdır.; $\mathbf{w} = (w_1, w_2 \ldots., w_K)$ ile gösterilen K kelimenin birleşimidir,
- N: doküman sayısıdır ve $d \in D = \{d_1, d_2, .., d_N\}$
- Konu (Topic - Z): Gizli değişkenlerdir.
- K: Konu sayısıdır ve $z \in Z = \{z_1, z_2, .., z_K\}$
- Kelime (Word - W): Veri setinin en küçük birimidir. Vektörlerle temsil edilir; $\{1, \ldots, V\}$, sözlükteki v'inci kelime V vektörü w ile $u \neq v$ olduğunda $w^v = 1$ ve

$w^u = 0$ şeklinde temsil edilir

- M: kelime sayısıdır ve $w \in W = \{w_1, w_2, .., w_M\}$

2.4.3. Olasılıksal Gizli Anlam Analizi (Probabilistic Latent Semantic Analysis-PLSA)

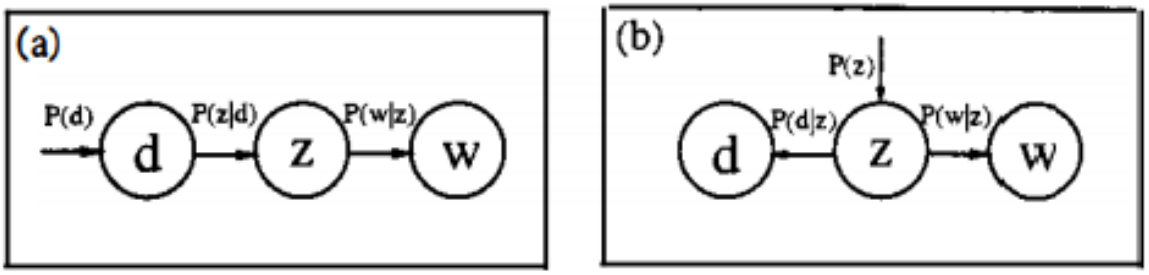
Olasılığa dayalı semantik indeksleme (Probabilistic Latent Semantic ındexing - PLSI) olarak da bilinen olasılıksal semantik analiz (PLSA), konu modeli kategorisinden bir tekniktir. Model 1999 yılında Hofmann [43] tarafından tanıtılmış ve başlangıçta metin tabanlı uygulamalar için kullanılmış ve geliştirilmiştir. Doğrusal cebir tekniklerinden türetilen ve kelime tekrarı verisini kullanan LSA'nın aksine, PLSA karışım model ayrışımına (mixture model decomposition) dayanır. PLSA standart LSA'nın aksine olasılıksal varyantı sağlam iyi istatistiksel temele sahip bir yöntemdir. PLSA'nın temel amacı, farklı kelime kullanım bağlamlarını tanımlamak ve ayırt etmektir [43]. PLSA yönteminde de tıpkı LSA yönteminde olduğu gibi kelime dizilim sırası önemsiz olduğundan girdi dokümanı kelime torbası (BOW) olarak düşünülebilinir. PLSA ve LSA arasındaki en önemli fark, uygun değer ayrıştırma / yaklaşımı belirlemek için kullanılan amaç fonksiyonudur. Bu fonksiyon LSA yönteminde örtük toplamsal Gauss gürültüsü (implicit additive Gaussian noise) varsayımı olan Frobenius normken (L2 norm), PLSA yönteminde Kullback-Leibler'dır [43]. PLSA istatistiksel bir model olan aspect model ile başlar. Aspect model gözlemlenmemiş bir sınıf değişkeni ile birlikte görünen değişkenleri ilişkilendiren bir gizli değişken modelidir [43]. Yöntem verilerin uygun üretken sürecini tanımlar.

PLSA üretken süreci adımları [44]:

1. P(d) olasılıklı D dokümanı seçilir.
2. Dokümandaki tüm W kelimeleri için $P(z|d_n)$ olasılıklı Z_i konusu ilgili dokümandaki koşullu multinominalden seçilir.
3. Son olarak önceden seçilmiş $P(w|z_i)$ olasılıklı Z konusundan W_i kelimesi seçilir.

$$P(d,w) = P(d)P(w|d),\ P(w|d) = \sum_{\mathbf{z}\in\mathbf{Z}} \mathrm{P(w \mid z)\ P(z \mid d)} \qquad (2.4)$$

$$P(d,w) = \sum_{z\in Z} P(z)P(d|z)P(w|z) \qquad (2.5)$$



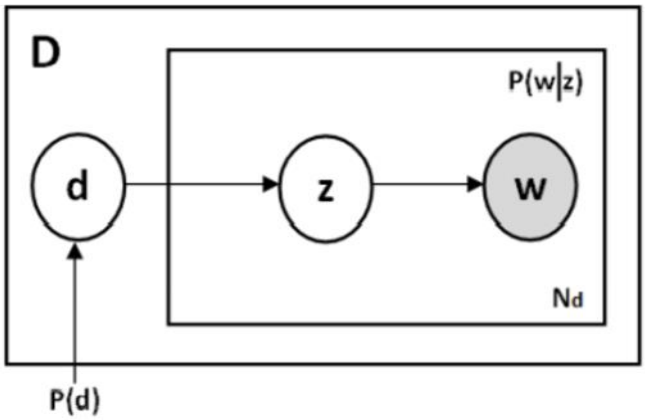
Şekil 2.5. PLSA yöntemi asimetrik ve simetrik temsili [43]

Şekil 2.5 a asimetrik PLSA gösterimini temsil ederken Şekil 2.5 b simetrik PLSA gösterimini temsil etmektedir. Asimetrik gösterim için P(d,w) Formül 2.4'deki gibi, simetrik gösterim için Formül 2.5'de gösterildiği gibi olmak üzere iki ayrı şekilde hesaplanabilir [43].

Gizli değişken modelindeki olasılıklar beklenti maksimizasyonu (Exception Maximization - EM) algoritması ile hesaplanır. Hesaplanan olasılıklar: konu olasılıkları P(z), konulardaki dokümanı olasılıkları P(d | z), konulardaki kelime olasılıkları P(w | z)dır. Bu algoritmanın 2 adımı [43]:

i) gizli değişkenler için olasılıkların son değerlerini belirleme (E),

ii) parametrelerin güncellenmesi (M)



Şekil 2.6. PLSA yöntemi genel yapısı [45]

Şekil 2.6'da PLSA yönteminin genel yapısını göstermektedir [45]. Şekilde görüldüğü gibi PLSA yönteminde d belgesi P(d) olasılıklı, z konusu P(z | d) olasılıklı olacak şekilde seçilir ve konulardaki her w kelimesi P (w | z) olasılığı ile üretilir.

LSA ve PLSA ilişkisini kurmak adına aspect model formül 2.6, formül 2.7 ve formül 2.8'de gösterildiği gibi yeniden tanımlanabilir [43].

$$U = (P(d_i|z_k))_{i,k} \quad (2.6)$$

$$V = (P(w_j|z_k))_{j,k} \quad (2.7)$$

$$\Sigma = diag(P(z_k))_k \quad (2.8)$$

Olasılık modeli ise formül 2.9'daki gibi tanımlanabilir [43].

$$P = U\boldsymbol{\Sigma}V^t \quad (2.9)$$

2.4.4. Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation - LDA)

LDA, derlemdeki dokümanların doğal grubu olan gizli konuları tespit etmeyi amaçlayan üretken olasılıklı konu modelidir. LDA, Blei et al. [46] tarafından tanıtılan denetimsiz bir makine öğrenme algoritmasıdır. LDA yönteminde model üç seviyeli Bayesian bir modeldir. Yöntemde kelime sırası önemsiz olduğundan doküman BOW olarak düşünülebilinir. Yöntemde dokümanlar gizli konular üzerindeki dağılımlar, konular ise kelimeler üzerine dağılımlar olarak düşünülür, dokümanlar konulardan oluşan listelerle ifade edilebilirler [46]. Başka ve en sade değiş ile LDA yönteminde dokümanlar konulardan, konular ise kelimelerden oluşur. LDA yöntemi derlemdeki tüm dokümanların kelimeleri için üretken bir süreç tasarlar.

Bu süreçte [46]:

M, doküman sayısını belirtir,

N, belirli bir belgedeki kelime sayısıdır,

α, dokümanın konu dağılımı,

β, konu başına kelime dağılımı,

θ, boyutu k olan dirichlet rassal değişkeni,

z, konu,

w, kelimedir.

LDA işlem adımları:

1. N, Poisson(ξ) dağılımı ile seçilir.
2. θ, Dir(α) (dirichlet) dağılımı ile seçilir.
3. Tüm N kelimeleri için (w_n) :

(a) konu (z_n) Multinomial(θ) dağılımı ile seçilir.

(b) Kelime (w_n) p(wn |zn,β)'den ((z_n) konusuna bağlı multinomial olasılık) seçilir.

θ'nın olasılık yoğunluğu formül 2.10'da verildiği gibi bulunur [46].

$$P(\theta|\alpha) = \frac{r(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k} r(\alpha_i)}\ \theta_1^{\alpha_1^{-1}} \dots \dots \theta_k^{\alpha_k^{-1}} \quad (2.10)$$

α ve β parametreleri verildiğinde, bir konu karışımının (θ), N'deki konular (z) ve N'deki kelimelerin(w) ortak dağılımı Formül 2.11 ile verilir [46]:

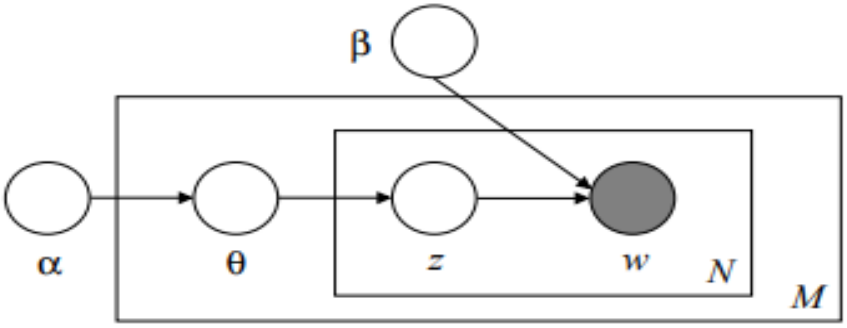
$$\mathrm{p}(\theta, \mathrm{z}, \mathrm{w}|\alpha, \beta) = \mathrm{p}(\theta|\alpha) \prod_{n=1}^{N} \mathrm{p}(z_n|\theta)\mathrm{p}(w_n|z_n, \beta) \quad (2.11)$$

D dokümanının marjinal dağılımı ise formül 2.12 ile elde edilir [46].

$$\mathrm{p}(\mathrm{w}|\alpha, \beta) = \int \mathrm{p}(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} \mathrm{p}(z_n|\theta)\mathrm{p}(w_n|z_n, \beta)\right) \mathrm{d}\theta \quad (2.12)$$

Son olarak derlemin olasılığı tekil marjinal doküman olasılıklarının çarpımı ile Formül 2.13 deki gibi bulunabilir [46].

$$\mathrm{p}(\mathrm{d}|\alpha, \beta) = \prod_{d=1}^{M} \int \mathrm{p}(\theta_d|\alpha) \left(\prod_{n=1}^{\mathrm{N}_d} \sum_{z_{dn}} \mathrm{p}(z_{dn}|\theta_d)\mathrm{p}(w_{dn}|z_{dn}, \beta)\right) \mathrm{d}\theta \quad (2.13)$$



Şekil 2.7. LDA yöntemi genel yapısı [46]

Şekil 2.7. LDA yönteminin grafiksel temsilini göstermektedir. Dış kutu (M) belgeleri temsil ederken, iç kutu (N) bir belge içindeki konuların ve kelimelerin tekrarlı seçimlerini temsil eder. α ve β parametreleri, derlem oluşturulmasında bir kez kullanıldığı düşünülen derlem düzeyi parametreleridir. θ(d) değişkenleri, belgede bir kez kullanılan belge düzeyinde değişkenlerdir. z(dn) ve w(dn) değişkenleri ise tüm belgelerdeki tüm sözcükler için bir kez kullanılan sözcük düzeyinde değişkenleridir [46]. LDA yöntemi ile doküman

benzerliği bulunurken konular dokümanların vektörel temsilleri olduğundan aradaki açının kosinüs benzerliğine bakılabilir.

