BAŞKENT UNIVERSITY

INSTITUTE OF SCIENCE AND ENGINEERING

EXPERIMENT RETRIEVAL IN GENOMIC DATABASES

DUYGU DEDE ŞENER

DOCTOR OF PHILOSOPHY THESIS

2019

EXPERIMENT RETRIEVAL IN GENOMIC DATABASES

GENOMİK VERİ TABANLARINDA DENEY GERİ GETİRİMİ

DUYGU DEDE ŞENER

Thesis submitted
in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy
in Department of Computer Engineering
at Başkent University

2019

This thesis, titled: “EXPERIMENT RETRIEVAL IN GENOMIC DATABASES”, has been approved in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY IN COMPUTER ENGINEERING, by our jury, on 15/01/2019.

Chairman	: Prof. Dr. Mehmet Reşit TOLUN
Member (Advisor)	: Prof. Dr. Hasan OĞUL
Member	: Prof. Dr. Nizami GASİLOV
Member	: Prof. Dr. Tolga CAN
Member	: Assoc. Prof. Dr. Yeşim AYDIN SON

APPROVAL
/ /2019

Prof. Dr. Ömer Faruk ELALDI
Director
Institute of Science and Engineering

BAŞKENT ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ
DOKTORA TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 28/01/2019

Öğrencinin Adı, Soyadı: Duygu DEDE ŞENER
Öğrencinin Numarası: 21310018
Anabilim Dalı: Bilgisayar Mühendisliği
Programı: Doktora
Danışmanın Unvanı/Adı, Soyadı: Prof. Dr. Hasan OĞUL
Tez Başlığı: Experiment Retrieval in Genomic Databases

Yukarıda başlığı belirtilen Doktora tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 82 sayfalık kısmına ilişkin, 28/01/2019 tarihinde şahsım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %10 'dur.

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
28/01/2019

Prof. Dr. Hasan OĞUL

ACKNOWLEDGEMENT

First and foremost I owe the deepest gratitude to my advisor Prof. Dr. Hasan Oğul for the continuous support of my study, for his patience, motivation and immense knowledge. His guidance always helped me in all of time of research and writing of this thesis.

I would also like to thank my committee members Prof. Dr. Nizami Gasilov and Assoc. Prof. Dr. Yeşim Aydın Son for their support along the thesis study.

I would like to thank our project collaborators Prof. Dr. Giovanni Felici and Dr. Daniele Santoni for their valuable ideas and contributions.

I am very grateful to my colleagues Didem Ölçer, Tülin Erçelebi Ayyıldız, Nihal Uğur, Buket Ünal, Oğul Göçmen, Koray Açıcı, Çağatay Berke Erdaş, Tunç Aşuroğlu and Mehmet Dikmen. You always have supported and encouraged me.

A special thanks to my family. I am grateful to you especially my mother, father and brother who have provided me through moral and emotional support in my life.

To my beloved husband Çağrı Şener, thank you for supporting me and believing in me. There is no difficulty we can’t go through with you.

And my 15 months old daughter Ela, I am so lucky to be your mother. I have been overcome all difficulties with your love.

This thesis study was supported by TUBITAK projects grant number 11E527 and 215E369.

ABSTRACT

EXPERIMENT RETRIEVAL IN GENOMIC DATABASES

Duygu DEDE ŞENER
Başkent University Institute of Science and Engineering
Department of Computer Engineering

Genomic data can be found in different formats such as experimental measurements, sequences, networks. Due to the rapid growth of such data in genomic repositories, retrieving relevant experiments has become an important issue to be addressed by researchers. To search an experiment through the databases, users generally use textual meta-data such as organism name, description, author, but this type of search is insufficient to represent the overall content of the experiment. Content-based search strategy has become an alternative solution for retrieving relevant experiments from huge data collections. This thesis study aims to develop retrieval models for different data types to find relevant experiments in genomic databases. The study has two main parts: time-series experiment retrieval framework and whole-metagenome sequencing sample retrieval framework. In the first part, different fingerprinting techniques and comparison metrics were used to retrieve relevant time-series experiments. The originality of this part consists in its attempt for taking gene expression profiles over the entire time points as a query and retrieving relevant samples from the data repository. The second part consists of developing a content-based retrieval framework for whole-metagenome sequencing samples. The framework involves different fingerprinting, feature selection methods and similarity measurements for a given data set. The main contribution of the study is extracting fingerprints based on two text mining methods. The experimental results showed that the proposed models have been successful in finding relevant experiments for genomic data in different formats. Experimental results also encourage the use of the proposed models in current database implementations.

KEYWORDS: Genomic database; gene expression database; time-series; content based search; information retrieval; fingerprinting; Arabidopsis; whole-metagenome sequencing.

Advisor: Prof. Dr. Hasan OĞUL

ÖZ

GENOMİK VERİ TABANLARINDA DENEY GERİ GETİRİMİ

Duygu DEDE ŞENER
Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Genomik veri; deneysel ölçüm, sekans verileri, ağ yapıları gibi farklı formatlarda saklanmaktadır. Genomik veri tabanlarında saklanan bu tür verilerin son yıllardaki hızlı artışı, deneylerin geri getirimi konusundaki ihtiyaçları gündeme getirmektedir. Kullanıcılar, veri tabanında bir deneyi ararken genellikle metin-tabanlı arama tekniğini kullanmaktadırlar. Fakat bu teknik, deney içeriğini temsil etmede yetersiz kaldığı için yeni yöntemlere ihtiyaç duyulmaktadır. Bu ihtiyaç doğrultusunda, içerik tabanlı arama yöntemleri benzer deneylerin geri getiriminde kullanılan alternatif yöntem olmuştur. Bu tez, farklı türlerde olan genomik verilerin veritabanlarında aranabilmesini sağlayan geri getirim modellerinin tasarımını amaçlayan bir çalışmadır. Çalışma, zaman serisi deney geri getirimi, bütün metagenom sekanslama örneklemlerinin geri getirimi olmak üzere iki temel kısımdan oluşmaktadır. Birinci kısım, zaman serisi deneylerin geri getirimi için farklı imza yöntemlerinin ve uygun benzerlik metriklerinin uygulanmasını içermektedir. Bu çalışma zaman serisi deneyinin tümünü sorgu olarak alan ve arama yapan ilk çalışma olma özelliğini taşımaktadır. İkinci kısımda ise, tüm metagenom sekanslama deneylerinin geri getirimi için farklı imza yöntemlerini, özellik seçim algoritmalarını ve benzerlik metriklerini içeren bir içerik tabanlı arama altyapısı geliştirilmiştir. Çalışmanın temel katkısı, deney imzalarını oluşturmada iki farklı veri madenciliği yönteminin kullanılmasıdır. Deneysel sonuçlar, geliştirilen modellerin benzer deneyleri bulmada başarılı olduklarını göstermektedir. Ayrıca, sonuçlar geliştirilen bu modellerin mevcut veri tabanı uygulamalarında kullanımları konusunda umut vaat etmektedir.

ANAHTAR KELİMELER: Genomik veri tabanı; gen ifade matrisi; zaman serisi veri; içerik tabanlı arama; bilgi geri getirimi; imza çıkarımı; arabidopsis; metagenom dizilim.

Danışman: Prof. Dr. Hasan OĞUL

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IR	Information Retrieval
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
TF	Transcription Factor
rRNA	Ribosomal Ribonucleic Acid
GEO	Gene Expression Omnibus
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
WMS	Whole-Metagenome Shotgun
DE	Differentially Expressed
LE	Lyapunov Exponent
GSEA	Gene Set Enrichment Analysis
ES	Enrichment Score
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
TPR	True Positive Rate
FPR	False Positive Rate
FS	Feature Selection
TF-IDF	Term Frequency Inverse Document Frequency
CAE	Correlation Attribute Evaluation
LSA	Latent Semantic Analysis
SVD	Singular Value Decomposition
LDA	Latent Dirichlet Allocation
JSD	Jensen Shannon Divergence
MSA	Multiple Sequence Alignment
CORE	Consistency of the Overall Residue Evaluation
TUBITAK	Türkiye Bilimsel ve Teknik Araştırma Kurumu

1. INTRODUCTION

This thesis study consists of four main chapters. Chapter 1 gives motivation and purpose of the study, terminology and background information, Chapter 2 describes different fingerprint extraction methods and convenient similarity metrics for retrieving time-series experiments, Chapter 3 consists of fingerprint extraction, feature selection and comparison approaches for whole metagenome sequencing sample retrieval. The final chapter is devoted to conclusion and future work.

1.1 Motivation and Purpose of the Study

In recent years, developments in biotechnology and computational biology lead to rapid growth in the accumulation of genomic data in public databases. The databases store the genomic data in various formats such as experimental measurements, sequence samples, structures or networks. Accessing and analyzing this type of data is one of the main tasks for the researchers and users. Researchers need to obtain biological knowledge from the data to produce new hypotheses to be applied in computational studies in their research field. The obtained data may be used in application of medical practices such as treatment for a specific disease or discovery of a new drug. Besides this, users expect to access the genomic data faster through efficient searching tools to make easy their lives. In this respect, there is a significant need for accessing and searching the data in the related repositories. Therefore, developing efficient retrieval models has become a popular research effort for researchers. Currently, meta-data based or keyword-based search is commonly used in large repositories. In this type of search, experiments are annotated by descriptive labels such as experiment name, author of the study, organism name and unfortunately users have limited searching options related with these labels. This case may cause some searching problems; because searching results highly depend on accessibility and accuracy of user-defined annotations. For instance, annotations may be missing or incorrect, because a user or a database administrator provides these labels and they may make some mistakes in filling the information of the experiment. In addition to this, these data may not represent the overall content of the searched data, so user requirements could not meet by the retrieval system. In this regard, new searching approaches are needed to build more representative queries to search an object in an efficient manner. Latest trend to overcome these problems is using query-by-example or content-based searching techniques rather than traditional meta-data techniques. In recent years, content-based search term has become popular for experiment retrieval in biological and biochemical sciences as in other research fields.

In this thesis study, developing content-based retrieval models for genomic data is mainly focused. It is aimed to develop retrieval models by using different data types and perspectives. The study has two main contents which are a retrieval framework for time-series experiments and a retrieval framework for whole-metagenome sequencing experiments. Retrieval processes consist of designing and development of targeted sub-models, creation of suitable comparing mechanisms, evaluating the proposed models with real datasets.

1.2 Information Retrieval: Terminology and Background

In the most general sense, information retrieval (IR) is defined as the study of obtaining relevant material in an unstructured form from data collections. Unstructured data represents raw and unorganized data type, while structured data refers to information, usually in text format, which can be organized and processed easily by data mining tools. Storing, organizing and searching information from the resources are the main tasks of the IR systems. The rapid expansion of the global resources of knowledge and use of web contents has made these tasks difficult to achieve. Furthermore, users expect to access knowledge faster by using more effective tools. In this respect, developing searching approaches in an efficient manner has become a basic research interest in IR field [1].

The main objective in IR is retrieving more relevant objects than irrelevant objects with the query object. Meta-data based search strategy is generally used in most of search engines. Meta-data, which are descriptive annotations such as name of the object, author of the study or any user-specified label, gives detailed information of the object to be searched. Although, it provides pretty much significant information about the object; it may be insufficient to represent the overall content of the object. In addition, annotations are generated by the users, so they may make some mistakes in filling the required information fields or some fields may be incomplete. This causes some searching problems. To handle these problems, it has been recommended to use query-by-example or content-based search approach. In this type of search, searched object is provided as a query instead of submitting any keyword to retrieve relevant objects with the query. Similarity between objects is calculated based on content similarity of query object and other objects in the repository. There are two main processes in content-based search strategy; creating fingerprints for representation of the object content and comparing these fingerprints with a related comparison metric in an efficient way. Different fingerprinting approaches have been used with respect to the representation of content of the object. Information retrieval, fingerprinting approach is defined as term of index. It allows representing the object content without need of any metadata in database search. Deriving a representative fingerprint and comparing these fingerprints in an efficient way are two main goals of a successful content-based search implementation. The key question in content-based search strategies is to find an approach to derive a representative fingerprint from the given object.

1.3 Biological Terminology and Background

DNA, or deoxyribonucleic acid, is the main component of all living organisms. It stores basic information of all cellular functions of organisms. It consists of four nucleotide bases named as Adenine (A), Guanine (G), Cytosine (C), and Thymine (T) in its double helix structure. The information stored in DNA depends on the order or sequences of those bases and the information is used to build different types of cells of an organism. Chromosomes are thread-like molecules that contain hereditary information of the organism (Figure 1.1). The chromosomes consist of long chains of DNA and related proteins. Moreover, a gene is a heredity element composed of DNA segments to store the information to build and maintain cells of an organism. It includes sequence of nucleotides on a given chromosome which codes a specific protein as given in the figure. Genome is defined as completed set of DNA that contains all of its genes. In addition to this, genomics is the study of genome characteristics associated with the organism and it has valuable knowledge about organisms. Genomic data has been gathered with various technologies and stored in different formats such as gene expression data, sequence data or networks.

Gene expression is defined as the synthesis of gene products, e.g protein, by the information provided genetic instructions in the cell. The expression levels of thousands of genes are measured simultaneously with the DNA microarray technology. The microarray technology helps researchers to understand fundamental units of life as well as to discover genetic causes of diseases occur in living organisms. Gene expression data are stored in matrices in which rows refer to expression levels of genes; columns represent samples or conditions such as environmental conditions or time points. (Figure 1.2). Time-series, so-called time-course, gene expression data represents the changes of gene expression measurements over a time period. Time-series data is stored as matrices in which rows represent genes; columns represent time range or period. A gene is defined as differentially expressed when its expression levels between two conditions changes significantly. Differentially expressed genes are genes whose expression levels are related with a factor such as a treatment, drug or a clinical outcome.