Çalışmada LDA modeli tanımlanırken doküman olarak deney açıklamaları kullanılır. Konu sayısı T adet olarak belirlenmiş olup kelime ya da terimler W ile temsil edilmiştir. Deney açıklaması $d = \{w_1, w_2 \ldots., w_K\}$ ile verilen K adet kelime içermektedir. Konu ise K kelime üzerine dağılımdır. Derlemde bulunan her bir deney açıklaması formül 2.14'de verildiği şekilde olasılık dağılımı olarak tanımlanır.

$$P(w_i) = \sum_{i=1}^{T} P(w_i|z = z_j)P(z = z_j) \qquad (2.14)$$

Her deney açıklamasındaki W_i kelimesinin olasılığı $P(W_i)$ olarak tanımlanır, mevcut deney analizi için $P(Z = Z_j)$ile temsil edilen Z_j konusundan bir kelime seçilir. Ayrıca, Z_j konusu verilen bir kelimenin örnekleme olasılığı $P(W_i|Z = Z_j)$ olarak tanımlanır. Çalışmada LDA modeli farklı sayıda konu ile uygulanmış ve derlemde ki her bir deney analizi, model tarafından oluşturulan konu dağılımları ile temsil edilmiştir. Çalışmada LDA modeli gerçekleştirilirken öğrenme metodu (learning_method) parametresi olarak Batch seçilmiştir. Bu öğrenme metodunda her iterasyonda yöntem parametreleri güncellenir [47,48].

LDA modelinde konu sayısı bilinmemekte ve konu sayısının belirlenmesinde kesin bir kural bulunmamaktadır, bu nedenle çalışmamızda en uygun konu sayısı deneysel olarak belirlenmiştir.

2.5. Yöntemlerin Karşılaştırılması

Bu kısımda detayları anlatılmış olan derlem tabanlı yöntemlerin karşılaştırılması ve kısa bir değerlendirilmesi yapılmıştır. Tablo 2.1'de LSA, PLSA ve LDA yöntemlerinin özellik ve sınırları verilmiştir. Tablodan anlaşılabileceği üzere yöntemlerin evrimleşme sürecinde bir önceki yöntemin eksikliklerini giderip daha güçlü ve daha doğru sonuçlar üretebilen yöntemler geliştirmek hedeflenmiştir.

Tablo 2.1. Yöntem Karşılaştırılması [29,45]

YÖNTEM	KARAKTERİSTİK	SINIRLAR
LSA	SVD'yi kullanarak tf-idf'nin boyutsallığını azaltır. Kelimelerin eş anlamlılarını bulabilir. Sağlam istatistiksel arka plana dayanmaz.	Konuları ve konu sayısını belirlemek zordur. Olasılık anlamıyla yükleme değerlerini yorumlamak zordur.
PLSA	Karışım bileşenleri, "konuların" temsilleri olan çok terimli rastgele değişkenlerdir. Kelime tek bir konudan üretilir; belgedeki farklı kelimeler farklı konulardan üretilebilir. PLSA çok anlamlılıkla başa çıkabilir.	Belgeler düzeyinde olasılık modeli yoktur.
LDA	Konulardaki kelimeler için çok terimli dağılım ve konular üzerinde Dirichlet dağılımı ile tam üretken model sağlar.	Konular arasındaki ilişkilerin temsili yapılamaz.

3.SONUÇLAR

Bu bölümde kullanılan veri seti, değerlendirme yöntemleri ve deneysel sonuçlar hakkında bilgi verilecektir.

3.1. Veri Seti

Çalışmada genomik veri tabanı GEO'dan seçilmiş deneylerden oluşturulmuş bir veri seti kullanılmıştır. Deneyler veri tabanında deney ile ilgili detaylı bilginin bulunduğu metin içerikleri (bilgi metinleri) ile tutulmaktadır. Veri tabanlarına daha önce yüklenmiş bir deneyi aramak için kullanıcılar genellikle deney ile ilgili organizma, yazar adı, deney açıklaması gibi metinsel üst veri (metadata) bilgilerini kullanmakta, metinsel benzerlikler ile veri tabanlarından bilgi geri getirimi (IR) yapılmaktadır. Çalışmada kullanılan veri seti 120 deney içermektedir, bu çalışmada kullanılan veri seti D.D. Şener ve H. Oğul'un [21] 2016'da yaptıkları çalışma ile aynıdır. Seçilen deneylerde kullanılan organizma uyaranlara hızlı yanıt vermesi sebebiyle Arabidopsis Thaliana bitkisidir. Deneyler zaman serisi profiline sahip; uyarana karşı zamana bağlı olarak bitkinin verdiği yanıtların gözlemlendiği mikro dizi verileridir. Çalışmada deneysel sonuçlar elde edilirken veri seti ilk olarak önişlem aşamasından geçirilmiştir. Önişlem olarak veri seti durma kelimelerinden ve noktalama işaretlerinden arındırılmış, deney metni, okuyanlara anlam ifade etmeyecek kelimelerden, kısaltmalardan ve rakamlardan temizlenmiştir, son olarak veri setindeki kelimeler kök ayırma (stemming) işleminden geçirilmiştir. Kök ayırma işlemi sırasında porter-stemmer kök ayırma algoritması olarak kullanılmıştır.

3.2. Değerlendirme Yöntemleri

Çalışmanın bu kısmında performans değerlendirme yöntemleri, biyolojiksel doğrulama yöntemi ve istatistiksel doğrulama yöntemleri anlatılacaktır.

3.2.1. Performans değerlendirme yöntemleri

Bu çalışmada D. D. Şener ve H. Oğul'un çalışmalarında [21] kullanılan temel gerçek bilgisi (ground truth) kullanılmıştır. Temel gerçek verisini elde etmek için veri setine biyolojiksel doğrulama yöntemlerinden olan Gen Kümesi Zenginleştirme Analizi (Gene Set

Enrichment Analysis-GSEA) uygulanmıştır [49]. GSEA yöntemi ile temel gerçek elde edilme sürecinde girdi; veri setinde kullanılan deneylerin gen ifade verisi ve zaman serisi bilgileridir. Çıktı ise ilgili gen setleri, deney ve deneysel zenginleştirme puanlarıdır. Temel gerçek verisini elde etmek için, gen seti benzerliği Formül 3.1'de verilen Jaccard kat sayısı ile bulunmuştur. Formülde; A ve B sırası ile zenginleştirilmiş gen kümelerini gösteren sorgu deneyi ve karşılaştırma deneyi için vektörlerdir. Yöntemde benzerlik, ortak zenginleştirilmiş gen setlerinin tüm gen setlerine oranı ile bulunmaktadır.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B| - |A \cap B|} \qquad (3.1)$$

Çalışmada karşılaştırılan iki deneyin ilgili olup olmadığına karar vermek için elde edilen gen seti benzerliği puanı 0.35 eşik değeri (threshold) ile değerlendirilmiştir. Bu eşik değer temel gerçek verileri ile elde edilen gen zenginleştirme puanlarının ortalamasına standart sapmanın eklenmesi ile bulunmuştur. Gen seti benzerlik puanı 0.35 eşik değeri üstünde olan deneyler ilgili olarak kabul edilirken eşik değer ve altında gen seti benzerlik puanına sahip olan deneyler ilgisiz deneyler olarak etiketlenmiştir.

3.2.2. Biyolojiksel doğrulama yöntemi

Biyolojik değerlendirme, sonuçların biyolojiksel doğruluğunun ispatı hakkında bilgi veren bir değerlendirmedir. Çalışmada biyolojik değerlendirme olarak Gen Ontology analizi (Gene Ontology-GO) kullanılmıştır [50]. GO, zenginleştirme analizine odaklanan bir yöntemdir. Bu analiz, bazı koşullar altında yukarı veya aşağı regüle (up-regulation ve down-regulation) edilen gen setlerini kullanır. Differential Expression (DE) olarak da isimlendirilen bu işlem çıktı olarak hangi terimlerin fazla veya az temsil edildiğini bulur. Çalışmada GO analizinde kullanılan yukarı veya aşağı regüle edilen genler DEBrowser aracı ile bulunmuştur [51]. Çalışmada DEBrowser aracınnda DE işlemini R dilinin DESeq2 paketi [52] ile |log2FoldChange (log2FC)|> 2 ve $p<0.05$ parametreleri ile gerçekleştirilmiştir. DEBrowser ile elde edilen gen listesi g:Profiler aracı [53] ile işlenmiş ve GO analiz sonuçları elde edilmiştir. GO analizi sonucunda sorgu ve karşılaştırma deneyleri için Biological Processes (BP), Molecular Function (MF) ve Cellular Component (CC)'lara ait GO terimleri elde edilmektedir. Çalışmada BP'lere ait GO terimleri kullanılmış ve ortak GO terimi olup olmadığına bakılarak biyolojik benzerlik bulunmuştur.

3.2.3. İstatistiksel doğrulama yöntemleri

İstatistiksel değerlendirme, sonuçların anlamlılık düzeyi hakkında önemli bir değerlendirmedir. Çalışmada istatistiksel değerlendirme için Wilcoxon Signed Rank Test [54] ve Paired t-test uygulanmıştır [55]. Wilcoxon Signed Rank Test, verilen örneklerin arasındaki farklılığın anlamlılığını değerlendirmek için kullanılan parametrik olmayan bir test iken Paired t-test ilişkili örneklerin ortalamaları arasındaki farkın anlamlılığını değerlendirmede kullanılan parametrik bir testtir [54,55]. Her iki testin sonucunda da bir p değeri üretilir, bu değer ne kadar düşük ise sonuçlar arasındaki fark o kadar anlamlıdır.

3.3.Deneysel Sonuçlar

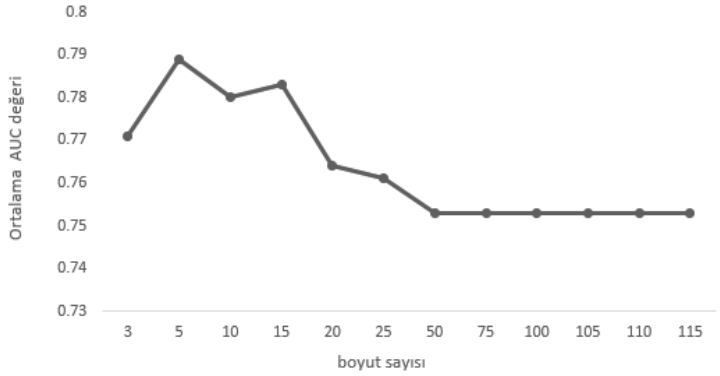
Çalışmada önişlem aşamasından geçirilen veri setindeki tüm deneyler sırası ile sorgu deneyi olmuş ve sorgu olan deney dahil olmak üzere tüm deneylerle karşılaştırılarak ayrı ayrı Jaccard, LSA, PLSA ve LDA yöntemleri kullanılarak olası tüm deney ikililerinin benzerlik oranları elde edilmiştir. Elde edilen benzerlik sonuçları ile yöntemlere özel tüm deneyler için ROC (Receiver Operating Characteristics) eğrileri çizilmiştir. ROC eğrileri sınıflandırıcıları seçmek, görselleştirmek, düzenlemek için kullanılan x ekseninde FPR (False Positive Rate - Yanlış Pozitif Oranı), y ekseninde TPR (True Positive Rate – Doğru Pozitif Oranı) bulunan eğrilerdir. FPR negatif olmasına rağmen yöntem tarafından doğru olarak bulunan örneklerin toplam negatif örneklere oranıyken, TPR doğru olup yöntem tarafından da doğru olarak bulunan örnek sayısının toplam doğru örnek sayısına oranıdır. AUC (Area Under Curve – Eğri Altında Kalan Alan) ise ROC eğrisi altında kalan ve yöntem başarısını gösteren skordur [56]. ROC eğrileri üzerinden AUC skorları hesaplanıp, deneylere özgü AUC değerlerinin ortalaması alınarak yöntem performansını belirten ortalama AUC hesaplanmıştır.

3.3.1. LSA, PLSA ve LDA yöntemleri için ideal boyutun belirlenmesi

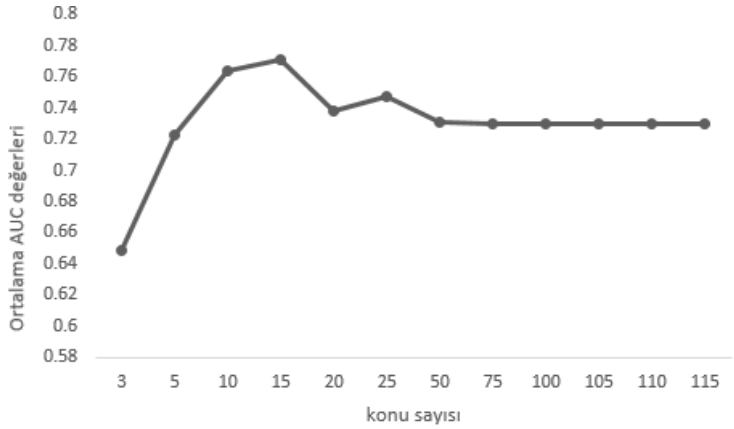
Çalışmada kullanılan anlamsal benzerlik tabanlı yöntemlerden olan LSA için ideal boyut sayısının, PLSA ve LDA yöntemleri için ise en iyi sonuçları veren ideal konu sayısının belirlenmesi önemli bir aşama olarak karşımıza çıkmaktadır. LSA yöntemi anlamsal içerik oluşturmak adına metinlerin vektör tabanlı temsillerini oluşturmaktır. Vektör gösterimi ise, sorgunun verimli olduğu ilgili kelimeyi seçmek için metinler arasındaki benzerliği hesaplar. Bu nedenle doğru boyut sayısı LSA performansı için önemlidir. PLSA ve LDA konu

modelleme yöntemlerinden olduğu için dokümanlar konulardan oluşmaktadır ve bu nedenle yöntem performansları için dokümanların en uygun sayıdaki konulara ayrılması önemlidir.

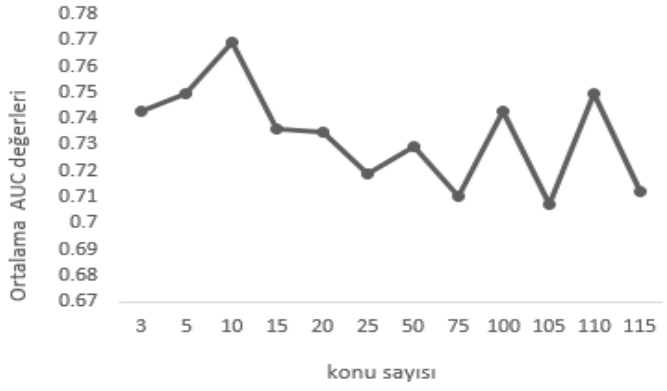
Yöntemlere özgü ideal boyut ve konu sayısının bulunması için LSA, PLSA ve LDA yöntemlerinde 3, 5, 10, 15, 20, 25, 50, 75, 100, 105, 110, 115 boyut ve konu sayısı değerleri ile benzerlikler bulunmuştur. Bulunan tüm benzerlikler için ortalama AUC hesaplanarak en iyi değerleri veren boyut ve konu sayısı belirlenmiştir. Şekil 3.1'deki grafikler yöntemler için en uygun boyut değeri ve konu sayılarının en yüksek AUC değerini sağlayan boyut ve konu sayısı değerleri olduğunu göstermektedir. Şekil 3.1.a. LSA yönteminde verilen boyut değerleri için AUC değerlerinin değişimini göstermekte ve en iyi performansın LSA yöntemi için en uygun boyut değeri olan 5 ile elde edildiğini belirtmektedir. Şekil 3.1.b. PLSA yönteminde verilen konu sayısı değerleri için AUC değerlerinin değişimini göstererek, PLSA yöntemi için en uygun konu sayısı değerinin en iyi performansı vermesi sebebi ile 15 olduğunu anlatmaktadır. Şekil 3.1.c. LDA yönteminde verilen konu sayısı değerleri için AUC değerlerinin değişimini inceleyerek LDA yönteminde en iyi performansın konu sayısı değeri 10 olarak belirlendiğine elde edilebildiğini göstermektedir. Ek1. Sırasıyla LSA, PLSA ve LDA yöntemleri için farklı boyut ve konu sayısı değerlerinin yöntemlere etkisini göstermektedir. Sonuçlar tüm yöntemler için en iyi sonucun bulunan en uygun boyut ve konu sayısı için elde edildiğini belirtir.



a.



b.



c.

Şekil 3.1. LSA, PLSA ve LDA yöntemleri için en uygun boyut ve konu sayılarının ortalama AUC ile ilişkisi

a. LSA yöntemi için boyut ortalama AUC ilişkisi
b. PLSA yöntemi için konu sayısı ortalama AUC ilişkisi
c. LDA yöntemi için konu sayısı ortalama AUC ilişkisi

3.3.2. Yöntem sonuçları

Çalışmada kullanılan metin tabanlı benzerlik ölçütü Jaccard yöntemi için deney tasarımına uygun olarak öncelikle, olası tüm deney ikilileri için benzerlik hesaplanmış ardından bu benzerlikler üzerinden tüm deneyler için ayrı ayrı AUC değeri hesaplanmış bu değerlerin ortalaması alınarak yöntem için ortalama AUC değeri belirlenmiştir. Elde edilen sonuçlar doğrultusunda sözcük tabanlı benzerlik yöntemi olan Jaccard yöntemi %73 başarı oranı ile en düşük performansı vermiştir. Tablo 3.1 verilen AUC eşik değerleri için koşulu sağlayan deney sayılarını göstermektedir. Tablo 3.1 incelendiğinde Jaccard yöntemi için deneylerin %45.83'ünün 0.70 üzeri AUC değerine sahip olduğu gözlenmiştir.

Tablo 3.1. Jaccard yönteminde AUC eşik değerine göre getirilen deney sayısı

Eşik Değeri	Eşik değerinden büyük ortalama AUC skoruna sahip toplam deney sayısı
0.3	120
0.4	118
0.5	103
0.6	72
0.7	55
0.8	46
0.9	43

Çalışmada kullanılan anlamsal benzerlik yöntemlerinden ilki olan LSA yöntemi uygulanırken veri seti detayları önceki bölümlerde anlatılmış önişlem aşamasından geçirilmiştir. LSA yönteminde, deney tasarımında belirtildiği gibi tüm deney ikilileri için benzerlik hesaplanmıştır. Bulunan benzerlikler üzerinden tüm deneylerin AUC değeri hesaplanmıştır. Ortalama AUC değeri ise yöntem performansını değerlendirmede kullanılmıştır. Elde edilen sonuçlar doğrultusunda anlamsal benzerlik yöntemi olan LSA %79 başarı oranı ile en yüksek performansı vererek en başarılı yöntem olarak bulunmuştur. Tablo 3.2 LSA için AUC eşik değerlerine göre getirilen deney sayılarını göstermektedir. Tablo 3.2 incelendiğinde LSA yöntemi için deneylerin %60'ının 0.70 üstünde AUC değerine sahip olduğu gözlenmiştir.

Tablo 3.2. LSA yönteminde AUC eşik değerine göre getirilen deney sayısı

Eşik Değeri	Eşik değerinden büyük ortalama AUC skoruna sahip toplam deney sayısı
0.3	120
0.4	120
0.5	115
0.6	99
0.7	72
0.8	54
0.9	47

Çalışmada kullanılmış bir diğer anlamsal benzerlik yöntemi PLSA ile deneysel sonuçlar elde etmek için veri seti ilk olarak önişlem aşamasından geçirilmiştir. PLSA yönteminde de diğer yöntemlerde olduğu gibi; olası tüm deney ikilileri için benzerlik hesaplanmış ardından bu benzerlikler üzerinden tüm deneyler için ayrı ayrı AUC değeri bulunmuş bu değerlerin ortalaması alınarak yöntem için ortalama AUC değeri hesaplanmıştır. Elde edilen sonuçlar doğrultusunda anlamsal benzerlik yöntemlerinden PLSA'nın 0.771 başarı oranına sahip olduğu belirlenmiştir. PLSA yönteminde AUC eşik değerine göre getirilen deney sayıları ise Tablo 3.3'de verilmiştir. Tablo 3.3'den anlaşılabildiği gibi PLSA yöntemi için deneylerin %54.16'sı 0.70 üstünde AUC değerine sahiptir.

Tablo 3.3. PLSA yönteminde AUC eşik değerine göre getirilen deney sayısı

Eşik Değeri	Eşik değerinden büyük ortalama AUC skoruna sahip toplam deney sayısı
0.3	120
0.4	120
0.5	113
0.6	89
0.7	65
0.8	52
0.9	46

Çalışmada kullanılan son yöntem olan anlamsal benzerlik ölçütü LDA ile deneysel sonuçlar elde etmek için veri setinin önişlemi tamamlanıp ardından olası tüm deney ikilileri için benzerlik hesaplanmıştır. Bulunan benzerlikler üzerinden tüm deneylerin AUC değeri hesaplanmış ve deneylerden elde edilen AUC değerlerinin ortalaması alınarak yöntem performansını veren ortalama AUC değeri hesaplanmıştır. Elde edilen sonuçlar

doğrultusunda anlamsal benzerlik yöntemi olan LDA yöntemi 0.768 performans oranına sahip olmuştur. LDA yöntemi önceki bölümlerde de anlatıldığı gibi dokümanları konuların, konuları kelimelerin karışımı olarak kabul eden bir yöntemdir bu nedenle yöntem uygulanırken doküman – konu ve konu – kelime matrisleri oluşur. Çalışmada kullanılan veri seti için LDA yöntemi ile elde edilen doküman – konu dağılımı EK-2'de verilmiştir. Tabloda sütunlar konuları temsil ederken satırlar dokümanları temsil etmektedir. Tabloda hangi dokumanın hangi konu ile hangi oranda temsil edilebildiği görülmektedir, ayrıca tabloda dokümanların baskın konularına da yer verilmiştir. Tablodan hareketle benzer deneylerin benzer konularla benzer oranlarda temsil edildiği ve baskın konularının aynı ya da benzer konular olduğu açık bir şekilde görülmektedir. EK-3 ise konuların içeriğindeki kelime dağılımını göstermektedir. Konu – kelime tablosunda ise satırlar kelimeleri sütunlar konuları temsil etmektedir, bu tabloda konuların içerdiği kelimeleri gösterilmektedir. Tablo 3.4 ise AUC eşik değerlerine göre getirilen deney sayılarını göstermektedir. Tablo 3.4 incelendiğinde LDA yöntemi için deneylerin %53.33'ünün 0.70 üzeri AUC değerine sahip olduğu belirlenmiştir.

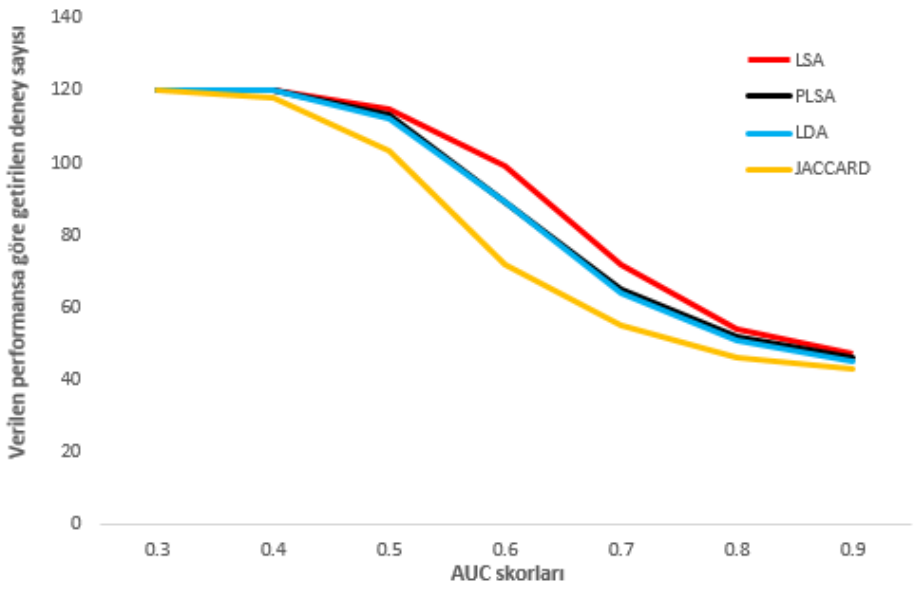
Tablo 3.4. LDA yönteminde AUC eşik değerine göre getirilen deney sayısı

Eşik Değeri	Eşik değerinden büyük ortalama AUC skoruna sahip toplam deney sayısı
0.3	120
0.4	120
0.5	112
0.6	89
0.7	64
0.8	51
0.9	45

3.3.3. Deneysel sonuçların yorumlanması

Yöntemlerin genel performansını görselleştirmek değerlendirmek ve karşılaştırmak için, belirli bir AUC değerinden daha iyi performans gösteren deneylerin sayısı Şekil 3.2'de verilmiştir. Şekil 3.2'de x ekseninde ROC eğrilerinden elde dilen AUC değerleri, y ekseninde ise bu değerlere karşılık gelen deney sayısı gösterilmektedir. Elde edilen sonuçlara göre sözcüksel benzerlik yöntemi Jaccard en düşük performansa sahip olurken, LSA en yüksek performansa sahip olmuştur. Çalışmada kullanılan anlamsal benzerlik yöntemleri PLSA ve LDA birbirlerine yakın performans skorlarına sahip olmuşlardır.