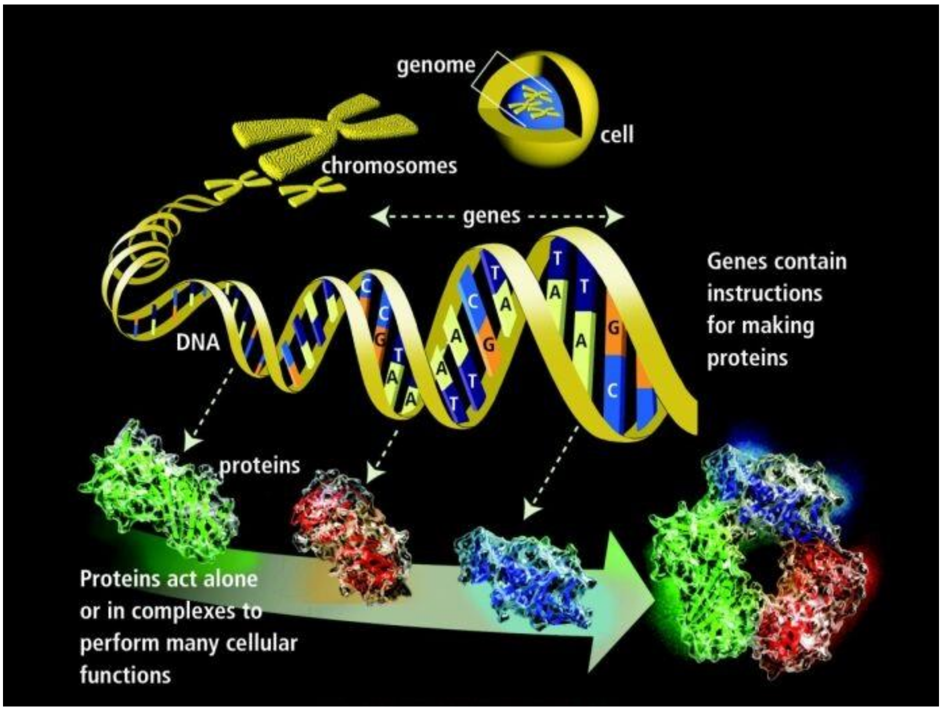

genome
cell
chromosomes
genes
Genes contain instructions for making proteins
DNA
proteins
Proteins act alone or in complexes to perform many cellular functions

Figure 1.1 Cell structure of a living organism[1]

[1] Quoted from TITILADE, Popoola Raimot and OLALEKAN, Elegbede Isa, The Importance of Marine Genomics to Life, *Journal of Ocean Research*, Vol. 3, no.1, p.1–13, 2015.

Differential expression of a gene is used to characterize its behavior. These profiles are generated for each experiment to represent their gene expression matrix as a single vector. Those profiles are used in database search instead of using whole gene expression matrices, so the computational efficiency can be reduced.

GEO (Gene Expression Omnibus) [2], ArrayExpress [3] and GenBank [4] are widely used public repositories that allow storing, retrieving and organizing functional genomic data. GEO was launched by NCBI (National Center for Biotechnology Information) in 2000 to provide gene expression datasets for researchers. Over 650.000 submissions has been hold in GEO. ArrayExpress has been used since 2002 and it consists of data from >50000 hybridizations and >1500 000 individual expression profiles. Furthermore, it has also two main parts called ArrayExpress Repository and ArrayExpress Data Warehouse. Moreover, GenBank is one of the most popular sequence databases that contain over 55.000 sequences from different organisms.

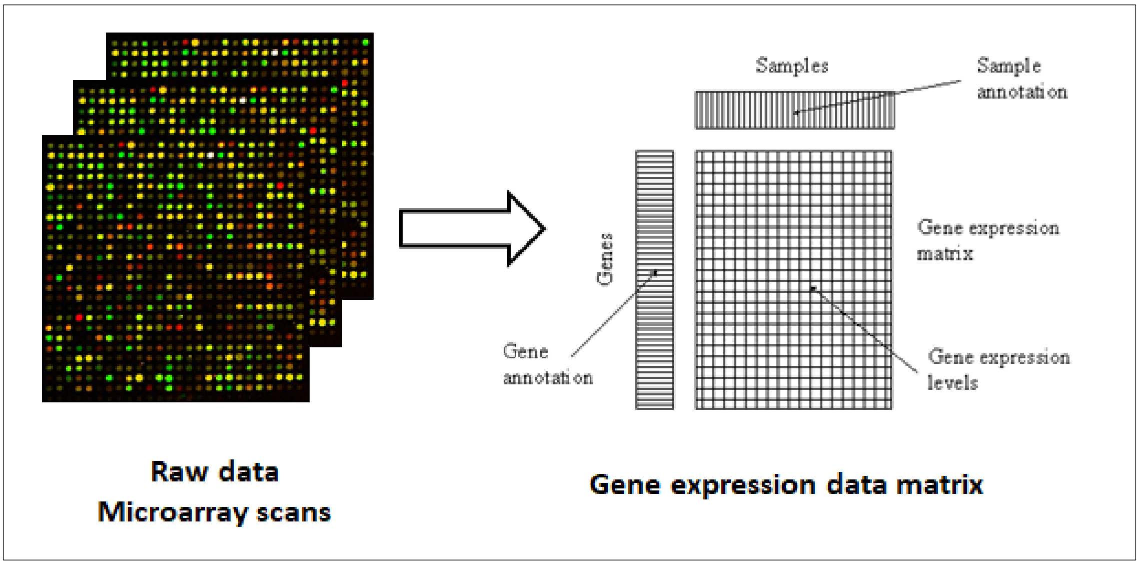


Figure 1.2 Obtaining gene expression data matrix from a collection of raw microarray data

Genes may consist of information about diseases. This information can be derived from a single gene or relationships among many genes. However, some diseases or phenotypic disorders cannot be expressed by individual gene. Beside this, genes generally work together like a piece of whole. In this regard, identifying gene sets or groups has become a major focus to interpret biological knowledge in biomedical research area. Gene sets are gene groups obtained based on biological knowledge. Obtaining biologically significance gene sets provide some specific information about biological pathways, protein-protein interactions or functionally related genes.

Metagenomics is discovering genetic content of microorganisms from different environmental samples with using bioinformatics tools and genomic technologies. [5]. Chen and Pachter defined metagenomics as “the application of modern genomics technique without the need for isolation and lab cultivation of individual species”. Metageomic data provide valuable information about organisms, so analyzing this data has become a significant research interest recently. This interest leads to some approaches raised for generating sequence data. There are two widely used sequencing approaches for generation of metagenomic samples. The first one is Sanger sequencing in which DNA is copied into plasmids and determination of the sequences is completed through the chain termination method. In second method, instead of DNA cloning, one of the next-generation sequencing (NGS) approaches, also called high-throughput sequencing, is used to obtain sequence reads. Although longer sequence reads can be generated by Sanger sequencing, it has some disadvantages based on the cloning process. On the other hand, NGS has a lower error rate than Sanger sequencing [5]. However, there are recent developments in NGS technology and huge amount of sequence data has been generated, using whole metagenome shotgun (WMS) sequencing in analyzing huge data collection is a more efficient way to get accurate information. Furthermore, targeted studies perform analysis such as phylogenetic profiling with a lower cost, while information about metagenomics can be obtained by WMS sequencing data analysis. Moreover, development of new analysis approaches to discover knowledge from organisms can be done with this method easily.

1.4 Statistical Significance Tests

Statistical significance tests are used to show that observed results are not occurred randomly; instead they are based on some statistical facts. These tests have become a quite important step in data analysis for various academic disciplines such as medicine, economics or computational biology.

Hypothesis testing, also called p-value approach, refer to defining research hypothesis or an observable event as a null and alternate hypothesis. The null hypothesis (H_0), which is opposite of the alternate hypothesis (H_1), is defined as a hypothesis that can be rejected or nullified. The null hypothesis claims that there are no statistical significance between given observations. The level of statistical significance of observed results is defined by p-value approach. In this approach, a probability that given null hypothesis is true is calculated. When a p-value less than or equal to 0.05 is obtained, the null hypothesis can be rejected, in other words the alternate hypothesis is accepted.

Statistical procedures are used for determining whether the difference between observations is zero. Statistical procedures have two hypotheses called the null hypothesis and the alternative hypothesis defined below. The former claims that difference between observations is zero, while the latter assumes that the difference is not zero. Paired t-test and Wilcoxon signed-rank test are widely used statistical procedures developed for discovering difference between observed results. In this study, these tests were used to show statistical significance between performances of fingerprinting approaches. When applying both of these tests the null hypothesis (H_0) and alternate hypothesis (H_1) are defined as follow;

$$H_0 = There\ is\ no\ difference\ between\ performances\ of\ used\ fingerprinting\ approaches.$$

$$H_1 = There\ is\ a\ difference\ between\ performances\ of\ used\ fingerprinting\ approaches.$$

The main goal is rejecting null hypothesis. Paired t-test, so-called the dependent sample t-test, is a widely used statistical procedure to analyze difference between observations. In this test, each object is measured twice such as case-control studies.

$$s_{\bar{x}} = \frac{s_{diff}}{\sqrt{n}} \tag{1.1}$$

$$t = \frac{\bar{x}_{diff}}{s_{\bar{x}}} \tag{1.2}$$

Let two given observations are represented by X and Y; each individual observation is given as x_i and y_i and total number of observations is n. In this test, difference between each pair is calculated, then mean difference ($\bar{x}_{diff}$) and standard deviation (s_{diff}) of the differences are obtained. Standard error for the mean difference $s_{\bar{x}}$ (1.1) is evaluated using the standard deviation. Then t-statistic (1.2) is calculated and obtained value is compared with the critical value from the t-distribution table. According to the value from the table, a p-value is obtained which is used for rejecting or accepting the null hypothesis specified before [6].

Wilcoxon signed-rank test is a non-parametric test which is alternative to paired-t test. In this test, firstly null hypothesis and a hypothesized value (in our case this value is 0) for comparison are defined. Paired score differences are calculated, and then ascending order of absolute value of the difference are obtained. Unlike the paired t-test, Wilcoxon test use those ranked values. If there are two observations that are equal to hypothesized value, the test ignore them [7].

$$S_+ = \sum_i^n \psi_i \, r|Z_i|, \; where\, \psi_i = \begin{cases} 1, \; Z_i > 0 \\ 0, \; Z_i < 0 \end{cases} \tag{1.3}$$

$$S_- = \sum_i^n \psi_i \, r|Z_i|, \; where\, \psi_i = \begin{cases} 0, \; Z_i > 0 \\ 1, \; Z_i < 0 \end{cases} \tag{1.4}$$

Difference between each observation shown as $Z_i = x_i - y_i$, $r|Z_i|$ is rank of absolute value of Z_i. Sum of the positive ranks is given as S_+ (1.3), while the sum of negative ranks is represented by S_- (1.4). After calculation of S_+ and S_-, the smaller one is selected and an appropriate p-value is calculated [8].

2. TIME SERIES EXPERIMENT RETRIEVAL

In this chapter, a content-based retrieval framework with suitable fingerprinting methods and comparison strategies for time-series microarray experiments is introduced. The chapter consists of three main parts. Motivation of the study and related work are defined in the first part. Methods are described in the second part and experimental results are given in the final part.

2.1 Introduction

Time-series gene expression data are obtained from microarray or similar experiments. They have been widely used to explore variety of genomic processes. Time-series gene expression data analysis is performed to observe variation of gene expression based on an environmental change or different time points. In this direction, there are basically two kind of approaches; mathematical approaches and network approaches for data analysis process. The former uses latent variables to model a gene behavior, while the latter further focuses on relationship between gene groups. There are plenty of studies based on the first approach to cluster genes [9–11], to classify gene profiles [12, 13] and to estimate expression using regression method [14]. The former approach consists of methods in which gene regulatory network is used to detect interactions in terms of the some environmental changes [15–17].

With the exponential growth of time-series experiments, data repositories to access the data has been increased recently. The increasing number of experiments in these repositories has created a fundamental need for retrieving biologically relevant experiments in an efficient way. Therefore, developing efficient retrieval models has become a popular research effort for researchers. Due to some searching problems of meta-data based search, there has been increasing interest about content-based search through gene expression repositories. There are two main processes in content-based search strategy; creating gene profiles for representation of the experiment content and comparing these profiles with a related comparison metric. Different approaches have been used with respect to the representation of experiment content. Some studies focus on co-expressed or differentially expressed gene list to obtain gene profiles while others obtain gene profiles by known gene-sets.

Content-based search approach has been widely used in searching through gene expression experiments in the data collection. The first study, proposed by Hunter et al [18] for content based search in gene expression databases, is a search tool named GEST (Gene Expression Search Tool). It compares two experiments using Bayesian-based similarity metric based on correlational structure and complex joint distributions of expression values. One experiment means a series of profile consists of more than one gene expression value at any condition. A simple algorithm called RaPiDS (Rapid Profile Database Search) to compare gene expression profiles is proposed by Horton et al. [19]. In their study, a profile means an experiment involves many genes. They use Spearman rank correlation (SRC) to calculate similarity for profile pairs. It has been shown that RaPiDS is a fast and efficient method for a reasonable sized database. Fujibuchi et al. [20] build a search engine named CellMontage using RaPiDS method. It is the first content-based search engine that detect similarities between expression profiles. A large number of microarray experiments were used to test system performance. GENE CHAnge browSER (GeneChaser) developed by Chen et al. [21] is a search engine for differentially expressed genes. It automatically analyzes given experiments and annotates them. The study consists of two search modules such as single gene search and multiple gene searches. In the former, any gene identifier is taken as input while in the latter function a gene list is given as a query then relevant gene list that contain differentially expressed genes with the query gene or gene list is obtained. In addition to this, Hibbs et al. [22] developed an algorithm named SPELL which is a web-based search procedure for large gene expression data. The proposed model retrieves genes that expressed together with the query genes and make some biological prediction. Engreitz et al. [23] proposed a content-based approach called ProfileChaser to retrieve gene expression experiments. A dimension reduction technique so called independent component analysis from their previous study [24] was used to enhance the speed of the experiment search. Reduced set of gene expression features are extracted by this transformation process, then differentially expression (DE) profile, that refers to changes in the expression level, for each experiment are generated. Finally, obtained profiles are compared by their novel weighted correlation coefficient. Bell and Sacan [25] use binary vector representation to retrieve gene expression experiments using content-based approach. In the study, it is showed that binary vector representation reduced the time needed for searching database. Besides that, in the study of Caldas et al. [26], an experiment is defined using gene sets and these gene sets are used as a query for searching process. The proposed retrieval model is based on representing experiments through the differential gene sets of each experiment. Suthram et al. [27] also used network-based gene-sets to obtain fingerprints for representing experiment content. The developed framework is used to compare and contrast diseases and they also identified functional modules in the human protein network. Georgii et al. [28] developed a retrieval framework which has targeted analysis at regulatory relationship of genes and regulatory model-based similarity measure. In addition to these studies, there is also a study that aims to propose a framework to discover relevant microRNA (miRNA) experiments through large data collections [29]. In order to detect differentially expressed miRNA profiles, they applied a normal-uniform mixture model and they developed a similarity metric to compare categorical fingerprints. Each miRNA experiment is represented by binary fingerprints that are vectors of differentially expressed of all the miRNAs given in the experiment. It is the first study developed for miRNA microarray experiment retrieval.

Current retrieval methods use different fingerprinting techniques and comparison strategies. Although, all methods provide valuable solutions for experiment retrieval, they considered that experiments have only two conditions such as control and treatment, so the proposed models cannot handle experiments with three or more conditions. In addition to this, there is pretty much time-series experiments in gene expression repositories. It is the fact that time-course experiments provide more depictive information especially for treatment studies. Unlike the mentioned studies above, Hayran et al. [30] used time-course content to build fingerprints for representing the experiments. They considered first and last time points to generate differential expression-based fingerprints, but time-course behavior should be defined using all time points in the retrieval process. To this end, a content-based retrieval framework, that takes into all time points for representing experiment content, was proposed in this chapter. The framework involves different fingerprinting techniques and comparison strategies. This study is the first approach that uses gene behavior across all time points in building fingerprints. The obtained results show that the proposed framework can retrieve biologically relevant experiments.

2.2 Methods

Four different fingerprint extraction methods and associated similarity metrics were used in the proposed retrieval system. This section consists of two subsections such as time-series fingerprint extraction methods and fingerprint comparison methods.

2.2.1 Time-series fingerprint extraction methods

In the proposed retrieval model, given in Figure 2.1, the first process is transforming experiment content into a representative fingerprint. Fingerprinting is a widely used technique to describe experiment content in a feature space. After transforming all experiments in the repository into a fingerprint, the next process is detecting similarity between obtained fingerprints through an appropriate comparison strategy. As can be seen from Figure 2.1, the system reports a ranked list of experiments which are similar to the query experiment based on a similarity score. Novelty of this study comes from using gene behaviors over all time points in translating time-series experiment into the fingerprints. Used fingerprint extraction methods are described in detail in the next section.

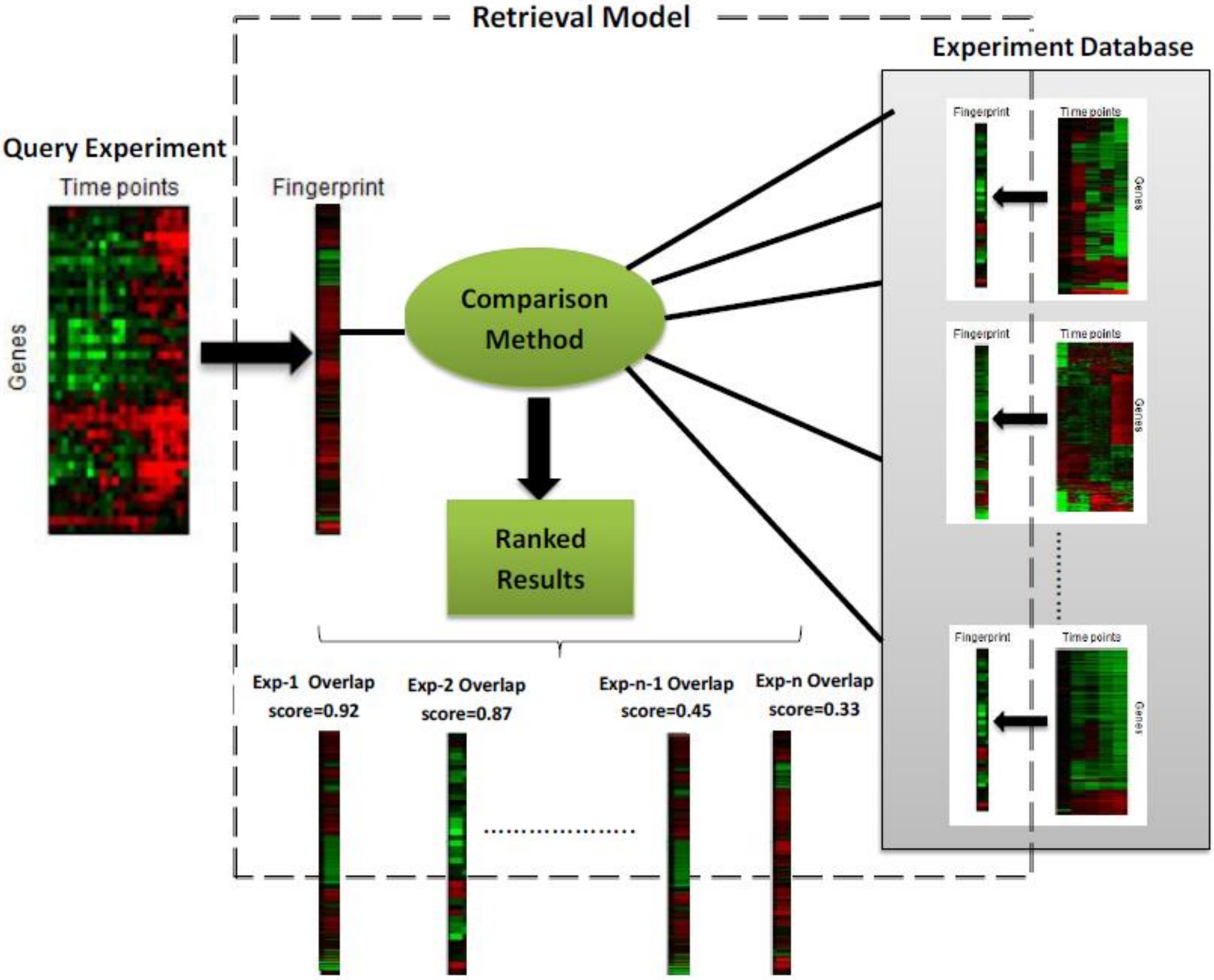


Figure 2.1 Overview of the proposed retrieval framework

2.2.1.1 Differentially expression profile-based method

Differentially expressed genes are genes that have expression levels changes significantly between two different samples or experimental conditions (normal and diseased cells etc.). In order to discover differentially expressed genes the ratio of expression level of a gene over two conditions is calculated. The calculated value, called log ratio, is a quantity for determining differential expression for a gene. In the “rule of two”, determining differentially expressed gene is stated as follows: The gene is considered as a differentially expressed gene, if its log ratio is greater than two or less than half [31]. The rule is the earliest uses of the quantity.

Discovering differentially expressed genes is one of the main goals of analyzing gene expression data to investigate causes of diseases and treatments of such diseases. Identifying and using differentially expressed genes in time-series data have been studied in various analysis techniques such as cluster analysis [32] and pointwise comparison [33]. In this study, an approach [34], called Normal Uniform Differential Gene Expression (NUDGE), was adapted to get the probabilities of genes being differentially expressed. The DE profiles represent the changes in the expression levels. The genes are modeled in two different groups such as differentially expressed and non-differentially expressed. To generate DE profiles the specified method is adapted into used time-series experiments.

$$r_i \sim p\, N(r_i|\mu,\sigma^2) + (1-p)U(r_i),\ i = 1,2\ldots,N \tag{2.1}$$

Each time-series experiment is represented by DE profile vectors. The DE of a gene i, called Z_{i}, is a measure of probability of the gene being differentially expressed between two conditions (first and last time point). The method aims estimating Z_{i} by fitting data into a normal-uniform mixture of flat and differentially expressed genes. The model formulization is given in the formula (2.1). In the formula, the observed normalized log ratio for gene i is shown by r_i, p denotes the prior probability of a gene being differentially expressed, $\mathrm{N}(\mathrm{r_i}|\mu,\sigma^2)$ is the Gaussian distribution with mean μ and variance σ^2 and $\mathrm{U}(\mathrm{r_i})$ is the uniform distribution on a finite interval and N is the number of genes.

The defined model is estimated by maximum likelihood method based on Expectation Maximization (EM) algorithm. The labels of genes are defined, $z_i, i = 1,\ldots,N$, in which if a gene is not differentially expressed z_i is , if it is z_i is 1. There are two steps in the algorithm; Expectation (E step) and Maximization (M step) step.

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

Firstly, the labels are estimated in iteration-k of E step as given in the formula (2.2).

$$\hat{p}^{(k)} = \frac{\sum_i(1-\widehat{z_\imath}^{(k)})}{N} \tag{2.3}$$

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

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

Then, the model parameters p, μ, and σ^2 are estimated in a maximization step (2.3, 2.4, 2.5). These steps are processed until a convergence is reached.

In order to generate DE profile vectors, rank-based binarization was used. The impact of the noise in raw data and processed instance generated by the normal-uniform mixture model can be decreased using the binary representation. Genes are listed in descending order according to the probability of differential expression. Genes which are located top k% on the list takes the value of 1, the rest takes the value of 0. This threshold was used to confirm that fixed percent of all genes are differentially expressed. To enhance the retrieval performance the value of k was set experimentally.

2.2.1.2 Transition model-based method

A fingerprint vector consists of different types of data e.g integer, float or categorical values. Time-course experiments generally have two or more time points. Representing a gene profile with a binary category such as differential or non-differential expression is not a sufficient way to represent these types of experiments, so different types of categories should be used in describing their profiles. To this end, a competent method developed by Sahoo et al. [35] was adapted into this study to organize time profiles. In this method, gene expression profiles are described by binary transitions of gene expression over time periods. As given in Figure 2.2, there are five transition models named as model 0, 1, 2, 3 and 4. Model 0 (Figure 2.2.a) represents no important changes in gene expression level during a time period. The one-step transition is shown by models 1 and 2. In model 1, the gene expression has increasing value from low to high (Figure 2.2.b), while in model 2 the expression value changes from high to low (Figure 2.2.c). Furthermore, two-step transitions are model 3 (Figure 2.2.d) and model 4 (Figure 2.2.e); in the former there is an increase followed by a decrease, in the latter one there is a decrease followed by an increase. However, gene expressions may be

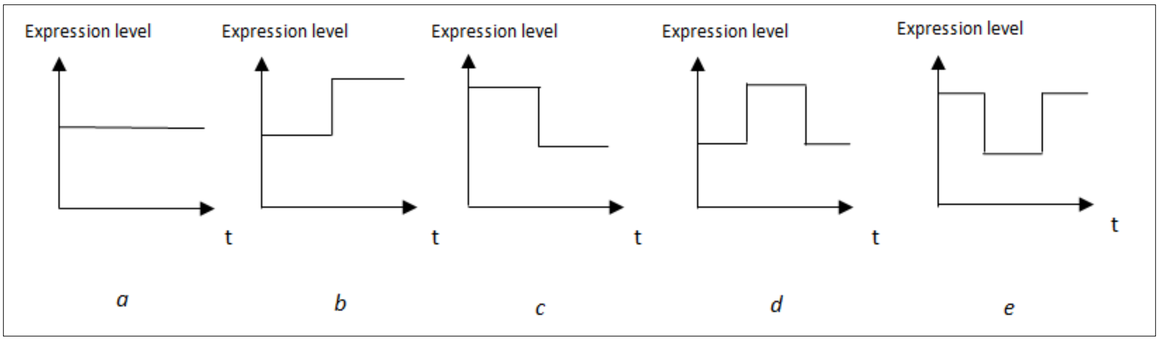


Figure 2.2 Transition models to represent expression levels of genes

(a) No change in expression

(b) Expression change from low to high

(c) Expression change from high to low

(d) Expression increase followed by a decrease

(e) Expression decrease followed by an increase

described by more than five transitions, in this study it is assumed that five models can accurately describe gene behaviors. Gene profiles are labeled by the model described above. Adaptive regression method is used in which one-step and two-step models are evaluated to select more convenient model that describe the data. All step positions are then assessed and the values of constant segments are calculated. Finally, to minimize the square error, collection of the step positions is performed.

$$SSE = \sum_{i=1}^{n}(E_i - \hat{E}_i)^2,\ \ SSR = \sum_{i=1}^{n}(\ \hat{E}_i - \overline{E})^2 \tag{2.6}$$

Gene expression values over n time points are shown by $E_{1,}E_2, \dots E_n$. Adjusted values of the adaptive regression are given as $\hat{E}_1$,$\hat{E}_2, \dots, \hat{E}_n$ and mean of the entire time points is depicted by $\overline{E}.$ In addition to this, SSE represents the sum of squares error, while the regression sum of squares are defined by SSR (2.6).

$$F = \frac{SSR/(m-1)}{SSE/(n-m)} \tag{2.7}$$

$$P = Pr(F_{n-m}^{m-1} > F) \tag{2.8}$$

For each transition model (one-step and two-step), a regression test statistic F_i (2.7) is described. The freedom degrees of SSE and SSR are represented by

$(m-1)$ and $(n-m)$ respectively. An F-distribution with those values follows the F-statistic; such as there is a random variable, named F_{n-m}^{m-1}, which has this distribution the corresponding P-value to the tail probability of this distribution is calculated as given in the formula (2.8).

$$F_{12} = \frac{(SSE_1 - SSE_2)/(m_2 - m_1)}{SSE_2/(n - m_2)} \tag{2.9}$$

F_{12} (2.9) also indicates a relative goodness of fit of a one-step versus a two-step pattern. This is an F-distribution whose p-value represents the probability of the same result on random data. F_1, F_2, F_{12} are then used to make decision about transition models of gene profiles such as observed data belong to one-step model if its P-value for F_1 is significant, but F_{12} does not have a significant P-value. Sometimes the data does not match with the one-step model, though it has significant P-value for F_2, in this case its model is represented as two-step model. Otherwise, the model belongs to ‘no change’ transition model.

2.2.1.3 Time warping method

Dynamic time warping is a distance measure originally developed for speech recognition in the 1970s [36, 37]. It has been used in many areas such as handwriting, online signature matching [38, 39], data mining and time-series clustering [40], computer vision and animation [41]. Time warping algorithm, similar to sequence alignment algorithms, is used to align two time-series. Sequence alignment and time warping are different from each other at a point such that the former considers base or residue similarity individually, while the latter considers the similarity of pairs of vector taken from a common k-dimensional feature space taken one from each time-series. In this study, feature space represents vectors of common set of k genes’ expression levels, since alignment of the gene profiles is the main purpose of the study.