Şekildeki en yüksek eğri, en etkili geri getirim performansını göstermektedir ve LSA yöntemine aittir. LSA yönteminin, birçok sorgu için biyolojik olarak ilgili deneyleri başarılı bir şekilde tespit etmede diğer yöntemlerden daha başarılı bulunduğu şekilde açıkça görülmektedir ve LSA yöntemi için 0.7'den büyük ortalama AUC'nin deneylerin yarısından fazlasında sağlanabildiği belirlenmiştir. EK 4'de tüm yöntemlerle elde edilen AUC değerleri gösterilmektedir.



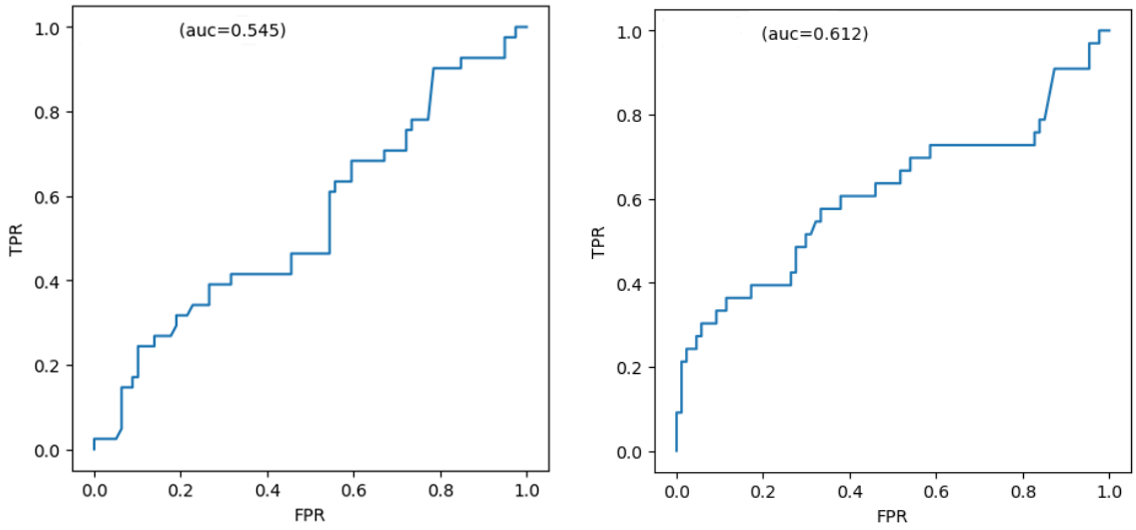
Şekil 3.2. Jaccard, LSA, PLSA, LDA yöntemleri için AUC eşik değerlerine göre getirilen deney sayıları

Önerilen sistemin geri getirim başarısını değerlendirmek için deney açıklamaları arasındaki ilgililik seçilen 2 sorgu deneyi için incelenmiştir. Seçilen sorgu deneyleri, kendisi dışında veri setindeki en az 2 ilgili deney için yüksek benzerlik puanına sahip deneylerdir. Her bir sorgu deneyi için temel gerçek tarafından sorgu deneyi ile ilgili olduğu belirlenmiş 2 deney ve yine temel gerçek tarafından ilgisiz olduğu belirlenen 2 deney alınmıştır. Tablo 3.5. Jaccard, LSA, PLSA ve LDA yöntemlerinin seçilen deneyler için buldukları deney benzerliklerini göstermektedir. Tablodan açıkça yöntemlerin benzerlik oranlarının ilgili deneyler için yüksek çıkarken ilgisiz deneyler için düşük çıktığı anlaşılmaktadır. Anlamsal benzerlik yöntemleri deney ilgililiğini başarılı bir şekilde yakalayabilirken Jaccard yöntemi deney ilgililğini bulmada zayıf kalmıştır. Şekil 3.3.a. ve Şekil 3.3.b. en başarılı bulan yöntem LSA için sırasıyla GSE6349 ve GSE35325-2 sorgu deneylerinin ROC eğrilerini göstermektedir. Şekiller aynı zamanda sorgu deneyleri için AUC değerlerini içermekte olup,

sorgu deneyleri için yöntemlerin ilgili deneyleri geri getirebildiğini göstermektedir. Tablo 3.5, Şekil 3.3.a. ve Şekil 3.3.b'den yola çıkılarak anlam tabanlı yöntemlerin deneylerin ilgili ya da ilgisiz olduğunu bulabildiği ancak LSA yönteminin daha başarılı olduğu kanısına varılabilir.

Tablo 3.5. Örnek sorgu deneyleri ile yöntem sonuclarının karşılaştırılması

Sorgu Deneyi	Karşılaştırma Deneyi	Jaccard	LSA	PLSA	LDA	Temel Gerçek
GSE6349	GSE30098	0.099	0.956	0.608	0.809	0.449
GSE6349	**GSE3350-2**	**0.100**	**0.964**	**0.612**	**0.808**	**0.525**
GSE6349	GSE18975-6	0.067	0.144	1.90E-08	0.046	0.258
GSE6349	GSE19271-7	0.029	0.102	1.21E-08	0.042	0.264
GSE35325-2	GSE18985-2	0.178	0.985	0.998	0.469	0.468
GSE35325-2	GSE19261-1	0.136	0.994	0.987	0.365	0.357
GSE35325-2	GSE576-1	0.068	0.173	0.003	0.052	0.316
GSE35325-2	GSE577-4	0.050	0.013	0.002	0.059	0.189



a. LSA yöntemi için GSE6349 sorgu Deneyi ROC eğrisi

b. LSA yöntemi için GSE35325-2 sorgu Deneyi ROC eğrisi

Şekil 3.3. LSA yöntemi için GSE6349 ve GSE35325-2 sorgu Deneyleri ROC eğrileri

Ayrıca Tablo 3.5'de ilgili deney çiftleri olarak gösterilen GSE6349–GSE30098, GSE6349 – GSE3350-2, GSE35325-2– GSE18985-2 ve GSE35325-2 – GSE19261-1 deney ikililerinin deney metinleri okunduğunda gözle görülür benzerlik fark edilebilmektedir. Deney isimlerinde geçen "-" ifadesi aynı GEO verisindeki deney numarasını gösterir. İlk sorgu deneyi olan GSE6349 yanal kökler oluşurken Arabidopsis köklerinin ksilem kutbu perisikl hücrelerinin ekspresyon verilerini gözlemlemek için gerçekleştirilmiştir [57]. Deneyde Arabidopsis bitkisinin kinase ACR4 benzeri reseptörlerini ana faktör olarak tanımlamak için yanal köklerdeki sıralı periskil hücrelerinin transkript profili kullanılmıştır. Aynı zamanda deneyde G1'den S'e geçişin durdurulmasını, oksin hormonunun izlenmesini, G1'den S'e ve G2'den M'e geçişi ya da asimetrik hücre bölünmesini içeren farklı zaman aralıklarında ksilem kutup periskil hücrelerinde GFP (Green Flourescent Protein) işaretçisi içeren Arabidopsis köklerine FACS (Fluorescence-activated cell sorting - Floresans tarafından aktive edilmiş hücrelerin ayrılması) uygulanmıştır. Bu deney ile ilgili olarak getirilen deneylerden biri GSE30098 deneyidir. Deney, kükürt eksikliğinin Arabidopsis bitkisi köklerinin ekspresyon analizine etkisi ile ilgilidir. Arabidopsis kök gelişimi ve uygulanan etki (stress) arasındaki ilişkiyi anlamak için söz konusu deney, genom çapında testler kullanılarak yapılmıştır [58]. Sorgu ile bu deney arasındaki en dikkat çekici benzerlik, farklı zaman dilimlerinde ve bazı çevresel faktörlere yanıt olarak bitkinin kök gelişim sürecini gözlemlemek için geliştirilmiş olmalarıdır. Birinci sorgu deneyi ile ilgili olarak geri getirilen ikinci deney ise GSE3350-2 deneyidir, bu deneyde yanal kök benzeri kullanılarak oksin hormonu ile uyarılmış hücre bölünmesinin arkasındaki mekanizma incelenmiştir [59]. Deneyler arasındaki benzerlik, oksin hormonu ve hücre bölünmesini araştırmak için yapılmış olmalarıdır. Ayrıca, önerilen sistemin aynı amaç ve aynı uyaranı kullanan deneyler arasındaki ilişkiyi çıkarabileceğini de açıkça söylenebilir. İkinci sorgu deneyi GSE35325-2 ise belirli rizobakterilerin Arabidopsis Thaliana üzerinde büyümeyi önleyen geçici etkilerini gözlemlemek için gerçekleştirilmiştir. Bu deney ile ilgili olarak geri getirilmiş GSE18985-2 deneyinde ise köklerinde giberellin (GA) eksikliği olan Arabidopsis bitkisinin GA hormonuna erken yanıt veren genlerini bulmak amaçlanmıştır. Bu iki deneyin ortak noktası ise büyüme engelleyici bir bakteri ve büyüme destekleyici bir hormonun etkilerinin gözlemlenmesidir. GSE35325-2 deneyi ile ilgili olarak geri getirilen 19261-1 deneyi ise ışığın bitki büyüme süreci üzerine etkilerinin incelendiği bir deneydir. Her iki deneyde ışığın bitki büyümesine etkisini farklı durumlar altında inceleyen çalışmaların sonuçlarıdır.

Tablo 3.6'da Jaccard, LSA, PLSA ve LDA yöntemlerinin performansları gösterilmektedir. Tablo 3.7'de ise yöntemlerden elde edilen deneylere özgü AUC değerlerinin ortalaması en yüksek değeri ve standart sapması gösterilmektedir. Tablo 3.6 ve Tablo 3.7 incelendiğinde LSA yönteminin hem en iyi performansa sahip olduğu hem de en düşük standart sapmaya sahip olduğu belirlenmiştir.

Tablo 3.6. Yöntem Performansları

Yöntem	Jaccard	**LSA**	PLSA	LDA
AUC Skor	0.73	**0.79**	0.771	0.768

Tablo 3.7. Deneylere Özgü AUC Skorlarının Yöntemlere Göre Ortalama Değerleri, En Yüksek Değerleri ve Standart Sapma Değerleri

	Jaccard	**LSA**	PLSA	LDA
Ortalama	0.73	**0.79**	0.771	0.768
En Yüksek	1	**1**	1	1
Standart Sapma	0.21	**0.18**	0.193	0.196

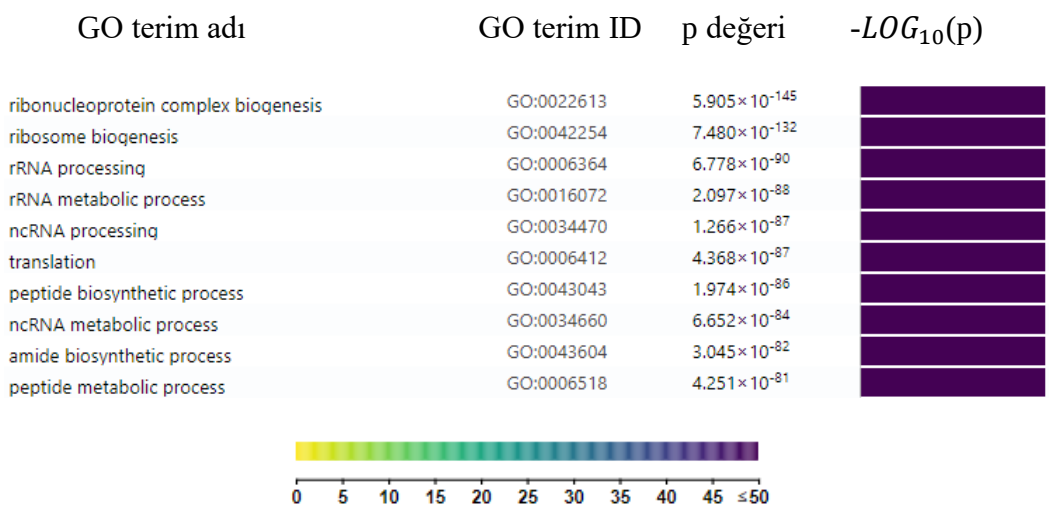
Elde edilen sonuçların geçerliliği hem istatistiksel hem de biyolojiksel olarak gösterilmiştir. Çalışmada önceki bölümlerde belirtildiği gibi çalışma için en başarılı bulunan yöntem LSA ve diğer yöntemler arasındaki ikili AUC puanları arasındaki fark için Wilcoxon Signed Rank Test ve Paired t-test uygulanmıştır. Tablo 3.8 elde edilen Wilcoxon Signed Rank Test ve Paired t-test sonuçlarını göstermektedir. İstatistiksel değerlendirme testleri ile elde edilen p değerlerinin LSA-PLSA arası Wilcoxon Signed Rank Test hariç 0.05'den düşük olduğu gözlemlenmektedir. Bu sonuç yöntemlerin sonuçları arasındaki farkın istatistiksel olarak anlamlı olduğunu göstermektedir.

Tablo 3.8. Wilcoxon Testi ve Paired t-testi p Değerleri Sonuçları

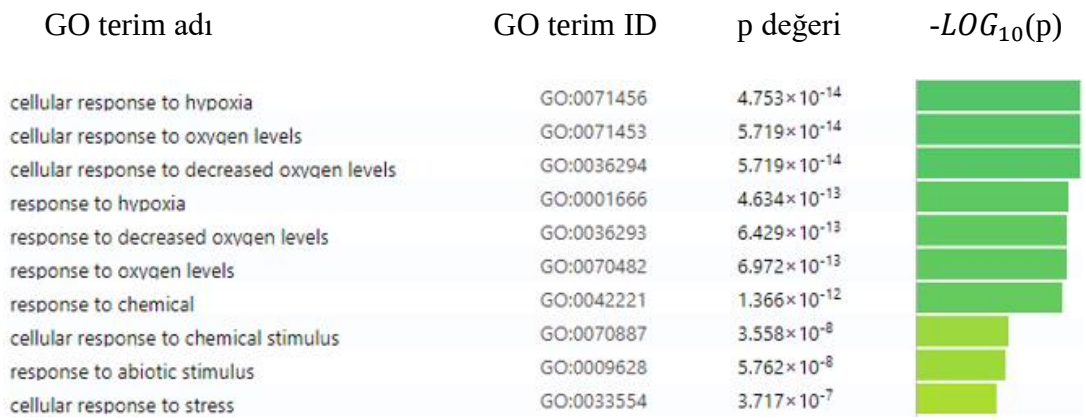
Yöntem	Wilcoxon Signed Rank Test p değeri	Paired t-test p değeri
LSA - Jaccard	5.834E-11	2.34E-11
LSA - PLSA	0.054	0.018
LSA - LDA	0.003	0.009

Yöntemler tarafından ilgili bulunan GSE6349 ve GSE30098 deney çiftinin biyolojik benzerliğine bakılmıştır. Biyolojik benzerlik için detayları önceki bölümlerde anlatılmış GO analizi uygulanmıştır. Şekil 3.4. ve Şekil 3.5. sırasıyla GSE6349 ve GSE30098 deneylerinin GO analizi sonuçlara göre deneylerde en fazla zenginleştirilmiş BP'lere ait 10 GO terimini göstermektedir. Şekle göre $-LOG_{10}(p)$ değeri renk koyulaştıkça yükselmektedir. GO Analizi

sonuçlarına göre GSE6349 ve GSE30098 deneyleri ortak 4 adet GO Terimine sahiptir. Ortak olan bu GO terimlerinin adı, ID ve p değerleri Tablo 3.9.'da verilmiştir. Yine yöntemler ile elde edilen sonuçlara göre ilgili bulunan GSE35325-2 ve GSE18985-2 deney çiftinin biyolojik benzerliği incelenmiştir. İnceleme sonucuna göre Şekil 3.6. ve Şekil 3.7. sırasıyla GSE35325-2 ve GSE18985-2 deneylerinin GO analizi sonuçlarına göre deneylerde en fazla zenginleştirilmiş BP'lere ait 10 GO Terimini göstermektedir. GO Analizi sonuçlarına göre GSE35325-2 ve GSE18985-2 deneyleri ortak 4 adet GO terimine sahiptir. Ortak olan bu GO terimlerinin adı, ID ve p değerleri Tablo 3.10.'da verilmiştir. GO Analizi yapılması sırasında g:profiler aracı kullanılmış olup sonuçları gösteren şekil ve veriler oradan alınmıştır. Benzer olarak bulunan deneylerde paylaşılan ortak GO Terimlerinin olması deneylerin biyolojik olarak benzer olduğunu kanıtlamaktadır.



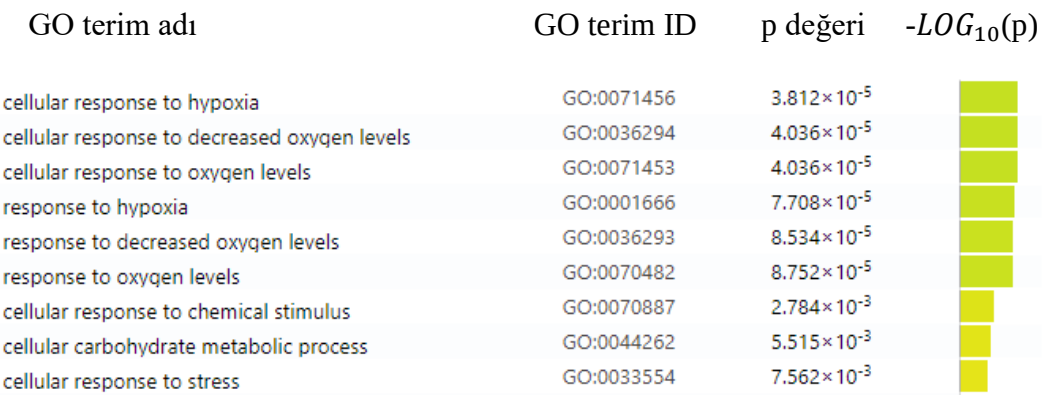
Şekil 3.4. GSE6349 Sorgu Deneyi İçin GO Analiz Sonuçları



Şekil 3.5. GSE30098 Sorgu Deneyi İçin GO Analiz Sonuçları

Tablo 3.9. GSE6349 ve GSE30098 Deneyleri İçin Ortak GO Terimleri ve p Değerleri

GO terim adı	GO terim ID	p değeri
alpha-amino acid metabolic	GO:1901605	1.123×10^{-4}
organic acid metabolic process	GO:0006082	3.682×10^{-2}
cellular amino acid metabolic process	GO:0006520	1.143×10^{-8}
oxoacid metabolic process	GO:0043436	4.014×10^{-2}



Şekil 3.6. GSE35325-2 Sorgu Deneyi İçin GO Analiz Sonuçları

Şekil 3.7. GSE18985-2 Sorgu Deneyi İçin GO Analiz Sonuçları

Tablo 3.10. GSE35325-2 ve GSE18985-2 Deneyleri İçin Ortak GO Terimleri ve p Değerleri

GO terim adı	GO terim ID	p DEĞERİ
cellular response to chemical stimulus	GO:0070887	2.784×10^{-3}
response to chemical	GO:0042221	1.365×10^{-2}
response to abiotic stimulus	GO:0009628	3.667×10^{-2}
response to stimulus	GO:0050896	4.446×10^{-2}

4. TARTIŞMA

Deneysel çalışmalar deneyi açıklamaya yönelik üst veri bilgileri ile veri tabanlarında tutulmaktadır. Kullanıcılar araştırmaları için veri tabanlarında saklanan yüksek miktardaki verinin içinden ilgili veriye ulaşmak için etkin arama yöntemlerine ihtiyaç duymaktadır. Mevcut veri tabanlarının kullandığı geri getirim yöntemlerinin sınırlamaları mevcuttur. Örneğin sadece sözcüksel eşleşme ve benzerlik ile gerçekleştirilen üst veri tabanlı aramada mantıksal operatörlerle (AND, OR) kullanıcın mantıksal ihtiyaçları yapılandırılırken sorgu ile uyum kaybedilebilir ya da anahtar kelime tabanlı aramalar için deney açıklamaları yeterli bilgiyi barındırmayabilir. Bu sınırlamaları aşmak için kelimelerin eş anlamlılarını ve çok anlamlı kelimeleri dikkate alan anlam tabanlı benzerlik yöntemleri önerilmiştir. Önerilen bu yöntemin dışında ise deneylerin soyut içeriklerinin benzerliğinden faydalanarak deney geri getirimi yapan içerik tabanlı yöntem mevcuttur.

Çalışma kapsamında veri tabanlarının deney benzerliklerini bulma yetersizliğine çözüm olarak genomik veri tabanlarından ilgili deneylerin geri getirimi için metin tabanlı bir geri getirim sistemi geliştirilmiştir. Geliştirilen sistemde sözcük tabanlı yöntem olarak Jaccard benzerliği, anlam tabanlı benzerlik yöntemleri olarak LSA, PLSA ve LDA yöntemleri kullanılıp yöntem performansları kıyaslanmıştır. Sistem performansı, GEO veri tabanından alınmış organizma olarak Arabidopsis Thaliana bitkisinin kullanıldığı, zaman serisi özellikli 120 mikro dizi deneyinin açıklama metinleri önişlem aşamasından geçirilerek test edilmiştir. Elde edilen sonuçların biyolojik ve istatistiksel anlamlılıkları doğrulanmıştır. Çalışma kapsamında elde ede edilen sonuçlar aynı veri seti ile gerçekleştirilmiş içerik tabanlı bir çalışma ile karşılaştırılmıştır. Tablo 4.1'de tüm yöntemler ve içerik tabanlı çalışma ile elde edilen sonuçlar verilmektedir. Tablo 4.1'de görülebileceği gibi çalışma sonucunda elde edilen deneysel sonuçlara göre 0.79 başarı ile en iyi yöntem olarak LSA belirlenirken, 0.73 başarı ile Jaccard en düşük performanslı yöntem olarak bulunmuştur. Sözcüksel benzerliğe dayalı yöntemlerin anlamsal benzerliğe dayalı yöntemlere göre daha düşük bir performansa sahip olması beklenen bir durumdur. Ancak anlam tabanlı yöntemler kendi içlerinde değerlendirildiğinde 0.79 başarı oranına sahip LSA yöntemi ile 0.771 başarı oranına sahip PLSA ve 0.768 başarı oranına sahip LDA yöntemlerinden daha başarılı sonuçlar elde edilmiştir. Anlam tabanlı yöntemlerinin gelişimlerine bakıldığında LDA yöntemi PLSA'nın, PLSA yöntemi LSA'nın eksik ve zayıf yönlerini gidermek için çıkan yöntemler olduğundan,

elde edilen sonuçlar beklenmedik sonuçlar olmuştur. Bu sonuçların sebebi veri setinin boyutunun küçük olmasına bağlanmıştır. Ancak PLSA ve LDA yöntemlerinin birbirlerine yakın sonuçlar üretmesi ise her iki yöntem de konu modelleme yöntemlerinden olduğundan beklenen bir sonuçtur. Yine Tablo 4.1'den hareketle içerik tabanlı deney getirimi hedefleyen [21] çalışmada başarı oranının 0.74 olduğu görülmektedir. Çalışmaların sonuçları incelendiğinde sözcük tabanlı deney getirim yönteminin (Jaccard) içerik tabanlı deney getiriminin gerisinde kaldığı, anlam tabanlı deney getirim yöntemlerinin ise içerik tabanlı deney getiriminden yüksek başarı oranına sahip olduğu gözlemlenmiştir.