An algorithm proposed by Aach and Church [42] was implemented to align time-series experiments. The algorithm, which is developed from the principle in Kruskal and Liberman [43], is an implementation of simple and interpolative time warping algorithms for expression data. The formulization of the approach is given as follows; there are two time-series a and b, a has n time points; $0,1,\ldots,n$ at times

t_0< t_1 <….< t_n, b has m time points; $0,1,\ldots,m$ at times u_0< u_1 <….< u_m. Each series is associated with a set of k genes then they are referred as being associated with a trajectory of feature vectors in k-dimensional feature space. While feature vectors of time-series a at time point $\mathrm{t_i}$ is shown by a_i, for series b at time point $\mathrm{u_j}$ is shown by $\mathrm{b_j}$. The algorithm aims to find the correspondence between the time points of each series that minimizes the overall distance $D(a,b)$ between trajectories. A representation of two aligned series in feature space is given in Figure 2.3.a.

$$i(0) = 0$$

$$i(h+1) = either\ i(h) + 1\ or\ i(h) \tag{2.10}$$

$$i(p) = n$$

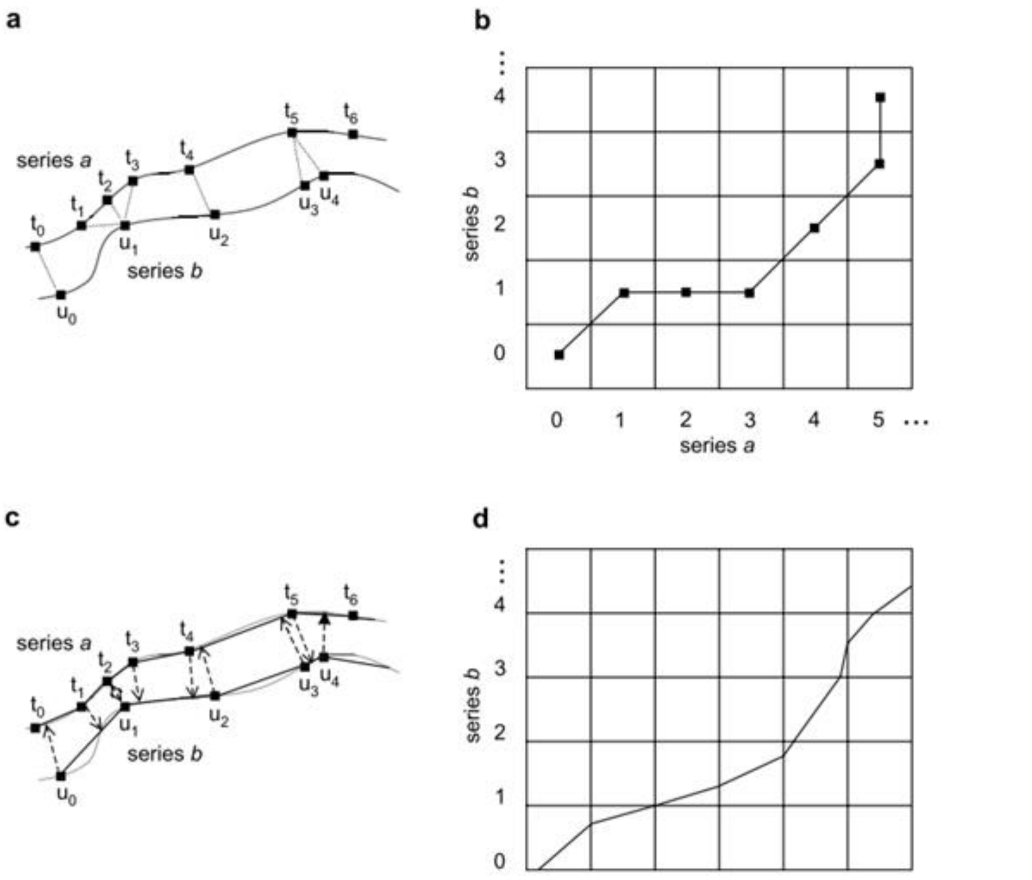
Order and continuity constraints are defined by warping paths through a table. As given in (2.10), $i(h)$ and $j(h)$ represents paths in simple warping algorithm, time points in given series are shown as $h = 0,1,\ldots,p$.

$$D_q(a,b) = \sum_{h=1}^{q} w(h) d(a_{i(h)} b_{j(h)}) \tag{2.11}$$

$$w(h) = \frac{1}{2}(t_{i(h)} - t_{i(h+1)} + u_{j(h)} - u_{j(h-1)}) \tag{2.12}$$

In Figure 2.3.b the warping path corresponding to Figure 2.3.a is depicted. As shown from the formula (2.11), $D_q(a,b)$ refers to the overall distance of the warping path. In addition, $W(h)$ (2.12) represents the average time spent between two trajectories.

$$e(x,y) = \sqrt{\sum_{i=1}^{k} f_i (x_i - y_i)^2} \tag{2.13}$$



(a) Two time series a and b in a two-dimensional feature space containing sample points from a continuous process, with sample points of each series mapped to each other by simple time warping.

(b) Dynamic programming matrix for simple warping and the optimal path corresponding to (a).

(c) Series a and b from (a) with sample points on each series mapped to interpolative points on the other by interpolative time warping algorithm

(d) Dynamic programming grid for interpolative warping and the optimal path corresponding to (c).

Figure 2.3 Representation of the time warping algorithm[2]

Distances for the algorithms on weighted Euclidean distance is defined as given in the formula (2.13) where x and y are k dimensional feature vectors, f_i is the feature weight. These weights can be specified as parameters. In the study $f_i = 1$ is used for all genes.

Optimal alignment score of two time-series is produced by the time warping algorithm. The alignment score is a powerful factor to assess the quality of the obtained alignment. The score is zero when two series are identical, while they are

[2] Quoted from: AACH, John and CHURCH, George M. ,Aligning gene expression time series with time warping algorithms. *Bioinformatics*. ,Vol. 17, no. 6, p. 495–508. ,2001.

different from each other the score diverges from zero. The average of alignment of all common gene pairs is taken as the overall alignment score between the given time-series experiment. Obtained score was used as the similarity measure for finding similarity between the experiment pairs.

2.2.1.4 Lyapunov exponent method

In recent years, separating chaos from noise has become one of the significant research issues. Lyapunov Exponents (LEs) measure the rate of convergence or divergence of nearby trajectories (the path that a moving object follows through space as a function of time) that represent chaos [44] in a system. In other words, they are used to quantify sensitivity to initial conditions in a dynamical system. While negative LEs indicate convergence, positive ones are indication of divergence. Chaotic behavior can be easily estimated on a time scale and the greatness of the LE is a marker of the time scale. There is a variety of methods developed for identifying chaos by using experimental time-series [45–47]. The Grassberger-Procaccia algorithm (GPA) [47] is one of the widely used methods to identify chaos in dynamic system. GPA is easy to implement, however it is sensitive to variations in its parameters such as embedding dimension, reconstruction delay. In many implementation of LE on time series a positive characteristic exponent shows chaos, therefore calculating only the largest LE of the given series is enough to identify chaotic system. On the other hand, existing methods for calculating LEs have some disadvantages such as being unreliable for small datasets, being difficult to implement or having high computational cost. For these reasons, to calculate largest LE a method [48] which is faster and easier to implement than other methods was used to get LE of the time series experiments in this study.

$$|| \delta_{x(t)} || = || \delta_{x(0)} || \, e^{\lambda t} \tag{2.14}$$

$$\lambda(i) = lim_{t \to \infty} \frac{1}{t} log \frac{||\delta_{x_i}(t)||}{||\delta_{x_i}(0)||} \tag{2.15}$$

Let the Lyapunov Exponent λ is defined as the average of the local separation of the adjacent curve degree in space (2.14). If λ is negative, different initial conditions tend to give the same output, so it is said that development is not chaotic. Otherwise, different initial conditions give separate outputs then movement is chaotic. Initially, there is a small difference $\delta_{x(0)}$ between two close points (x_1, x_2), one of them is set as reference point, located on two close curves. At the end of time t, these points diverge from each other and the difference between them becomes $\delta_{x(t)}$. Lyapunov Exponent can be calculated as given in the formula (2.15) (||…|| indicates Euclidean distance). In phase space, due to a λ represents convergence and divergence at each dimension, LE spectrum λ_1 of d-dimensional dynamic system (R^d) is calculated as follows; $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. In chaotic system, there is at least one largest LE and if the exponent is greater than 0, behavior of the system is chaotic, otherwise it is a deterministic system.

2.2.2 Fingerprint comparison methods

After having obtained fingerprints for each experiment, the next process is comparing these fingerprints with an appropriate similarity metric. A convenient similarity metric was used based on the fingerprint extraction method used. Detailed description of each comparison metric was given in the next sections.

2.2.2.1 Overlap similarity metric

Overlap similarity metric is adapted for comparison of fingerprints generated by Transition model-based fingerprint extraction method. In spite of its simplicity, it is a widely used metric for categorical data [49]. The mentioned extraction method defines gene expression profiles over categorical values, so the overlap metric was selected as an appropriate comparison metric for these values. The overlap score ranges between 0 and 1; if there are no similarity between compared objects the score is 0; while perfect match between them is represented by the value of 1.

$$S(X,Y) = \sum_{k=1}^{d} S_k(X_k, Y_k)/d \qquad (2.16)$$

$$S_k(X_k, Y_k) = \begin{cases} 1, & if\ X_k = Y_k\ and\ X_k, Y_k \geq 1 \\ 0, & otherwise \end{cases} \qquad (2.17)$$

Let X and Y be fingerprint vectors to be compared, the overlap score between these vectors is given in (2.16) and (2.17). In this metric, the similarity measurement is calculated considering only common genes, called as d, in compared experiments. As given in (2.17), individual gene behaviors are considered similar when the labels differ from 0. The value of 0 is not regarded as a similarity because it represents no change in time expression value over a time period. This choice was made since the most of the genes in an experiment do not have differentially expression profiles in terms of any specific environmental condition. Considering these genes as similarity between experiments may cause a dominating factor among other categorical labels. In addition to this, the similarity between experiments that have differentially expressed genes point more valuable relevance of the compared experiments. Due to these reasons, the original overlap metric was adapted to be applied for the studied case.

2.2.2.2 Tanimoto similarity metric

Tanimoto distance, so-called Jaccard, is used for comparison of fingerprints obtained with Differential Expression Profile-based fingerprinting method. It is originally used for comparison of unordered sets. Similarity between two unordered sets is calculated as the ratio of their common elements to the number of all different elements. Usually, similarity metrics are defined over binary valued vectors, so vectors that have categorical features should be converted into binary features to implement Tanimoto coefficient.

$$\textit{Tanimoto Coefficient} = \frac{a+d}{a+d+2(b+c)} \qquad (2.18)$$

Rogers and Tanimoto [50] defined Tanimoto similarity measurement, as given in the formula (2.18), for binary valued vectors. Tanimoto coefficient can be described over fingerprint vectors such as; X and Y are fingerprint vectors of two different experiments. Contingency table [51] for those vectors is given in Table 2.1. The table consists of comparing results of the values for X and Y:

a= number of times X_i=1 and Y_i=1
b= number of times X_i=0 and Y_i=1

c= number of times X_i=1 and Y_i=1
d= number of times X_i=0 and Y_i=0

In order to use this similarity metric; fingerprint vectors obtained with Differentially Expression Profile-based Method are converted into binary vectors as described in the previous section. The Tanimoto scores range between 0 and 1; 0 means no similarity, 1 shows a perfect match between compared experiments.

Table 2.1 Contingency table values for two fingerprint vectors

		Fingerprint vector of Y		
		1	0	sum
Fingerprint vector of X	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	$a+b+c+d$

2.2.2.3 Pearson correlation coefficient

Pearson coefficient was used for determining whether there is a correlation between Lyapunov Exponents of two compared experiments. It is a widely used measure of the linear dependence between two variables.

$$s(X,Y) = \frac{n\sum x_i y_i - \sum x_i - \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{2.19}$$

If there is a correlation between experiments, it can be stated that those experiments are similar to each other. Let X and Y be compared fingerprint vectors obtained with the LE fingerprinting method and x_i refers to Lyapunov score of gene i of vector X, while y_i is the Lyapunov score of gene i of vector Y. In addition to this, n is the number of genes in each fingerprint vectors. The measure (2.19) gives a value between +1 and -1, where value of 1 points a positive correlation, 0 refer no correlation and -1 is represents a negative correlation. A positive correlation between similar experiments is expected.

2.3 Results

This section consists of three sub-sections: Data, Evaluation Criteria and Empirical Results. Used time-series experiments are given in the first sub-section, second sub-section describes evaluation criteria of the proposed system and the empirical results are given in the final sub-section.

2.3.1 Data

In order to establish a data repository 120 Arabidopsis time-series experiments from GEO were collected. The datasets were obtained using different platforms and time points range between 3 and 24. In order to minimize cross-platform effects, scaling process for each time-series experiment was performed such that mean is 0 and standard deviation is 1.

2.3.2 Evaluation criteria

An obvious definition, so-called ground truth, is a basic need to evaluate performance of a system. Actual relevance of compared objects is defined according to defined ground-truth. Determination of relevance of retrieved entities is performed using ground-truth information. The most important task in evaluation process is describing the relevance information between compared experiments. This task is usually performed by labelling the experiments based on some environmental factors, such as disease or healthy classes, response to a stimulus; however, it is not an efficient way for time-series experiments. For instance, treatments of patients with the same disease may be different and they may not be related directly to the label of the experiment. Moreover, each treatment affects distinct gene regulation, while some gene-sets may be co-regulated by same treatment in patients with different disease. That is to say defining relevance between time-series experiments should be based on gene-sets rather than static labelling. Therefore, two time-series experiments are considered as biologically relevant when they share common enriched gene-sets. To adapt this consideration into this study, a well-known method named Gene Set Enrichment Analysis (GSEA) [52] was used to get enriched gene-sets between compared experiments. GSEA is a knowledge-driven and analytical method to analyze genome-wide expression profiles at the level of gene sets. It generates set of genes that share common biological functions or regulation. The main goal of the method is determining how members of a gene set are distributed among a given gene list e.g they are located at the top or bottom of the list. GSEA method has some basic steps: it considers that there are expression datasets of experiments, given in a heat map (Figure 2.4), from two different classes. Genes are sorted according to their correlation between their expression level and the class they belong (Figure 2.4.A). Enrichment score (ES) that refers to the degree of overrepresentation at top or bottom of the gene list is calculated (Figure 2.4.B). Then, significance level of ES is estimated by obtaining a nominal p-value. Finally, adjustment of multiple hypothesis testing is made through with getting a normalized enrichment score (NES) and calculating false discovery rate (FDR) for each NES [52]. NES is a main statistic to assess gene set enrichment results and it provides the comparison of the results over the obtained gene sets.

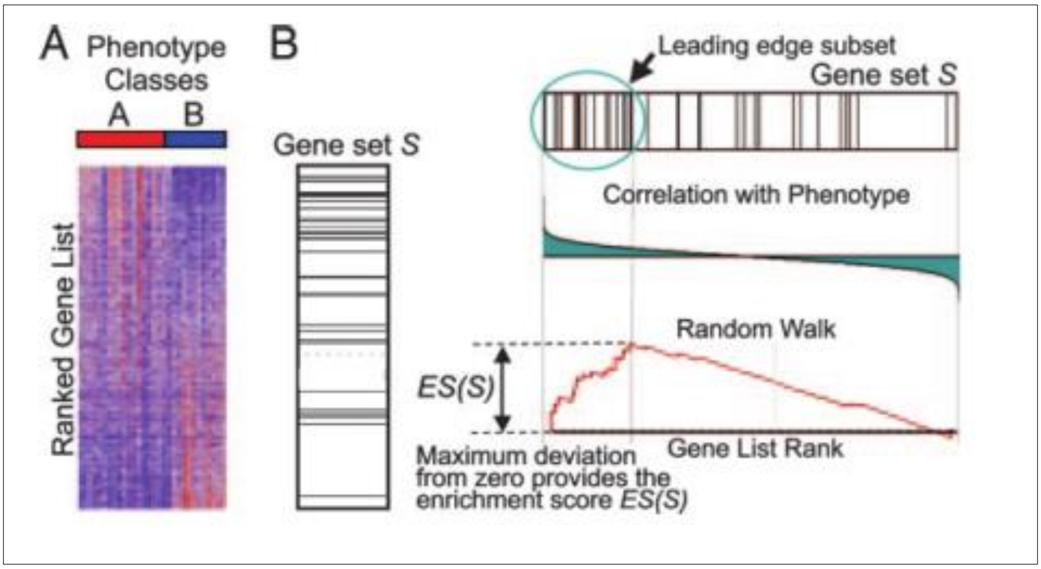


Figure 2.4 GSEA method overview[3]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.20}$$

[3] Quoted from: SUBRAMANIAN, Aravind, et al., 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, p.15545–15550, 2005.

After having obtained gene-sets, the next process is finding similarity between these gene sets. As given in formula (2.20), Jaccard coefficient is calculated between enriched gene sets, named as A and B, of two different compared experiments. The coefficient is calculated by dividing number of common gene sets of compared experiments by the number of all gene sets. A threshold of 0.3 was selected regarding a Gaussian distribution of the Jaccard index values of all experiment pairs. The threshold was obtained by summing the mean of all values and the standard deviation of the data. The true relevance between experiments was depicted by the obtained threshold.

To evaluate the system retrieval performance, Receiver Operating Characteristic (ROC) curves was also used in this study. In recent years, ROC curves are commonly used in biomedical, machine learning and data mining fields. Although, it is used to visualize and organize classifiers based on their performance, it can be used to evaluate and compare algorithms [53]. ROC graphs shows relation between true positives rates (TPR) plotted on X axis and false positive rates (FPR) plotted on Y axis. TPR represents ratio of positives correctly classified to total positives, FPR is the ratio of negatives incorrectly classified to total negatives. In Figure 2.5, a simple ROC graph is given to show performance of five distinct classifiers. In the graph, upper left corner (0, 1) denotes perfect classification and diagonal line represents random guess. Upper side of the diagonal line shows better classification, while lower side shows worse classification. So, it can ben stated that A, B, D classifier have better performance than E and C classifiers according to the graph. Also, C’s performance is random. Area under ROC curve (AUC) is calculated to compare ROC performance of classifiers. Its value ranges between 0 and 1. The higher AUC score shows better retrieval performance, value of 1 refers to perfect case.

In addition to these evaluation processes, statistically significance tests such as Paired t-test and Wilcoxon signed-rank test were performed. It was aimed to observe that whether the differences between performances of used fingerprint extraction methods was statistically significant. It is expected that obtained p-values should be below the value of 0.05. This value demonstrates that the difference between performances of fingerprint extraction methods with regard to AUC score is statistically significant.

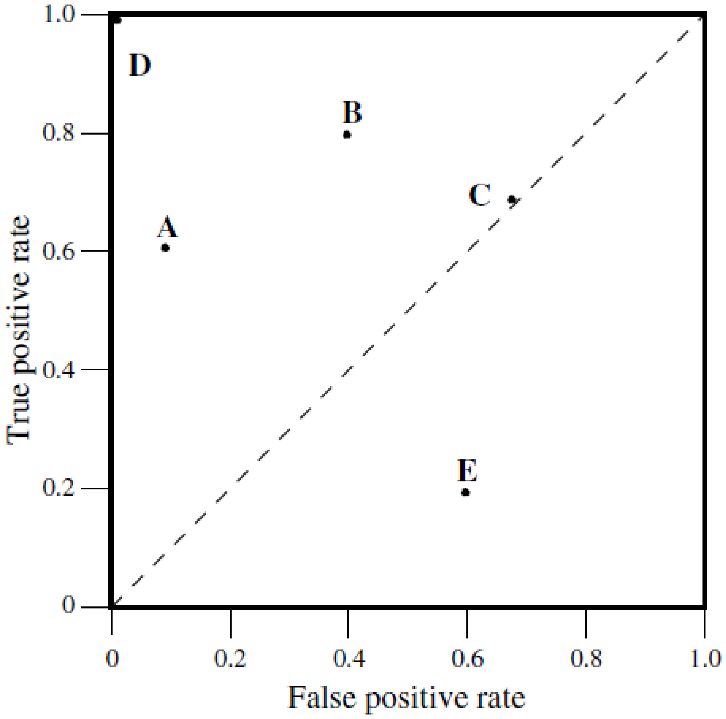


Figure 2.5 A basic ROC graph showing five different classifiers[4]

2.3.3 Empirical results

In this study, four different fingerprint extraction methods and three fingerprint comparison methods, given in Table 2.2, were used. A similarity metric is needed for each fingerprint extraction methods except Time Warping method, since it generates an alignment score that can be used as similarity score for experiment pairs to be compared. In addition to this, when performing the fingerprinting method Transition Model-based method, the parameter k, which is the rank-based binarization parameter to select top %*k* of genes, was selected as 1, since the best retrieval performance was achieved with this value.

In order to perform retrieval task, all experiments in the data collection are taken as a query respectively and a ranked list of retrieved experiments, based on a similarity score calculated by an associated similarity metric, is obtained. It is

[4] Quoted from: FAWCETT, Tom, An introduction to ROC analysis, Pattern Recognition Letters, ,Vol.27, no.8, p.861–874, 2006.

expected that retrieved experiments that have higher similarity with the query experiment be at the top of the list. As stated previously, the system performance is evaluated by ROC curves. For each relevant experiment an AUC score was calculated. Higher AUC score indicates the better system performance.

Retrieval performances of all fingerprint extraction methods were given in Figure 2.6. The performances are given as ROC plots in which horizontal axis represents AUC scores; the vertical axis depicts number of experiments with a corresponding AUC score. According to the plots, Time Warping method has become more successful in retrieving relevant experiments. Moreover, it can be seen that an AUC score greater than 0.6 was obtained for the majority of the experiments for each fingerprinting method. Average AUC scores for Time Warping method is 0.77, while it is 0.73 for Transition Model-based method. Besides this, AUC of 0.70 and 0.68 for Differentially Expression Profile-based method and Lyapunov Exponent method are obtained respectively.

Table 2.2 Fingerprint Extraction and Comparison Methods

Fingerprint Extraction Method	Fingerprint Comparison Method
Differentially Expression Profile-based Method	Tanimoto
Transition Model-based Method	Overlap
Time Warping Method	---
Lyapunov Exponent Method	Pearson

Statistical significance tests, a Paired t-test and a non-parametric Wilcoxon signed- rank test, were performed to detect whether the difference between performances of used methods were statistically significant. AUC score difference between fingerprint extraction methods pairs were used to perform p-value calculation. Compared method pairs and related p-values using two different tests were given in Table 2.3. As can be seen from the table, in most of the results, it was observed that p value was below 0.05 which is the threshold of statistical significance. In addition to this, this result is strong evidence that difference between AUC scores of Time Warping method, that has the best retrieval performance, and other method's AUC score is statistically significant.

Some examinations also were performed to evaluate the system performance based on manual annotations. Before indirect evaluation based on gene-sets, a direct evaluation based on textual relevance was performed to discover biological sense of the fingerprinting approaches. Transition-model based fingerprinting approach and related similarity metric, named Overlap metric, were selected to evaluate the system performance based on textual relevance between retrieved experiments. To this end, three specific experiments from the collected dataset were taken as query experiments. For each query experiment, a ranked list was obtained from the experiment collection. When selecting query experiment it is expected that at least two relevant experiments which have higher overlap scores than other experiments should be retrieved.

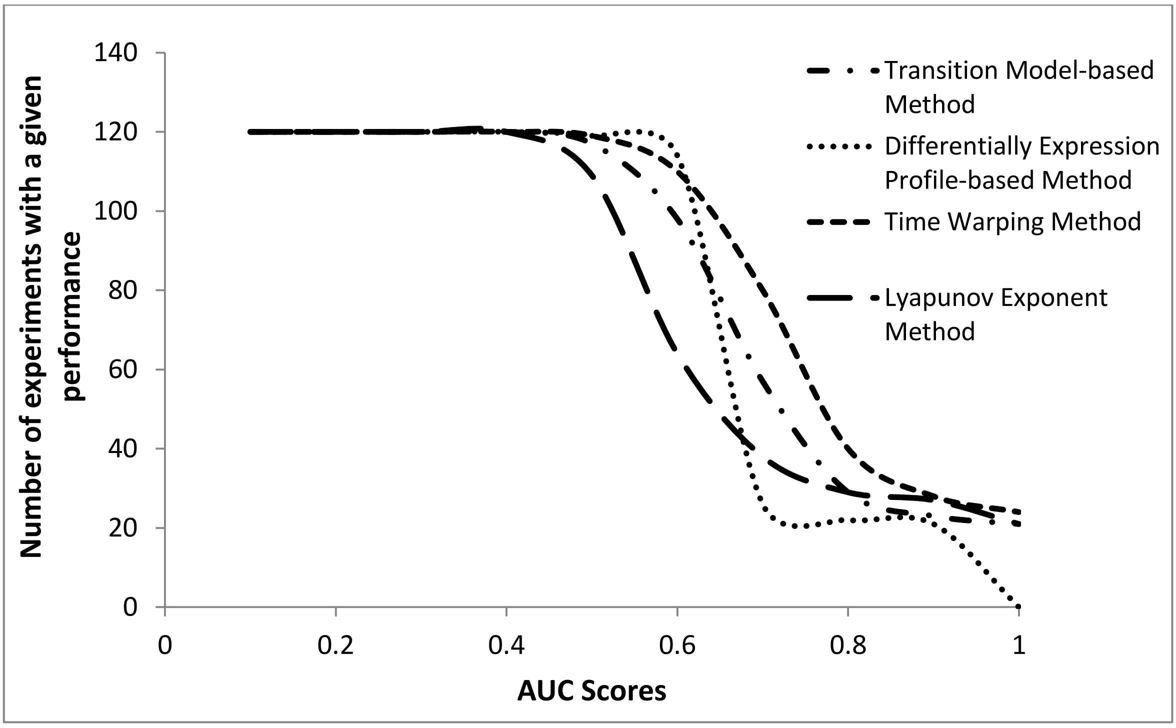


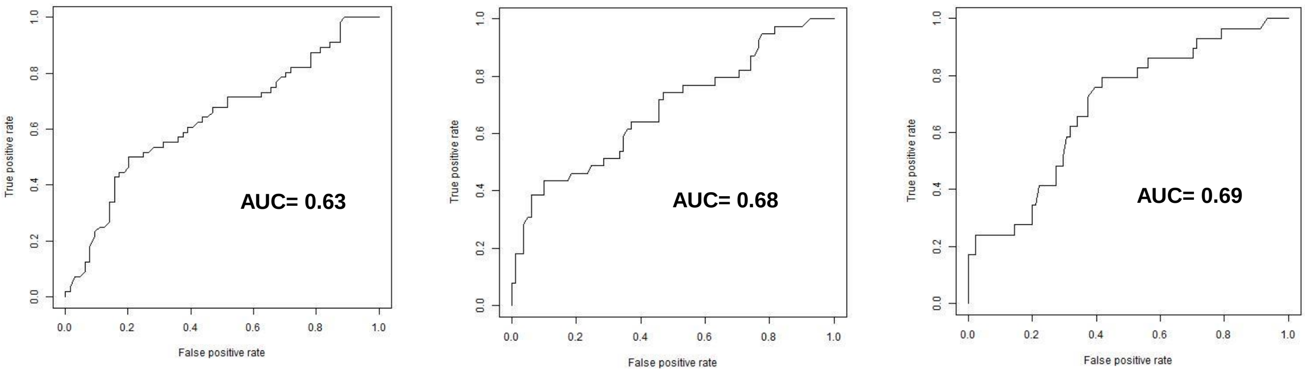
Figure 2.6 Retrieval performances of all fingerprinting methods

Table 2.3 Statistically Significance Tests of Fingerprint Extraction Methods

Method Pair		p-value	
		Paired t-test	**Wilcoxon signed-rank test**
Transition Model-based Method	Differentially Expression Profile-based Method	0.04861	0.0004649
Transition Model-based Method	Time Warping Method	**1.49E-06**	**3.669E-09**
Transition Model-based Method	Lyapunov Method	0.0002567	9.241E-07
Differentially Expression Profile-based Method	Time Warping Method	**2.064E-11**	**3.626E-13**
Differentially Expression Profile-based Method	Lyapunov Method	0.1509	0.004291
Time Warping Method	Lyapunov Method	**5.54E-15**	**1.008E-12**

The first query experiment is about transcriptional regulation based on the MYB46-mediated. In this experiment, transcriptome profiles were generated in terms of secondary wall development at different time periods such as 1h, 3h and 6h [54]. It has the accession number of GSE16143-2 in which "-" points to experiment number in same GEO entry. Moreover, the second experiment, GSE3350-1, is about analyzing of structures of auxin-induced cell division. Lateral root initiation was used to measure expression levels at three different time points 2h and 6h [55]. Finally the third experiment, GSE18975-7, was studied for observing natural variations of downstream auxin responses. Gene expression measurement of Arabidopsis Seedlings grown in liquid culture was performed at time points of 0, 30 min, 1h and 3h [56].

For each selected query experiments, ROC performances are depicted in Figure 2.7. The obtained AUC scores were 0.63, 0.68, and 0.69 respectively. In addition to this, most relevant and least relevant retrieved experiments with each query are given in Table 2.4. The first two rows represent most relevant experiments, while the others represent the least relevant experiments with the query. As stated before, ranked list was generated based on the overlap score. Overlap score of experiment pairs represents the system prediction, while Jaccard score refers to true relevance obtained with gene set-based comparison. According to these scores, it can be observed that true relevance and predicted relevance of retrieved experiments have a powerful correlation.