Tablo 4.1. Çalışmada kullanılan yöntemler ve içerik tabanlı yöntem performansları

Yöntem	Sözcük Tabanlı	Anlam Tabanlı			İçerik tabanlı
	Jaccard	**LSA**	PLSA	LDA	
AUC Skor	0.73	**0.79**	0.771	0.768	0.74

Gelecek dönemlerde bu çalışmada uygulanmamış konu modelleme yöntemleri (örn. CTM) ile çalışma tekrarlanarak, çalışmaya katkı verilebilir. Veri seti boyutunun yöntem performanslarına etkisi hakkında çalışmalar yapılabilir. Farklı veri türleri için (örn. miRNA, metagenom dizisi) çalışma tekrarlanarak çalışma kapsamı genişletilebilir. Biyologların anahtar kelime tabanlı arama yöntemlerinde sık kullandığı kelimler ile deney geri getirimi performansı hesaplanıp anlam tabanlı yöntemlerle karşılaştırılması yapılarak çalışma genişletilebilir.

KAYNAKLAR

[1] T. Barrett, R. Edgar, “Mining microarray data at NCBI’s Gene Expression Omnibus (GEO),” Methods in molecular biology, vol. 338, pp. 175-190, 2006, doi:10.1385/1-59745-097-9:175.

[2] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, “GenBank,” Nucleic Acids Res. , vol. 41, pp. D36-D42, 2013, doi: 10.1093/nar/gks1195.

[3] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. R. Serra, S. A. Sansone,“ArrayExpress—a public repository for microarray gene expression data at the EBI,” *Nucleic Acids Res.*, vol. 33, no. 1, pp. 68-71, 2003, doi: https://doi.org/10.1093/nar/gkg091

[4] S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D.C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, P. Zhang, “The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology. research materials and community,” Nucleic Acids Res, vol. 31 no. 1, pp. 224–228, 2003, doi: 10.1093/nar/gkg076.

[5] H. Oğul, “Content-Based Retrieval of Microarray Experiments, Pattern Recognition” in *Computational Molecular Biology: Techniques and Approaches*, ELLOUMI, M. John, Hoboken, New Jersey, Wiley & Sons, 2015, p.315–334.

[6] D. Fenstemacher, “Introduction to bioinformatics,” *Journal of the American Society for Information Science and Technology*, vol. 56, no. 5, pp. 440–446, 2005, doi: 10,1002/asi.20133.

[7] Your Genome. “What is a geneome?” yourgenome.org. http://www.yourgenome.org/facts/what-is-a-genome (Accessed: July 29, 2021).

[8] National Human Genome Research Institude. “What’s a Genome?” genome.gov. https://www.genome.gov/About-Genomics/Introduction-to-Genomics (Accessed: July 29, 2021).

[9] National Human Genome Research Institude. “A Brief Guide to Genomics.” genome.gov. https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics (Accessed: July 29, 2021).

[10] National Human Genome Research Institude. “Gene Expression.” genome.gov. https://www.genome.gov/genetics-glossary/Gene-Expression (Accessed: July 27, 2021).

[11] S. H. BAL, F. Budak, “Mikroarray Teknolojisi,” *Uludağ Üniversitesi Tıp Fakültesi Dergisi*, vol. 38, no. 3, pp. 227–233, 2012.

[12] M.A. Tuncel, “A statistical framework for the analysis of genomic data, ” M.S thesis, Dept. Electronics, Informatics and Bioengineering, Polytechnic University of Milan, Milan, Italy, 2017.

[13] Gene Expression Omnibus, “Series GSE576.” ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE576 (Accessed: July 25, 2021).

[14] D. Dede Şener, “EXPERIMENT RETRIEVAL IN GENOMIC DATABASES,” Ph.D. thesis, Dept. Computer. Eng., Başkent Univ., Ankara, Türkiye, 2019.

[15] ML Wiki, “Information Retrieval.” mlwiki.org. http://mlwiki.org/index.php/Information_Retrieval (Accessed: July 29, 2021).

[16] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, S. Kaski, “Probabilistic retrieval and visualization of biologically relevant microarray experiments,” *Bioinformatics*, vol.25, no.12, pp. i145–i153,2009, doi: 10.1093/bioinformatics/btp215.

[17] J.M. Engreitz, A.A.Morgan, J.T. Dudley, R. Chen, R. Thathoo, R.B. Altman, A.J. Butte, "Content-based microarray search using differential expression profiles," *BMC Bioinformatics*, vol. 11, pp. 603-614, 2010, doi: https://doi.org/10.1186/1471-2105-11-603

[18] A. Hayran, H. Oğul, E.E. Özkoc, "Content-based search in time-series microarray databases," in *Database and Expert Systems Applications (DEXA) 25th International Workshop*, Munich, Germany, Sept. 1-5, 2014, pp. 89–93, doi: http://dx.doi.org/10.1109/DEXA. 2014.33

[19] S. Seth, N. Välimäki, S. Kaski, A. Honkela, "Exploration and retrieval of whole-metagenome sequencing samples," *Bioinformatics*, vol. 30, no. 17, pp. 2471-2479, 2014, doi:10.1093/bioinformatics/btu340.

[20] D.D. Şener, H. Oğul , "Inferring Similarity Between Time Series Microarrays: A Content Based Approach,"in *IEEE 2nd International Conference on Cybernetics (CYBCONF)*, Gdynia, Poland, june 24-26, 2015, p. 201-205, doi: 10.1109/CYBConf.2015.7175932.

[21] D.D. Şener, H Oğul, "Retrieving Relevant Time-Course Experiments: a Study on Arabidopsis Microarrays", *IET Systems Biology*, vol. 10, no. 3, pp. 87-93, 2016, doi: 10.1049/iet-syb.2015.0042.

[22] K. Açıcı, Y.K. Terzi, H. Oğul, "Retrieving relevant experiments: The case of microRNA microarrays," *BioSystems*, vol. 134, pp. 71–78 2015, doi: 10.1016/j.biosystems.2015.06.003

[23] D.D. Şener, D. Santoni, G. Felici, H. Oğul H, "A Content-Based Retrieval Framework for Whole Metagenome Sequencing Samples," *Journal of Integrative Bioinformatics*, vol. 15, no. 4, pp. 20170067, 2018, doi: https://doi.org/10.1515/jib-2017-0067.

[24] W. Fujibuchi, L. Kiseleva, T. Taniguchi, H, Harada, P. Horton, "CellMontage: similar expression profile search server, " *Bioinformatics*, vol. 23 no. 22, pp. 3103-3104, 2007. doi: https://doi.org/10.1093/bioinformatics/btm462

[25] J.M. Engreitz, R. Chen, A.A. Morgan, J.T. Dudley, R. Mallelwar, A.J. Butte, "ProfileChaser: searching microarray repositories based on genome-wide patterns of differential Expression, " *Bioinformatics*, vol. 27, no. 23, pp. 3317-3318, 2011.

[26] G. Williams, "SPIEDw: a searchable platform-independent expression database web tool," *BMC genomics*, vol. 14, no. 1, pp. 765-770, 2013, doi: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-765.

[27] H. S. Le, Z. N. Oltvai, Z. Bar-Joseph, "Cross-species queries of large gene expression databases," *Bioinformatics*, vol.26, no19, pp. 2416-2423, 2010, doi: 10.1093/bioinformatics/btq451.

[28] J. Caldas, N. Gehlenborg, A.Faisal, A. Brazma, S. Kaski, "Probabilistic retrieval and visualization of biologically relevant microarray experiments," *Bioinformatics*, vol.25, no. 12, pp. i145-i153 , 2009, doi:10.1093/bioinformatics/btp215.

[29] S. Lee, J. Song, Y. Kim , "An empirical comparison of four text mining methods," *The Journal of Computer Information Systems*, vol. 51, no. 1, pp. 1–10, 2010.

[30] D. Gautam, Z. Cai, V. Rus, "Effect of Domain Corpus Size and LSA Vector Dimension: A Study in Assessing Student Generated Short Texts in Virtual Internships without Participant Data," in *32th FLAIRS Conference*, Sarasota, Florida, USA, May 19-22, 2019, P. 179-184.

[31] Z. Tong, H. Zhang, "A text mining research based on LDA topic modelling," *Comput. Sci. Inf. Technol.*, vol. 6, pp. 201–210, 2016.

[32] Y. Kalepalli, S. Tasneem, P.D.P. Teja, S. Manne, "Effective Comparison of LDA with LSA for Topic Modelling," in *4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Melur, India, May. 13–15, 2020, p. 1245–1250, doi: 10.1109/ICICCS48265.2020.9120888.

[33] V. Rus, R. Banjade, N. Niraula, "Similarity Measures based on Latent Dirichlet Allocation," in *Computational Linguistics and Intelligent Text Processing (CICLing 2013)*, Berlin, Heidelberg, 2013, p. 459-470. https://doi.org/10.1007/978-3-642-37247-6_37

[34] L. Liu, L. Tang, W. Dong, S. Yao, W. Zhou "An Overview of Topic Modeling and Its Current Applications in Bioinformatics," *SpringerPlus*, vol. 5, pp. 1608-1629, 2016, doi: http://doi.org/10.1186/s40064-016-3252-8.

[35] W.H. Gomaa, A. A. Fahmy , "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013, doi: 10.5120/11638-7118.

[36] M.K. Vijaymeena, K. Kavitha , "A survey on Similarity Measures in Text Mining", *Machine Learning and Applications: An International Journal (MLAIJ)*, vol. 3, no. 1, 2016, doi: 10.5121/mlaij.2016.3103.

[37] datacamp "Latent Semantic Analysis using Python**."** datacamp.com. https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python (Accessed: July 26, 2021).

[38] I.T. Medeni, T.D. Medeni, "TOPIC MODEL IMPLEMENTATION TO FIND RELATED DOCUMENTS IN CORPORATE ARCHIVES IN REAL LIFE: "A CASE SCENARIO ON KNOWLEDGE RETRIEVAL", *INTERNATIONAL JOURNAL OF eBUSINESS AND eGOVERNMENT STUDIES*, vol. 5, no. 1, 2013, ISSN: 2146-0744. [Online]. Available: https://dergipark.org.tr/en/download/article-file/257109

[39] Statistic How To, "Jaccard Index / Similarity Coefficient." statisticshowto.com. https://www.statisticshowto.com/jaccard-index/ (Accessed: July 30 2021).

[40] T.K. Landauer, P.W. Foltz, D. Laham, "An introduction to latent semantic analysis,", *Discourse Processes*, pp. 259–28, 1998.

[41] Latent Semantic Analysis@ CU Boulder, "What is LSA." lsa.colorado.edu. http://lsa.colorado.edu/ (Accessed : July 26, 2021).

[42] N.A. Albatayneh, K.I. Ghauth, F.F. Chua, "A Semantic content-based forum recommender system architecture based on content-based filtering and latent semantic analysis*,"* in *Recent Advances on Soft Computing and Data Mining*, T. Herawan, R. Ghazali, M. Deris, Cham, Switzerland, Springer International Publishing: 2014, pp. 369-378, doi:10.1007/978-3-319-07692-8_35

[43] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Fifteenth Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden, July 30 – Aug.,1, 1999, p. 289-296.

[44] D. Oneata, "Probabilistic latent semantic analysis, " in Proceedings of The Fifteenth conference on Uncertainty, Stockholm, Sweden, 1999, p. 1-7.

[45] R. Alghamdi , K. Alfalqi , "A Survey of Topic Modeling in Text Mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, pp. 147-153, 2015.

[46] D.M. Blei, Y.N. Andrew, M.I. Jordan, "Latent dirichlet Allocation," T*he Journal of Machine Learning Research*, vol. 3, pp. 993- 1022, 2003.

[47] ScikitLearn. "sklearn.decomposition.LatentDirichletAllocation." scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html (Accessed: July 30 2021).

[48] ScikitLearn. "2.5.7. Latent Dirichlet Allocation (LDA)." scikit-learn.org. https://scikit-learn.org/stable/modules/decomposition.html#latentdirichletallocation (Accessed: July 30 2021).

[49] A. Subramanıan, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545-15550, 2005, doi: https://doi.org/10.1073/pnas.0506580102

[50] M. Ashburner et al., “Gene Ontology: tool for the unification of biology”, *Nat. Genet*., vol. 25, no. 1, pp. 25–29, 2020, doi:10.1038/75556

[51] A. Kucukural, O. Yukselen, D.M.Ozata, M.J. Moore, M. Garber, “DEBrowser: interactive differential expression analysis and visualization tool for count data”, *BMC Genomics,* vol. 20, no. 6, 2019, doi: https://doi.org/10.1186/s12864-018-5362-x.