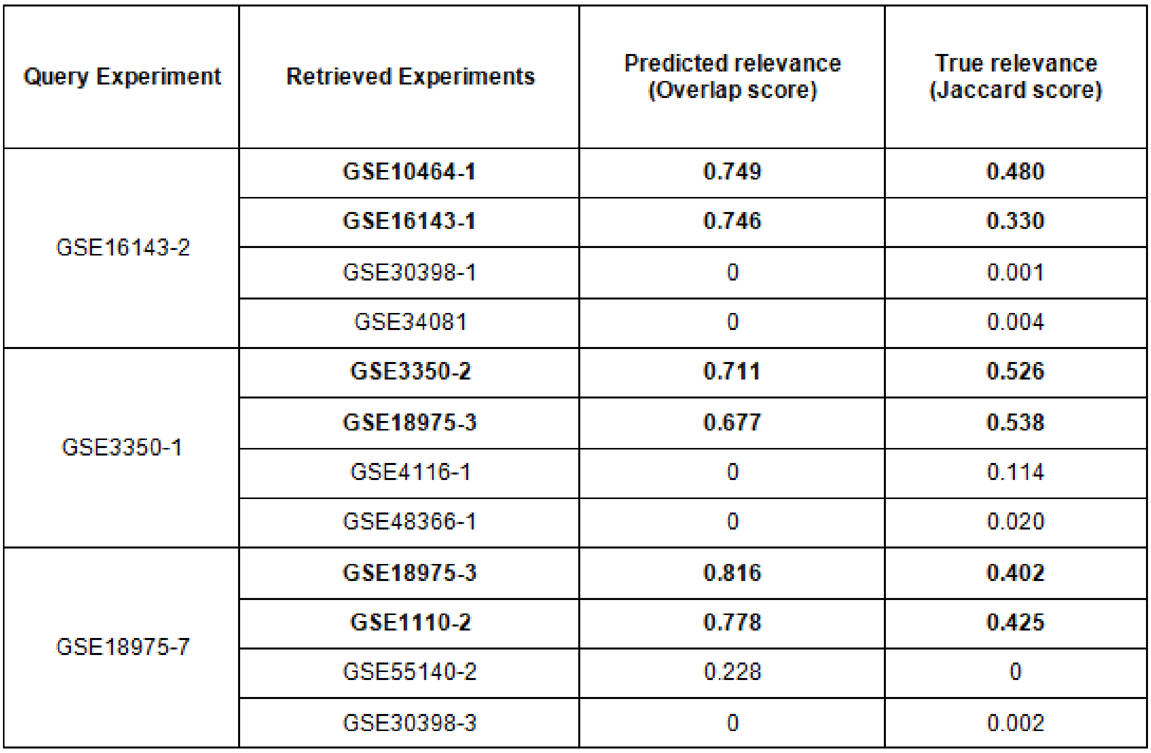


(a) GSE16143-2
(b) GSE3350-1
(c) GSE18975-7

Figure 2.7 ROC curves of sample query experiments

After having obtained retrieved experiments, evaluation of the retrieval was performed by biological relevance. Manual annotations of the experiments were compared to discover relevance between them. The first query sample, GSE10464-1 was an experiment which was done for discovering the gene expression changes in response to paraquat [57]. In both experiments, after applying different treatments, Arabidopsis seedlings were harvested at nearly same time periods. It can be observed that the system retrieves relevant experiment with the query using same stress response. Moreover, GSE16143-1 was reported as second most relevant experiment. It is the part of same GEO entry with the query experiment. The research on this experiment was conducted using two different conditions such as with and without dexamethasone treatment. The query experiment uses the treatment; the retrieved one is an experiment without the treatment. Similarity between these experiments comes from the conditions used in conducting the research rather than treatments for the experiments.

Table 2.4 Most relevant and least relevant experiments for sample queries



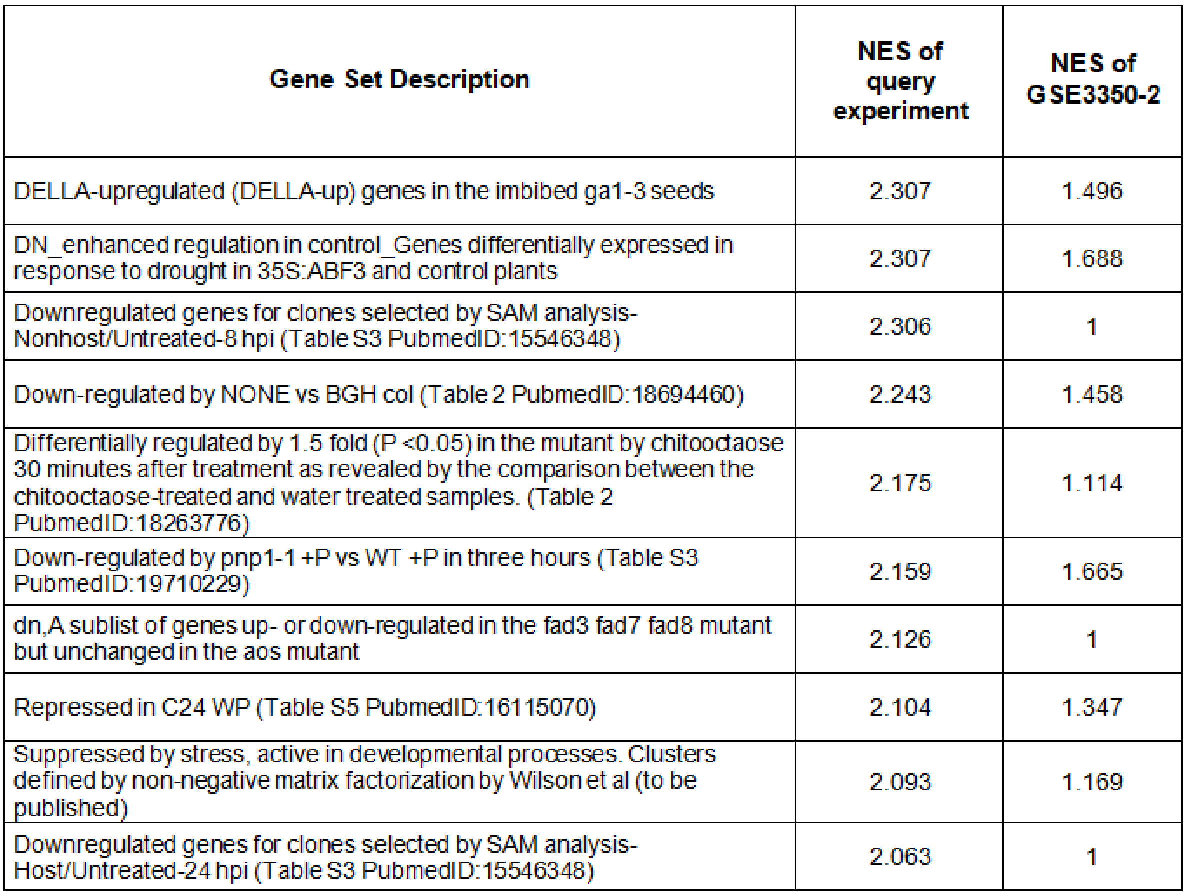
Query Experiment	Retrieved Experiments	Predicted relevance (Overlap score)	True relevance (Jaccard score)
GSE16143-2	**GSE10464-1**	**0.749**	**0.480**
	GSE16143-1	**0.746**	**0.330**
	GSE30398-1	0	0.001
	GSE34081	0	0.004
GSE3350-1	**GSE3350-2**	**0.711**	**0.526**
	GSE18975-3	**0.677**	**0.538**
	GSE4116-1	0	0.114
	GSE48366-1	0	0.020
GSE18975-7	**GSE18975-3**	**0.816**	**0.402**
	GSE1110-2	**0.778**	**0.425**
	GSE55140-2	0.228	0
	GSE30398-3	0	0.002

GSE3350-2, which is the most relevant experiment with the second query, is the experiment in same GEO entry as in the previous query experiment. The system defined them as relevant experiments, because they are obtained almost in same environmental conditions and setup. Furthermore, GSE18975-3 was given as second most similar experiment with the query. It is the study of natural variation of auxin response in different time points such as 30 min, 1h and 3h [56]. There is an interesting point that auxin response observation was also performed for the query experiment. Their similarity is based on having same treatment with different purposes.

Moreover, GSE18975-3 and GSE1110-2 were retrieved as the most relevant experiments with the third query experiment, named GSE18975-7. The former is obtained from the same GEO entry, they have same environmental conditions and it is expected that they are more relevant. The latter has important relevance with the query experiment, because it was studied in same environmental conditions with the query experiment to observe auxin treatment in Arabidopsis. That is to say, the system has succeeding in finding relevant experiments with the same treatment or environmental conditions.

In addition to manual annotation based evaluation, GSEA-based evaluation was also performed. The query experiment GSE3350-1 and GSE3350-2 the most relevant experiment were taken as sample experiments for this purpose. Table 2.5 shows first 10 common gene sets and related NES between the query and relevant experiment. In this study, a gene list from the study of Yi et al. [58] was used in performing GSEA, because current GSEA implementation does not support gene sets of Arabidopsis organism. According to the results, there are common gene sets with high NES between the query and the relevant experiment.

Table 2.5 Gene sets enriched for both query (GSE3350-1) and first relevant experiment (GSE3350-2)



Gene Set Description	NES of query experiment	NES of GSE3350-2
DELLA-upregulated (DELLA-up) genes in the imbibed ga1-3 seeds	2.307	1.496
DN_enhanced regulation in control_Genes differentially expressed in response to drought in 35S:ABF3 and control plants	2.307	1.688
Downregulated genes for clones selected by SAM analysis-Nonhost/Untreated-8 hpi (Table S3 PubmedID:15546348)	2.306	1
Down-regulated by NONE vs BGH col (Table 2 PubmedID:18694460)	2.243	1.458
Differentially regulated by 1.5 fold (P <0.05) in the mutant by chitooctaose 30 minutes after treatment as revealed by the comparison between the chitooctaose-treated and water treated samples. (Table 2 PubmedID:18263776)	2.175	1.114
Down-regulated by pnp1-1 +P vs WT +P in three hours (Table S3 PubmedID:19710229)	2.159	1.665
dn,A sublist of genes up- or down-regulated in the fad3 fad7 fad8 mutant but unchanged in the aos mutant	2.126	1
Repressed in C24 WP (Table S5 PubmedID:16115070)	2.104	1.347
Suppressed by stress, active in developmental processes. Clusters defined by non-negative matrix factorization by Wilson et al (to be published)	2.093	1.169
Downregulated genes for clones selected by SAM analysis-Host/Untreated-24 hpi (Table S3 PubmedID:15546348)	2.063	1

3. WHOLE-METAGENOME SEQUENCING SAMPLE RETRIEVAL

This chapter consists of a content-based retrieval framework developed for whole-metagenome sequencing samples. The chapter is organized as follows; motivation of the study and related work are defined in the first part, while k-mer extraction, k-mer selection methods, fingerprint extraction and comparison methods are described in the second part and the experimental results are given in the final part.

3.1 Introduction

Analyzing metagenomic data has become a significant research interest with the rapid development in sequencing technologies. There are two main approaches in studying metagenomic samples; some studies concentrate on targeted sequencing of particular genes like 16S rRNA, others focus on whole-metagenomes [59, 60]. Phylogenetic profiling information can be obtained easily by targeted studies at a lower cost through the former approach, while much more information such as inhabitant genetics of the community can be gained by the latter approach. Targeted sequencing has some disadvantages such that it does not provide any information about other genes except 16S rRNA gene and there may be conflicts between generated phylogenetic trees. Lately, an alternative and more informative approach, called whole-metagenome shotgun (WGS) sequencing, was proposed to obtain vast number of DNA reads of all organisms. There have been a great number of studies about WGS sequencing by which DNA reads of all organisms can be produced. Qin et al. [61] stated that there is a relationship between type II diabetes disease and gut metagenome samples. An automated analysis platform, called MG-RAST, was developed by Meyer et al. [5] to accumulate and access data, make quality control and analysis of almost 3000 metagenomic sequence samples. In addition to this, iMicrobe project [62] provides microbial datasets and computational frameworks for researches. Although, these repositories have some analysis modules, they include neither any search function nor comparison tool for sequencing samples.

Detecting similarities between metagenomic samples through huge data collections is a remarkable research area in bioinformatics. Recently, a variety of studies has proposed the content-based approach using distinct perspectives. Huson et al. [63] developed a software tool, named MEGAN, for analysis of metagenomic datasets. The main objective in this study is discovering taxonomic and functional content of the sequences. Firstly, a sequence comparison tool, such as BLAST, is used to align set of DNA sequences and known sequences. MEGAN uses NCBI taxonomy to process comparison results. A MEGAN file that consists of information for analyzing and obtaining graphical and statistical output is generated. In order to evaluate the assignment of the reads and generate the results at varied stages of NCBI taxonomy, LCA (Lowest Common Ancestor) algorithm is performed. Finally, matching sequence of species and taxa are done in the NCBI tree in which species-specific sequences are closer to the leaves of the tree while widely conserved sequences are closer to the root. Wang et al. [64] proposed a naïve Bayesian classifier to classify sequences without aligning them. They tested the system performance using large volume datasets in terms of sequence length. Liu et al. [17] proposed an approach, called MetaDistance, which classifies sequences and selects features of these sequences. They also describe the data normalization method to be applied before the proposed method. It is stated that the method is appropriate for small size datasets and unbalanced classes. In addition to this, Su et al. [65] developed a tool named Meta-Storms to build a database of metagenomic samples and search samples through the database. The proposed system was evaluated using a large number of samples. It succeeded in organizing a database and developing a search system.

A common point of the methods mentioned above is that they use some annotations, taxas or a priori knowledge to analyze metagenomic samples. Although, there are some unknown or unculturable organisms, for example 99% of bacteria, referred methods could not be applied on these organisms. Therefore, new approaches that do not rely on any information or annotations have been suggested on this deficiency. These approaches are called reference-free, unlike alignment-based approaches they use raw read content of the samples to represent them in a feature space. Recently, k-mer (substring of length k) representations are the widely used technique for sample representation among reference-free studies. In this context, Maillet et al. [66] first proposed a method, named Compareads, for finding similar metagenomic samples in a data collection using k-mer approach. The proposed method succeeded in finding similarities between samples. Although this approach is faster than traditional methods, its computational cost is quite high because of storing all k-mer information. Selecting informative features of metagenomic data is quite an important step in data analysis. Some studies use two classes of samples, while others work on all large number of features instead of using any feature selection technique [67–69]. For example, Qin, J. et al [61] make analysis on human gut samples using almost 5 million genes. Moreover, Seth et al. [70] proposed a retrieval system for extracting informative k-mers instead using all k-mers. They applied feature extraction and selection method to find similar experiments from data collection. In addition, Weitschek et al. [71] developed an alignment-free distance for finding similarity between reads. It is clearly seen that alignment-free distance is an efficient way for sequence read comparison. Besides this, Polychronopoulos et al. [72] presented a method based on k-mer analysis combined with rule-based classification approaches to classify bacterial genomes. Dubinkina et al. [73] proposed a dissimilarity approach for detecting similarities among metagenomic samples using short k-mer spectra. It was stated that the proposed approach achieved in detecting similarities between samples and it can be easily adapted into sample analysis pipelines.

As mentioned above reference-free approaches for retrieval of metagenomic samples have promising results. To this end, in this chapter developing an efficient retrieval model using raw read content is mainly aimed. The chapter introduces a content-based retrieval framework developed to retrieve metagenomic experiments. There are four main steps; k-mer extraction and selection methods, fingerprinting methods and comparison metrics for obtained fingerprints. A data collection consisting of real metagenomic samples was used to assess the system retrieval performance. Each experiment in the collection was taken as a query experiment; a ranked list was generated based on the similarity between the query and other experiments. The main objective of the system is retrieving relevant samples from the repository. Relevance between samples is defined as if the patients, named positive samples, are retrieved by the system they are called relevant samples; otherwise the retrieved experiments are called irrelevant.

The basic contribution presented in this study is extracting fingerprints based on two text mining methods, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), which have not been used for retrieving metagenomic samples from this data collection. The experimental results show that LSA is an encouraging fingerprinting technique to represent the experiment content in the feature space and to find similarity between experiments from the experiment.

3.2 Methods

The proposed retrieval system first takes an experiment as a query then it retrieves a ranked list based on the similarity between the query and other experiments in the collection. Each experiment from the data collection is taken as a query respectively. Retrieval process consists of progressive steps such as extraction of k-mers, selection of informative k-mers, fingerprint extraction and comparison of those fingerprints. The overall view of the system is given in Figure 3.1. Firstly, k-mer frequency vectors of the query experiment and other experiments in the collection were obtained. Having extracted k-mer frequency vectors, k-mer selection process is performed for values of k which is greater than 6. After that, two different fingerprint extraction methods were applied to obtain the fingerprints of the experiments. Finally, comparison of these fingerprints was performed to detect similarity between experiments. Beside these processes, direct comparison of frequency vectors was also performed to compare obtained results with fingerprint extraction results. Each process is described in the next section.

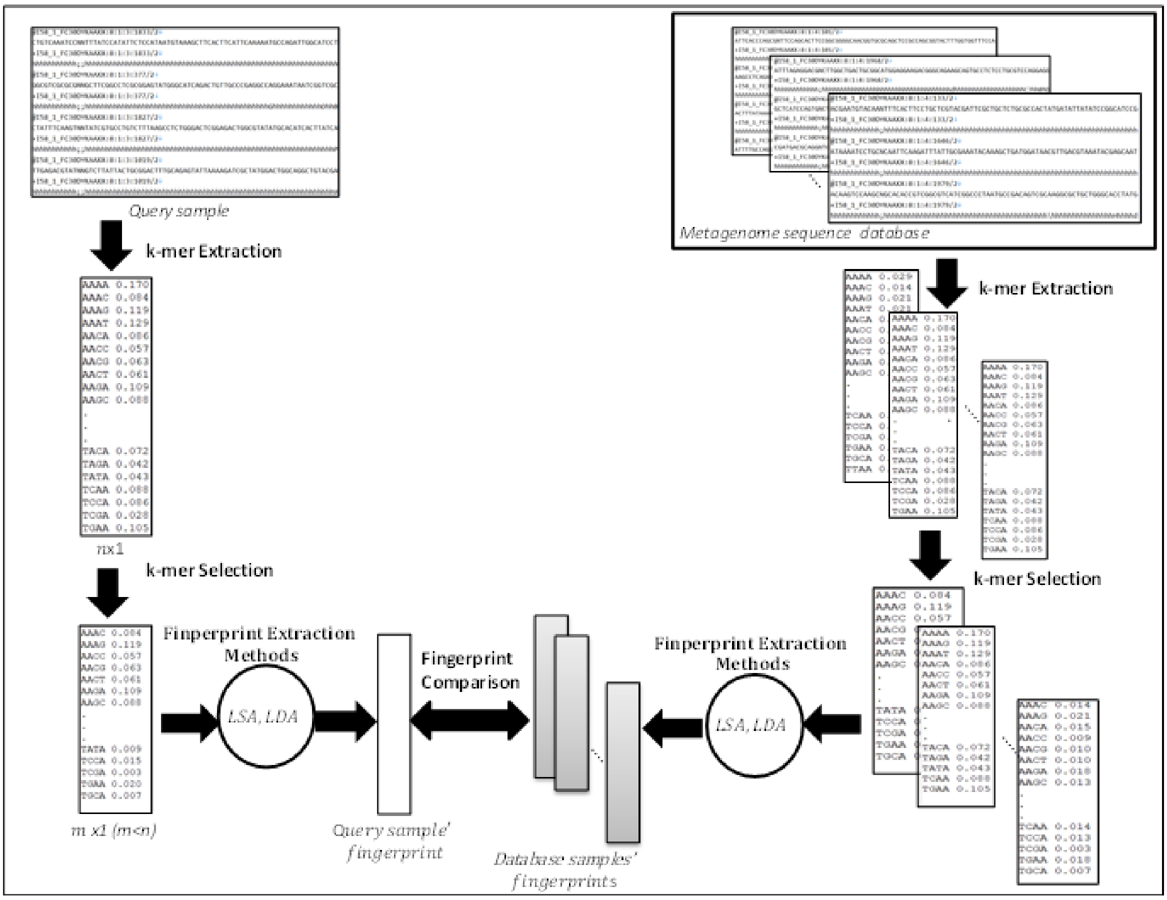

Query sample
k-mer Extraction
nx1
k-mer Selection
m x1 (m<n)
Finperprint Extraction Methods
LSA, LDA
Fingerprint Comparison
Query sample' fingerprint
Database samples' fingerprints
Finperprint Extraction Methods
LSA, LDA
k-mer Selection
k-mer Extraction
Metagenome sequence database

Figure 3.1 General view of the proposed framework

3.2.1 K-mer extraction

The term k-mer, so-called n-gram, refers to all ordered substrings length of k in a string. Extracting all possible k-mers in a sequence read is the main step of metagenome analysis applications. In DNA sequence, which consists of A, C, G and T nucleotides, there are 4^k possible nucleotide subsequences of length k to be extracted. As stated before, the first process of the proposed system is k-mer extraction, with k range between 2 and 13, for experiment representation in the feature space. K-mer frequency is the ratio of total number of the current k-mer to the total number of all k-mers. In DNA sequencing process, which strand is processed and read direction are not known, reverse complement of k-mers was considered in extracting k-mer occurrences in the current experiment. To this end, for each k-mer, read and its reverse complement were calculated, after that the one which comes first lexically was selected as the corresponding k-mer for the current experiment.

3.2.2 K-mer selection

High dimensional nature of data in bioinformatics makes Feature Selection (FS) process necessary to improve performance of data analysis applications, because it has been agreed that the best system performances cannot be achieved with using all features. FS process is applied to select informative features that are relevant with the specific analysis task to be accomplished. In other words, it is the process of eliminating irrelevant and redundant data from the data collection. The main difficulty in FS is selecting set of features which depend on the whole dataset.

In this study, three different FS techniques, selecting features based on Term Frequency Inverse Document Frequency (tf-idf) scores, Correlation Attribute Evaluation (CAE)-based and combinatorial feature selection approaches were used for selecting feature vectors for k>6. Each method is described in detail in the following sections.

3.2.2.1 Selecting features based on term frequency-inverse document (tf-idf) frequency scores

Term Frequency Inverse Document Frequency, so called tf-idf, is a widely used word weighting approach in text mining and information retrieval applications. Term specificity measure was originally introduced by Jones in 1972 [74] and it has been lately known as inverse document frequency. Let a corpus consists of many documents and each document involves different number of words. Basic idea behind this measure is that if a word occurs in many documents of the corpus, it is less important than other terms that occur rarely in the corpus. Tf-idf is evolved from idf measure to find importance of a word within a document collection.

NOO= number of occurrences of k-mer r in experiment e
TNK= total number of k-mers in experiment e
TNE= total number of experiments
NOE= number of experiments in which r occur

$$tfidf_{r,e} = tf_{r,e} * idf_r \qquad (3.1)$$

$$tf_{r,e} = \frac{NOO}{TNK} \tag{3.2}$$

$$idf_r = log_{10}\frac{TNE}{NOE} \tag{3.3}$$

In order to apply tf-idf approach into the developed system, a term is represented by a k-mer, a document is represented by an experiment. For each experiment in the collection, tf-idf scores were calculated as given in the formula (3.1). The product of two terms $tf_{r,e}$ (3.2) and idf_r (3.3) is given as $tfidf_{r,e}$ (3.1). Frequency of a k-mer is shown by $tf_{r,e}$ which points that how often a k-mer (r) occur in the experiment (e). In addition, the importance of a k-mer in the collection is measured by the term inverse document frequency idf_r. All tf-idf scores were calculated for each k-mer, and then these k-mers were sorted in descending order based on the obtained scores. Finally, the first N k-mer in the ranked list was selected.

3.2.2.2 Correlation attribute evaluation (CAE)-based feature selection

Correlation Attribute Evaluation (CAE) method was used as another feature selection method in this study. In this approach, evaluation of an attribute is performed by calculating Pearson correlation between the attribute and the class. The fundamental principle based on this FS method is selecting a subset of features consists of features which are highly correlated with the class, but uncorrelated with each other.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{n})(\sum Y^2 - \frac{(\sum Y)^2}{n})}} \tag{3.4}$$

Let X and Y be two subsets with the size of n, r (3.4) is Pearson Correlation coefficient between them. Pearson correlation is a correlation coefficient widely used in linear regression. It gives the relationship between two sets of data and its score ranges between -1 and 1, perfect match is shown by the value of 1, the negative relationship is represented by the value of -1, while 0 points no relationship between the subsets.

In the proposed retrieval model, CAE technique was applied to get ranked k-mer list based on correlation of the k-mer with the corresponding class. After that, the first N k-mer were selected with regard to the cut-off value determined experimentally. Several runs were performed to observe how the retrieval performance is influenced by the changing number of selected features for k values greater than 6.

3.2.2.3 Combinatorial feature selection approach

A combinatorial approach paired with a robust metaheuristic solution algorithm, the study of Bertolazzi et al. [75], was adapted into this study to tackle feature selection before retrieving samples. Similar methods have already been used with success in other applications regarding genetic and biological sequences [72, 76]. The adapted method, named IP-GRASP (Greedy Randomized Adaptive Search Procedure with a short memory), is a heuristic algorithm based on GRASP. It has been designed for using data set composed of binary and integer features. It is assumed that there is a real-valued data matrix, called A, consists of m rows and n columns in which samples are represented by rows, features represented by columns. Value of a feature on sample i is denoted by the item a_{ik}. The main goal of the method is to gain maximum information by selecting a small number of features. The idea is based on a measure of information using the Euclidean distance.

$$I(A) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \sum_{k=1}^{n} (a_{ik} - a_{jk})^2 \tag{3.5}$$

Information measurement given by $I(A)$, in the formula (3.5), is related to the variance calculated through each pair of samples of the data.

$$\sum_{k=1}^{n} d_{ij}^{k} x_k - \alpha \geq 0, \;\; \forall\, i, j, c(i) \neq c(j) \tag{3.6}$$

$$d_{ij}^{k} = \begin{cases} 1, & if\ a_{ik} \neq a_{jk} \\ 0, & otherwise \end{cases} \tag{3.7}$$

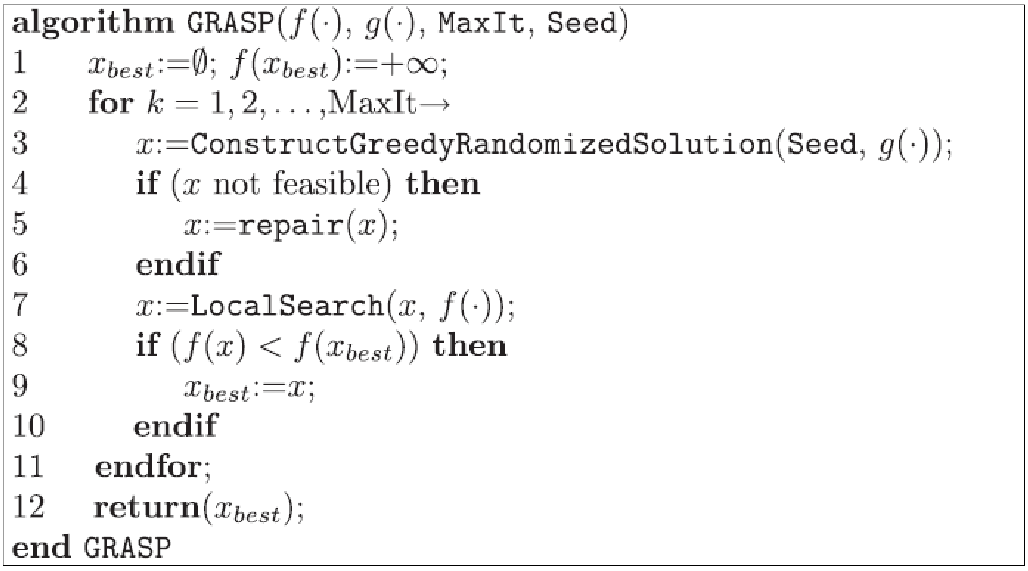
$$\sum_{k=1}^{n} x_k \leq \beta, \tag{3.8}$$

$$x_k = \begin{cases} 1, & if\ k \in N' \\ 0, & otherwise \end{cases}$$

$$x \in \{0,1\}^n, \alpha \epsilon \mathbb{R}^+$$

The main task is defined as reducing dimension of the data matrix, so such problem can be formalized as given in the formula (3.6). When applying the method to the original matrix A, it is considered that reduced number of dimension, so-called target dimension, is β and subset of columns is represented by N' which consists of features selected by the method. The problem can be formulized using a binary variable x_k defined in (3.8) to represent selected features. Minimal threshold quantity α is also depicted to provide separation samples projected on reduced dimension, especially when the reduced dimension β, takes the value of 1. This threshold quantity is selected as large as possible. Another issue that addressed for describing the general formula of the model is that each object may belong to one or more classes in supervised learning problems. Therefore, the proposed system should point the correct class of the object. Thus, samples that belong to different classes are used, the others are eliminated. The class of an object is represented a mapping c which shows the class of an object. As given in the formula (3.6), samples that belong to different classes are selected in the construction of the model. Finally, to represent objects by binary features instead of using a distance function between samples, a generic element of the constraint matrix d_{ij}^{k} (3.7) is defined. In this way, controlling of the value of a feature for two samples is equal or not is performed.

Greedy Randomized Adaptive Search Procedure (GRASP) with some modifications is applied to solve problems stated above. The algorithm is an iterative method and there are two steps in each iteration such as construction of a solution and local search. Basic steps of the algorithm are given in Figure 3.2 as a pseudocode. Firstly, a solution named x is built in the construction phase. A finite solution set named X, an objective function $f: X \rightarrow \mathbb{R}$ to be minimized are given in the model. In order to find a local minimum for the solution, its neighborhood is detected in the local search. Maximum number of iteration is named as *MaxIt*, while initial seed for the pseudo-random number generator is named as *Seed*. When adding a new element to the solution, the algorithm uses a greedy function named $g: C \rightarrow \mathbb{R}$ to select an element from a candidate list C based on its benefit. Benefit of each element is changed at each iteration; therefore the procedure is called as an adaptive procedure. The model select one of the candidates from the list and this list is called the restricted candidate list (RCL). When a row shows a value of 1 in a column, this means that the column covers this row of binary matrix. The selection process for a column consists of number of rows that are not covered by that column. Rows covered by a column are memorized by the model and cover process is performed based on order derived in the previous iteration. Each row is considered not covered by any column and has largest order. A randomized selection is done between those rows and columns that cover rows are stored in the RCL. After that, selection of a column from the RCL list is performed using a weight distribution. Columns have higher weights, if they appear in any of the best solutions.



```
algorithm GRASP($f(\cdot)$, $g(\cdot)$, MaxIt, Seed)
    $x_{best}$:=$\emptyset$; $f(x_{best})$:=$+\infty$;
    for $k = 1, 2, \ldots$,MaxIt→
        $x$:=ConstructGreedyRandomizedSolution(Seed, $g(\cdot)$);
        if ($x$ not feasible) then
            $x$:=repair($x$);
        endif
        $x$:=LocalSearch($x$, $f(\cdot)$);
        if ($f(x) < f(x_{best})$) then
            $x_{best}$:=$x$;
        endif
    endfor;
    return($x_{best}$);
end GRASP
```

Figure 3.2 Basic steps of GRASP algorithm[5]

[5] Quoted from: BERTOLAZZI, Paola et al., Integer programming models for feature selection: New extensions and a randomized solution algorithm, European Journal of Operational Research, Vol.250, no.2,p.389–399, 2016

3.2.3 Fingerprint extraction methods

Fingerprinting is a widely used technique to represent an object by summarizing its content in a feature space. Extracting fingerprints of experiments was performed to get feature vector representation of experiments before detecting similarities among the experiment collection. This process was applied using different text mining techniques such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Applying these techniques, terms used for text mining applications were matched as; a k-mer refers to a word, while a document represented by an experiment. Moreover, the corpus is represented by the data repository or collection.