[52] M.I. Love, W. Huber, S. Anders, “Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2,” *Genome Biology*, vol. 15, pp. 550-570, 2014. doi: 10.1101/002832

[53] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arat, P. Adler, H. Peterson, J. Vilo, “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update),” *Nucleic Acids Research*, vol. 47, no. W1, pp. W191–W198, 2019, doi: 10.1093/nar/gkz369

[54] B. Rosner, R.J. Glynn, M.L.T. Lee, “The Wilcoxon signed rank test for paired comparisons of clustered data” *Biometrics,* vol. 62, no. 1, pp. 185–192, 2006, doi: 10.1111/j.1541-0420.2005.00389.x

[55] T. K. Kim, “T test as a parametric statistic,” *Korean journal of anesthesiology*, vol.. 68, no. 6, pp. 540-546, 2015, doi:10.4097/kjae.2015.68.6.540.

[56] T. Fawcett, “An introduction to ROC analysis, ” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006, doi: https://doi.org/10.1016/j.patrec.2005.10.010.

[57] D.I. Smet et al., “Receptor-like kinase ACR4 restricts formative cell divisions in the Arabidopsis root,” *Science*, vol. 322, no. 5901, pp. 594-597, 2008, doi: 10.1126/science.116015

[58] A.S.I. Pascuzzi, T. Jackson, H. Cui, J.J. Petricka, W. Busch, H. Tsukagoshi, P.N. Benfey, “Cell identity regulators link development and stress responses in the Arabidopsis root. Developmental cell,” vol. 21, no.4, pp. 770-782, 2011, doi: https://doi.org/10.1016/j.devcel.2011.09.009.

[59] S. Vanneste, B.D. Rybel, G.T.S. Beemster, K. Ljung K, I.D. Smet, G. V. Isterdael, M. Naudts, R. Iida, W. Gruissem, M. Tasaka, D. Inzé, H. Fukaki, T. Beeckman, "Cell cycle progression in the pericycle is not sufficient for SOLITARY ROOT/IAA14-mediated lateral root initiation in Arabidopsis thaliana," *Plant Cell*, vol. 17, no. 11, pp. 3035-3050, 2005, doi: 10.1105/tpc.105.035493

EKLER

EK 1: LSA, PLSA ve LDA Yöntemleri İdeal Boyut Ve Konu Sayılarını Belirleme
EK 2: LDA Doküman Konu Dağılımı
EK 3: LDA Konu Kelime Dağılımı
EK 4: JACCARD, LSA, PLSA ve LDA Yöntemleri AUC Sonuçları

EK 1: LSA, PLSA VE LDA Yöntemleri İdeal Boyut Ve Konu Sayılarını Belirleme

Tablo EK1.1. LSA Boyut Değişimi

Deney	Boyut											
	3	5	10	15	20	25	50	75	100	105	110	115
GSE1111	0.594	0.639	0.495	0.497	0.439	0.439	0.41	0.391	0.391	0.391	0.391	0.391
GSE18975-6	0.872	0.895	0.773	0.721	0.758	0.779	0.831	0.831	0.831	0.831	0.831	0.831
GSE25171-2	0.509	0.543	0.537	0.476	0.428	0.46	0.481	0.487	0.485	0.485	0.485	0.485
GSE3326-1	0.675	0.709	0.651	0.574	0.558	0.471	0.512	0.513	0.512	0.512	0.512	0.512
GSE45215-1	0.562	0.633	0.568	0.598	0.556	0.585	0.589	0.596	0.595	0.595	0.595	0.595

Tablo EK1.2. PLSA Konu Sayısı Değişimi

Deney	Boyut											
	3	5	10	15	20	25	50	75	100	105	110	115
GSE19263-2	0.455	0.537	0.513	0.552	0.409	0.431	0.492	0.452	0.453	0.453	0.453	0.453
GSE19266-2	0.59	0.598	0.702	0.76	0.609	0.685	0.59	0.652	0.652	0.652	0.652	0.652
GSE26679-1	0.594	0.643	0.513	0.58	0.511	0.558	0.551	0.561	0.564	0.564	0.564	0.564
GSE3416	0.456	0.555	0.495	0.513	0.456	0.463	0.482	0.509	0.509	0.509	0.509	0.509
GSE40140-2	0.68	0.651	0.749	0.815	0.736	0.808	0.559	0.451	0.451	0.451	0.451	0.451

Tablo EK.1.3. LDA Konu Sayısı Değişimi

Deney	Boyut											
	3	5	10	15	20	25	50	75	100	105	110	115
GSE16143-2	0.562	0.592	0.644	0.521	0.476	0.523	0.486	0.509	0.498	0.432	0.5	0.513
GSE30098	0.465	0.553	0.623	0.523	0.552	0.521	0.497	0.465	0.54	0.323	0.564	0.453
GSE30223-2	0.58	0.467	0.575	0.554	0.51	0.417	0.385	0.513	0.463	0.519	0.419	0.524
GSE3416	0.57	0.522	0.629	0.507	0.527	0.413	0.459	0.481	0.556	0.386	0.493	0.533
GSE5747	0.462	0.566	0.616	0.526	0.543	0.475	0.394	0.491	0.526	0.337	0.498	0.502

EK 2: LDA Doküman Konu Dağılımı

Tablo Ek 2. LDA Doküman Konu Dağılımı

Doküman	Konu 0	Konu 1	Konu 2	Konu 3	Konu 4	Konu 5	Konu 6	Konu 7	Konu 8	Konu 9	Baskın Konu
Doc0	0.01	0.01	0.6	0.29	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc1	0.01	0.01	0.6	0.29	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc2	0.01	0.01	0.6	0.29	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc3	0.01	0.01	0.61	0.29	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc4	0.01	0.01	0.61	0.29	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc5	0.01	0.01	0.01	0.46	0.44	0.01	0.01	0.01	0.01	0.01	3
Doc6	0.01	0.16	0.45	0.31	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc7	0.01	0.16	0.45	0.31	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc8	0.01	0.01	0.01	0.28	0.64	0.01	0.01	0.01	0.01	0.01	4
Doc9	0.01	0.01	0.01	0.27	0.65	0.01	0.01	0.01	0.01	0.01	4
Doc10	0.01	0.01	0.01	0.3	0.62	0.01	0.01	0.01	0.01	0.01	4
Doc11	0.01	0.01	0.01	0.34	0.01	0.01	0.15	0.46	0.01	0.01	7
Doc12	0.01	0.01	0.01	0.34	0.01	0.01	0.15	0.46	0.01	0.01	7
Doc13	0.01	0.01	0.53	0.38	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc14	0.01	0.01	0.53	0.38	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc15	0.01	0.01	0.01	0.9	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc16	0.01	0.01	0.01	0.9	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc17	0.02	0.02	0.02	0.02	0.02	0.02	0.8	0.02	0.02	0.02	6
Doc18	0.41	0.01	0.01	0.48	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc19	0.41	0.01	0.01	0.48	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc20	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc21	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc22	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc23	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc24	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc25	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc26	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	1
Doc27	0.24	0.01	0.23	0.45	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc28	0.24	0.01	0.23	0.45	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc29	0.01	0.01	0.51	0.4	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc30	0.01	0.01	0.5	0.41	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc31	0.01	0.01	0.54	0.39	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc32	0.01	0.01	0.54	0.38	0.01	0.01	0.01	0.01	0.01	0.01	2
Doc33	0.01	0.01	0.01	0.01	0.01	0.01	0.88	0.01	0.01	0.01	6
Doc34	0.01	0.01	0.01	0.01	0.01	0.01	0.88	0.01	0.01	0.01	6
Doc35	0.01	0.01	0.01	0.01	0.01	0.01	0.88	0.01	0.01	0.01	6

Doc36	0.01	0.01	0.01	0.4	0.51	0.01	0.01	0.01	0.01	0.01	4
Doc37	0.01	0.01	0.52	0.33	0.01	0.01	0.01	0.01	0.01	0.07	2
Doc38	0.01	0.01	0.52	0.33	0.01	0.01	0.01	0.01	0.01	0.07	2
Doc39	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.91	9
Doc40	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.9	9
Doc41	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.91	9
Doc42	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.91	9
Doc43	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.91	9
Doc44	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.91	9
Doc45	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.91	9
Doc46	0.01	0.01	0.01	0.29	0.01	0.01	0.61	0.01	0.01	0.01	6
Doc47	0.01	0.01	0.01	0.29	0.01	0.01	0.61	0.01	0.01	0.01	6
Doc48	0.01	0.01	0.01	0.35	0.55	0.01	0.01	0.01	0.01	0.01	4
Doc49	0.01	0.01	0.01	0.35	0.55	0.01	0.01	0.01	0.01	0.01	4
Doc50	0.01	0.01	0.01	0.01	0.01	0.01	0.88	0.01	0.01	0.01	6
Doc51	0.01	0.01	0.01	0.01	0.01	0.01	0.88	0.01	0.01	0.01	6
Doc52	0.01	0.01	0.01	0.01	0.01	0.01	0.88	0.01	0.01	0.01	6
Doc53	0.01	0.01	0.01	0.01	0.01	0.01	0.48	0.01	0.01	0.4	6
Doc54	0.01	0.01	0.01	0.01	0.01	0.01	0.48	0.01	0.01	0.4	6
Doc55	0.01	0.01	0.01	0.31	0.58	0.01	0.01	0.01	0.01	0.01	4
Doc56	0.01	0.01	0.01	0.31	0.58	0.01	0.01	0.01	0.01	0.01	4
Doc57	0.01	0.01	0.01	0.31	0.58	0.01	0.01	0.01	0.01	0.01	4
Doc58	0.45	0.01	0.01	0.32	0.01	0.01	0.15	0.01	0.01	0.01	0
Doc59	0.46	0.01	0.01	0.31	0.01	0.01	0.15	0.01	0.01	0.01	0
Doc60	0.02	0.02	0.56	0.3	0.02	0.02	0.02	0.02	0.02	0.02	2
Doc61	0.01	0.01	0.01	0.36	0.01	0.01	0.57	0.01	0.01	0.01	6
Doc62	0.01	0.01	0.01	0.36	0.01	0.01	0.57	0.01	0.01	0.01	6
Doc63	0.01	0.01	0.15	0.75	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc64	0.01	0.01	0.15	0.75	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc65	0.01	0.01	0.01	0.89	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc66	0.01	0.01	0.01	0.29	0.01	0.01	0.12	0.01	0.01	0.5	9
Doc67	0.01	0.01	0.01	0.28	0.01	0.01	0.12	0.01	0.01	0.5	9
Doc68	0.01	0.01	0.01	0.27	0.01	0.01	0.65	0.01	0.01	0.01	6
Doc69	0.01	0.01	0.01	0.26	0.01	0.01	0.65	0.01	0.01	0.01	6
Doc70	0.01	0.01	0.01	0.51	0.01	0.01	0.01	0.01	0.38	0.01	3
Doc71	0.01	0.01	0.01	0.51	0.01	0.01	0.01	0.01	0.37	0.01	3
Doc72	0.01	0.01	0.01	0.9	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc73	0.01	0.01	0.01	0.9	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc74	0.01	0.01	0.01	0.69	0.01	0.01	0.21	0.01	0.01	0.01	3
Doc75	0.02	0.02	0.27	0.61	0.02	0.02	0.02	0.02	0.02	0.02	3
Doc76	0.01	0.01	0.01	0.37	0.01	0.01	0.53	0.01	0.01	0.01	6
Doc77	0.01	0.01	0.01	0.35	0.01	0.01	0.55	0.01	0.01	0.01	6

Doc78	0.02	0.02	0.36	0.48	0.02	0.02	0.02	0.02	0.02	0.02	3
Doc79	0.01	0.01	0.01	0.53	0.01	0.01	0.4	0.01	0.01	0.01	3
Doc80	0.01	0.01	0.01	0.53	0.01	0.01	0.4	0.01	0.01	0.01	3
Doc81	0.01	0.01	0.01	0.54	0.01	0.01	0.4	0.01	0.01	0.01	3
Doc82	0.01	0.01	0.01	0.9	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc83	0.01	0.01	0.01	0.9	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc84	0.44	0.01	0.01	0.31	0.01	0.01	0.17	0.01	0.01	0.01	0
Doc85	0.43	0.01	0.01	0.31	0.01	0.01	0.17	0.01	0.01	0.01	0
Doc86	0.01	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc87	0.01	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc88	0.01	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc89	0.01	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc90	0.01	0.01	0.01	0.91	0.01	0.01	0.01	0.01	0.01	0.01	3
Doc91	0.01	0.01	0.01	0.01	0.01	0.01	0.87	0.01	0.01	0.01	6
Doc92	0.01	0.01	0.01	0.01	0.01	0.01	0.87	0.01	0.01	0.01	6
Doc93	0.01	0.01	0.09	0.46	0.38	0.01	0.01	0.01	0.01	0.01	3
Doc94	0.01	0.01	0.09	0.46	0.38	0.01	0.01	0.01	0.01	0.01	3
Doc95	0.56	0.01	0.01	0.35	0.01	0.01	0.01	0.01	0.01	0.01	0
Doc96	0.56	0.01	0.01	0.34	0.01	0.01	0.01	0.01	0.01	0.01	0
Doc97	0.01	0.01	0.01	0.56	0.01	0.01	0.34	0.01	0.01	0.01	3
Doc98	0.02	0.02	0.02	0.02	0.86	0.02	0.02	0.02	0.02	0.02	4
Doc99	0.02	0.02	0.02	0.02	0.86	0.02	0.02	0.02	0.02	0.02	4
Doc100	0.02	0.02	0.02	0.02	0.86	0.02	0.02	0.02	0.02	0.02	4
Doc101	0.01	0.01	0.01	0.27	0.01	0.01	0.12	0.01	0.51	0.01	8
Doc102	0.01	0.01	0.01	0.27	0.01	0.01	0.12	0.01	0.52	0.01	8
Doc103	0.01	0.01	0.01	0.27	0.01	0.01	0.12	0.01	0.52	0.01	8
Doc104	0.02	0.02	0.02	0.45	0.02	0.41	0.02	0.02	0.02	0.02	3
Doc105	0.02	0.02	0.02	0.02	0.84	0.02	0.02	0.02	0.02	0.02	4
Doc106	0.02	0.02	0.02	0.02	0.84	0.02	0.02	0.02	0.02	0.02	4
Doc107	0.02	0.02	0.02	0.02	0.83	0.02	0.02	0.02	0.02	0.02	4
Doc108	0.02	0.02	0.02	0.02	0.84	0.02	0.02	0.02	0.02	0.02	4
Doc109	0.02	0.02	0.02	0.02	0.84	0.02	0.02	0.02	0.02	0.02	4
Doc110	0.02	0.02	0.02	0.02	0.02	0.8	0.02	0.02	0.02	0.02	5
Doc111	0.02	0.02	0.02	0.02	0.02	0.8	0.02	0.02	0.02	0.02	5
Doc112	0.02	0.02	0.02	0.02	0.02	0.81	0.02	0.02	0.02	0.02	5
Doc113	0.02	0.02	0.02	0.02	0.02	0.81	0.02	0.02	0.02	0.02	5
Doc114	0.02	0.02	0.38	0.5	0.02	0.02	0.02	0.02	0.02	0.02	3
Doc115	0.01	0.01	0.01	0.62	0.23	0.01	0.06	0.01	0.01	0.01	3
Doc116	0.01	0.01	0.01	0.34	0.01	0.01	0.16	0.01	0.01	0.41	9
Doc117	0.35	0.01	0.01	0.41	0.01	0.01	0.16	0.01	0.01	0.01	3
Doc118	0.38	0.01	0.01	0.39	0.01	0.01	0.15	0.01	0.01	0.01	3
Doc119	0.01	0.01	0.01	0.88	0.01	0.01	0.01	0.01	0.01	0.01	3

EK 3: LDA Konu Kelime Dağılımı

Tablo Ek 3. LDA Konu Kelime Dağılımı

	Konu 0	Konu 1	Konu 2	Konu 3	Konu 4	Konu 5	Konu 6	Konu 7	Konu 8	Konu 9
Kelime 0	inocul	steril	rosett	gene	pi	flc	infect	box	cold	circadian
Kelime 1	Jasmon	Cultiv	Night	Cell	Ler	Fri	Leav	Sa	Acclim	Cyt
Kelime 2	agrobacterium	seed	light	plant	flower	columbia	day	pm	subzero	nicotinamid
Kelime 3	edr1	Subject	Media	Use	Shift	Flower	Week	İmpact	Toler	Seed
Kelime 4	stamen	deionis	murashig	time	long	flc3	arabidopsi	flat	ice1	toc
Kelime 5	ambient	yellow	skoog	respons	day	delet	thaliana	perciv	temperatur	oscil
Kelime 6	befor	bay0	week	arabidopsi	wild	introgress	pathogen	er	rschew	wild
Kelime 7	Temperatur	Wash	Minut	Express	Type	Long	Replic	Fulli	Tenela	Type
Kelime 8	ga	auxin	leav	plate	comparison	shift	collect	moder	access	seedl
Kelime 9	shift	water	tween	root	salt	frigida	elong	innoculum	freez	c24
Kelime 10	cultur	liquid	standard	day	short	sf2	independ	orontii	columbia0	period
Kelime 11	powderi	treatment	oxygen	transcript	mutant	locu	drought	reproduct	import	light
Kelime 12	mildew	fei0	ethanol	rna	no3	allel	inocul	heavi	chill	abund
Kelime 13	visibl	sha	continu	mutant	phytagel	day	mock	golovinomyc	idenpend	plate
Kelime 14	antisens	bur0	rel	stress	ft2	synchron	spore	mildew	avail	identifi

EK 4: Jaccard, LSA, PLSA ve LDA Yöntemleri AUC Sonuçları

Tablo Ek 4. Jaccard, LSA, PLSA ve LDA Yöntemleri AUC Sonuçları

Sorgu Deneyi	Jaccard	LSA	PLSA	LDA
GSE10016_1	0.624	0.529	0.607	0.418
GSE10016_2	0.657	0.638	0.753	0.596
GSE10016_3	0.608	0.476	0.586	0.419
GSE10464_1	0.525	0.733	0.696	0.782
GSE10464_2	0.595	0.67	0.74	0.831
GSE10502	0.512	0.769	0.718	0.606
GSE10876_1	0.547	0.643	0.467	0.695
GSE10876_2	0.525	0.673	0.608	0.491
GSE1110_1	0.563	0.636	0.659	0.582
GSE1110_2	0.592	0.54	0.608	0.589
GSE1111	0.494	0.639	0.611	0.578
GSE13739_1	1	1	1	1
GSE13739_2	0.524	0.661	0.599	0.541
GSE15876_1	0.527	0.773	0.759	0.728
GSE15876_2	0.371	0.686	0.789	0.543
GSE16143_1	0.469	0.669	0.672	0.435
GSE16143_2	0.415	0.527	0.556	0.644
GSE1766	0.71	0.715	0.701	0.652
GSE18624_1	0.632	0.686	0.594	0.646
GSE18624_2	1	1	1	1
GSE18975_1	0.565	0.669	0.792	0.731
GSE18975_2	0.548	0.71	0.606	0.662
GSE18975_3	0.588	0.605	0.61	0.624
GSE18975_4	0.543	0.644	0.623	0.707
GSE18975_5	0.678	0.778	0.528	0.732
GSE18975_6	0.881	0.895	0.598	0.827
GSE18975_7	0.594	0.689	0.632	0.61
GSE18985_1	0.583	0.703	0.702	0.668
GSE18985_2	0.58	0.665	0.674	0.672
GSE19261_1	0.603	0.724	0.809	0.738
GSE19261_2	0.741	0.796	0.862	0.769
GSE19263_1	0.528	0.647	0.509	0.754
GSE19263_2	0.434	0.566	0.552	0.746
GSE19264_1	0.996	1	1	0.996
GSE19264_2	0.777	0.927	0.917	0.842
GSE19264_3	0.588	0.793	0.609	0.683
GSE19265	0.478	0.451	0.58	0.682
GSE19266_1	0.669	0.697	0.721	0.647
GSE19266_2	0.563	0.706	0.76	0.595

GSE19271_1	0.727	0.911	0.907	0.975
GSE19271_2	0.669	0.804	0.572	0.7
GSE19271_3	0.926	0.946	0.892	0.838
GSE19271_4	0.727	0.911	0.907	0.975
GSE19271_5	0.504	0.562	0.459	0.562
GSE19271_6	0.404	0.566	0.699	0.487
GSE19271_7	0.677	0.891	0.596	0.827
GSE19700_1	1	1	1	1
GSE19700_2	1	1	1	1
GSE20044_1	0.628	0.708	0.486	0.588
GSE20044_2	1	1	1	1
GSE21684_1	1	1	1	1
GSE21684_2	0.835	0.962	0.856	0.941
GSE21684_3	0.578	0.777	0.631	0.65
GSE22274_1	0.624	0.567	0.488	0.548
GSE22274_2	0.541	0.618	0.606	0.527
GSE25171_1	0.992	0.992	0.992	0.992
GSE25171_2	0.626	0.543	0.529	0.542
GSE25171_3	0.992	0.992	0.992	0.992
GSE26679_1	0.626	0.693	0.58	0.574
GSE26679_2	0.439	0.479	0.535	0.455
GSE27281	1	1	1	1
GSE29642_1	0.996	0.996	0.996	0.996
GSE29642_2	0.996	0.996	0.996	0.996
GSE30097_1	1	1	1	1
GSE30097_2	0.457	0.658	0.696	0.627
GSE30098	0.429	0.53	0.659	0.623
GSE30223_1	0.996	1	1	1
GSE30223_2	0.483	0.635	0.59	0.575
GSE30398_1	1	1	1	1
GSE30398_3	1	1	1	1
GSE3326_1	0.42	0.709	0.508	0.612
GSE3326_2	0.724	0.842	0.736	0.777
GSE3350_1	0.535	0.545	0.672	0.563
GSE3350_2	0.501	0.553	0.64	0.551
GSE34081	1	1	1	1
GSE3416	0.502	0.557	0.513	0.629
GSE35325_1	0.572	0.724	0.729	0.505
GSE35325_2	0.499	0.612	0.612	0.433
GSE3959	0.611	0.536	0.517	0.579
GSE39597_1	1	1	1	1
GSE39597_2	0.996	1	1	1
GSE39597_3	0.996	1	1	1
GSE40140_1	0.783	0.845	0.869	0.746

GSE40140_2	0.642	0.821	0.815	0.782
GSE4116_1	0.996	1	1	1
GSE4116_2	0.996	1	1	1
GSE45183_1	1	1	1	1
GSE45183_2	1	1	1	1
GSE45183_3	1	1	1	1
GSE45183_4	1	1	1	1
GSE45183_5	1	1	1	1
GSE45215_1	0.586	0.633	0.542	0.478
GSE45215_2	0.72	0.79	0.77	0.694
GSE46208_1	1	1	1	1
GSE46208_2	1	1	1	1
GSE4733_1	0.782	0.664	0.667	0.678
GSE4733_2	1	1	1	1
GSE47981	1	1	1	1
GSE55140_1	0.996	1	1	1
GSE55140_2	1	1	1	1
GSE55140_3	0.996	1	1	1
GSE55835_1	0.663	0.757	0.675	0.637
GSE55835_2	0.536	0.729	0.592	0.549
GSE55835_3	1	1	1	1
GSE5747	0.598	0.521	0.532	0.616
GSE576_1	0.996	1	1	1
GSE576_2	1	1	1	1
GSE576_3	0.399	0.428	0.516	0.621
GSE576_4	0.998	1	0.992	0.987
GSE576_5	0.448	0.467	0.489	0.592
GSE577_1	0.999	1	0.994	0.997
GSE577_2	0.519	0.526	0.596	0.796
GSE577_3	0.993	0.997	0.966	0.994
GSE577_4	0.996	1	1	1
GSE6349	0.577	0.545	0.66	0.607
GSE7642	1	1	1	1
GSE8365	0.488	0.694	0.487	0.518
GSE9674_1	0.488	0.611	0.489	0.524
GSE9674_2	0.615	0.666	0.532	0.535
GSE9996	0.875	0.892	0.939	0.828
ortalama AUC	0.73	0.79	0.771	0.768