3.2.3.1 Latent semantic analysis (LSA)

Lexical matching is a popular way that search engines use to retrieve relevant information from data collections. However, this approach is straightforward and fast; it fails to retrieve much relevant information. New approaches are required to overcome problems of existing retrieval techniques. Latent semantic analysis (LSA) is one of the proposed approaches for this purpose by assuming that there is a latent structure in the data that can be discovered with statistical techniques. It is a data mining approach which uses linear algebra techniques to discover relationships between documents and terms. LSA was firstly presented by Dumais et al. [77] and Deerwester, et al. [78] to be applied in IR studies. Common approach in this approach is that documents are represented vectors of terms in the vector space model. All documents are turned into a feature vector in which each term count in the document is represented as a distinct feature and stored in term-document matrix. There are four fundamental steps in LSA;

- Term-document matrix construction; frequency of each term in terms of the document is calculated.
- Term-document matrix transformation; obtained frequency values in the previous step are transformed to represent the importance of each term for a document in the corpus.

- Dimension reduction; to get latent structure of the transformed matrix Singular Value Decomposition (SVD) is applied. In this way, x largest singular values are extracted.
- Retrieval process in reduced space; the retrieval process is performed.

$$A = U * S * V^T \tag{3.9}$$

SVD is used to decompose rectangular term document matrix $\mathrm{A}\,(m\,x\,n)$ into three distinct matrixes (Figure 3.3). Matrix U $(m\,x\,m)$ is a real unitary matrix, S $(m\,x\,n)$ is a rectangular diagonal matrix with entries in descending order and V $(n\,x\,n)$ is a unitary matrix as given in the formula (3.9) [79]. Left and right singular vectors of A are given by U and V which means that A is represented using orthogonal indexing dimensions.

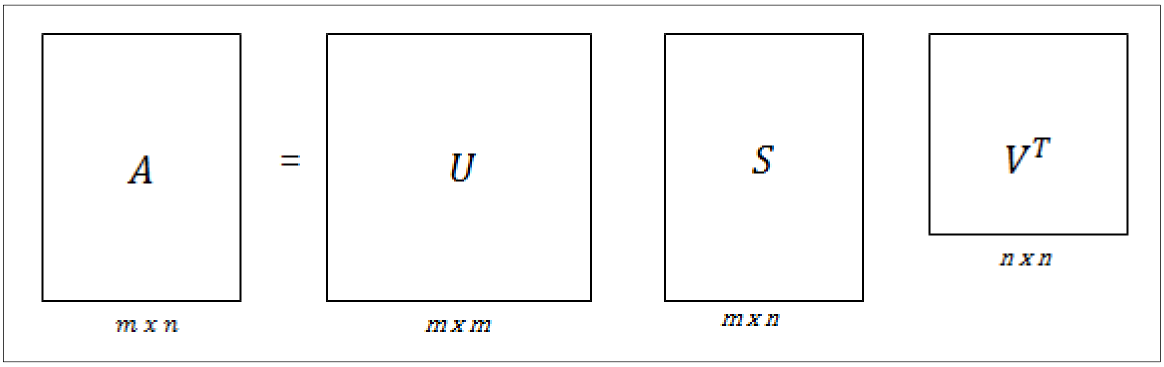


Figure 3.3 Singular Value Decomposition of matrix A

$$A_k = U_k * S_k * V_k^T \tag{3.10}$$

A shortened SVD is used by LSA which means that k largest singular values and related vectors are represented in the reduced space given in the formula (3.10). Term vectors in this reduced space are given in the rows of U_k, document vectors are given in the V_k. The number of reduced dimension is depicted by the parameter named d. In the developed framework, the model was run for different values of d to get best retrieval performance, since there is no absolute rule for selecting this parameter.

3.2.3.2 Latent dirichlet allocation (LDA)

Topic is defined as allocation over a definite vocabulary. Topics models were developed for detecting topics, hidden variables, which occur in a collection of documents called corpus. Generally, documents involve more than one topic with different proportions. Observing topics of documents directly is not possible, so they are called hidden variables. In topic models, topics are defined over generated words by the model. Topic models are generative models which aim to propose a model in a mathematical framework by which analyzing documents and detecting topics of documents based on word statistics could be performed. It is a useful methodology for interpreting structure of data information. It was first described in information retrieval, although it has been applied in various application areas such as visualization, statistical inference and bioinformatics. The first topic model, called Probabilistic Latent Semantic Analysis/Indexing (pLSA or PLSI), was proposed by Hoffman et al. [80] as an alternative method to Latent Semantic Analysis/Indexing (LSA/LSI). In this model, each document can be generated by only one topic. Moreover, pLSA has some deficiencies such as it does not have probabilistic model at the level of documents and it does not provide any generative model for document representation which is the list of numbers (mixing proportions for topics). Thus, when size of corpus grows, number of model parameters also increase that cause overfitting problem. Therefore, Latent Dirichlet allocation (LDA) was introduced by Blei et al. [81] to handle the problems of pLSA.

In recent years, with the rapid evaluation of topic models, researchers have started to use topic models in the field of bioinformatics. Due to its achievement in the analyzing large scale data, it has become a preferable approach in this field. As in many research areas, detecting hidden knowledge from the data structure is a significant research problem should be addressed by researchers in the field of bioinformatics. Clustering, classification and feature extraction of biological data are main tasks for using topic models. There are some studies which use topic model in analyzing biological data. For example, Caldas et al. [26] studied LDA in retrieving microarray genomic data. In this study, the microarray samples are represented as vector of number of differentially expressed genes and each experiment corresponds to a document which consists of different topic distributions. Moreover, Chen et al. [82–84] used LDA for analyzing gene sequence data. DNA sequences are represented by k-mer frequencies and each sequence corresponds to a document while each k-mer corresponds to a word. The aim of the study is to extract topics distributions for each genome sequence. Chen et al. [85] also studied LDA model with background distribution (LDA-B) in discovering functional groups. LDA-B is an extension of LDA which is constructed by adding background distribution of shared functional elements. In addition to this, La Rosa et al. [86] proposed a new alignment-free method based on Probabilistic Topic Modeling for genome sequences. They represented sequence experiments with using fixed length k-mers and applied LDA to classify genome sequences with different sequence length.

In this study, LDA was used as a second fingerprinting technique to be applied for k-mer frequency vectors of the experiments. The model terms are defined such as [81];

- Vocabulary is a vector of words; $\{1, \ldots., V\}$. Words are shown by unit vectors such that v^{th} word in the vocabulary is shown as w^{v} =1 and w^{u}=0 for $u \neq v$.
- There are N words in each document. Words are given as $\boldsymbol{w} = \{w_1, w_2, \ldots., w_N\}$ in which w_n is the n^{th} word in the document.
- There are M documents in the corpus $D = \{\mathbf{w_1}, \mathbf{w_2}, \ldots., \mathbf{w_M}\}$.

It is assumed that the corpus consists of M metagenomic experiments and T topics. A k-mer is represented by w, while there is a sample d contains of K k-mers shown by $d = \{w_1, w_2, \ldots., w_K\}$. In addition to this, a topic is a distribution over the k-mers of the samples.

$$P(w_i) = \sum_{i=1}^{T} P(w_i | z = z_j) P(z = z_j) \tag{3.11}$$

All metagenomics experiment are represented by generated topics with the probability distribution given in (3.11). $P(w_i)$ is the probability of a k-mer w_i in a given document, while selecting a k-mer from topic z_j for the current sample is represented by $P(z = z_j)$. Furthermore, probability of sampling a k-mer given the topic z_j is defined as $P(w_i|z = z_j)$. In summary, firstly topics are specified before any data is defined in the model. Generation process of each k-mer is performed in two steps;

- A distribution is selected randomly over topics.
- For each k-mer in the experiment
    - A topic is selected randomly from the topics in step 1.
    - A k-mer is selected from the corresponding distributions over the vocabulary.

The application of LDA model is given in Figure 3.4. Firstly, k-mer extraction of each sample is performed. Then, the model is implemented with different number of topics. Model fitting is performed by Gibbs Sampling [88] as defined in [87] which recommend a value of $50/k$ for α and 0.1 for β, where k represents the number of topics and α and β are model hyperparameters. After this process, each sample is represented by generated topic distributions. In addition to this, number of topics is determined experimentally, because there is no efficient way for setting it.

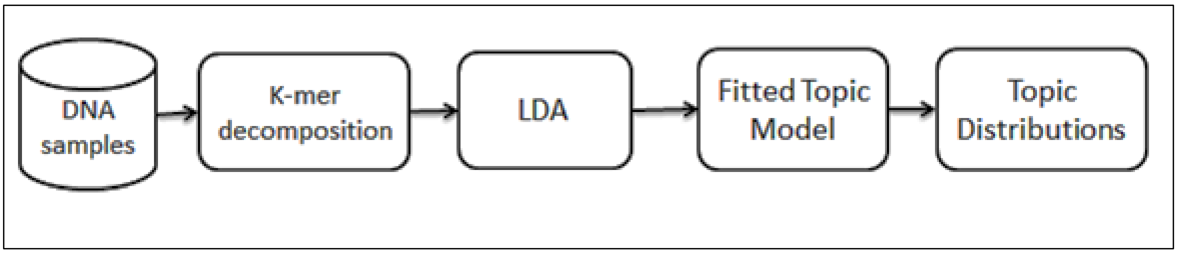


Figure 3.4 LDA steps in the proposed retrieval system

3.2.4 Nullomer analysis

Nullomers, so-called absent k-mers, are substring of sequences that do not occur in a sequence read. It is a well-known fact that every possible k-mer may not occur in concerned read, so nullomers of compared samples may provide valuable information about similarity or dissimilarity between them. If any two experiment share common absent k-mers, they may have some biological relevance. Therefore, nullomer analysis provides detecting structure of a sequence with investigating a question: whether their existence is a statistical matter or outcome of any feature of the sequence [89]. There has been a recent interest in nullomer analysis in DNA sequences over the past decade. Nullomer analysis can provide valuable information various biological researches like drug target identification, environmental monitoring and forensic applications [90].

In order to discover relevance between experiments, the set of absent sequences of a given size were computed. In this study, nullomer analysis consists of two sections such as simple nullomer analysis and 1-order, so-called high order, nullomer analysis. In the first section, after having detecting absent k-mers, with a given size, for each experiment, absent k-mer vectors are generated. Then, transformation of these vectors to binary vectors, in which value of 1 refers to an absent k-mer, the value of 0 represents the k-mer occurs in the current experiment, was performed. After transformation process, the related similarity coefficient was calculated to get similarity scores between compared experiments.

The second section in nullomer analysis is high order nullomer analysis introduced by Vergni and Santoni [89]. This study is an extension of nullomer analysis to investigate structure of nullomers in depth. High order nullomers are absent short sequences whose mutated sequences are still nullomers. Those nullomers are named as first order, second order nullomers etc. because, a short sequence (a k-mer) is not in the whole genome with its possible one letter or two letter mutations. In this study, 1-order nullomers of each experiment was investigated. For instance, if there is a short sequence $S = ACGAATTAGGGCCTGAG$, it is very easy to recognize that all sequences in length of 2 is present in the sequence, but there are some absent sequences in length of 3 e.g AAA, TTT. These absent sequences are called simple nullomers. Regarding first order nullomers, they occur in length of 4 in sequence S such as nullomer $AAAA$ is a first order nullomer since all possible sequences obtained with the mutation of one nucleotide are absent sequences.

In order to understand the implementation of the first order nullomers for comparing experiments, a simple example is defined as follows; Let A and B are compared experiments, 1-order nullomers of them are represented by X and Y subsets of nullomers given as; $X = \{ACGTAA, ATGGAT\}$, $Y = \{ACCGTA, CCCGAGC\}$. For each nullomers of a given sample, the set of dinucleotides (dinuc) at each position was considered: let k be the size of considered sub-words, the first dinucleotide starts at position 1 (first and second nucleotide), then shifting along the sequence the second dinucleotide starts at position 2 (second and third nucleotide) and so on till the last dinucleotide starting at position $(k-1)^{th}$ and k^{th} nucleotide). For each dinucleotide position the related dinucleotide frequencies were computed.

For given experiments frequency distributions are calculated as follows;

<u>For experiment A;</u>
dinuc 1-2: AC: 0.5, AT: 0.5 and others dinuc are 0
dinuc 2-3: CG: 0.5, TG: 0.5 and others dinuc are 0 etc.
….
dinuc (n-1)-(n): AA: 0.5, AT:0.5 others are 0
<u>For experiment B;</u>
dinuc 1-2: AC: 0.5, CC: 0.5 and others dinuc are 0
dinuc 2-3: CC: 0.5, CG: 0.5 and others dinuc are 0 etc.
….
dinuc (n-1)-(n): TA: 0.5, GT:0.5 others are 0

Finally, (k-1) distributions for sample A and (k-1) distributions for sample B were obtained. In order to compute the distance between a couple of metagenomic samples, the similarity of dinucleotide composition at each position of the related set of nullomers (1-order nullomers) was evaluated by applying the Jensen Shannon Divergence (JSD). For each dinucleotide position, the JSD was calculated between two related frequency distributions and then (k-1) JSD were summed up. Two distance values for each sample pairs were obtained, the former refers to simple nullomers results and the latter represents 1-order nullomers results.

3.2.5 Fingerprint comparison

Similarities between experiments are calculated in two ways: direct comparison of frequency vectors and calculating similarity scores between obtained fingerprints with an appropriate similarity metric. The similarity metric differs according to the used fingerprint extraction method.

$$D_{sqrt}(X,Y) = \sum_K (\sqrt{f_n(k,X)} - \sqrt{f_n(k,Y)})^2 \tag{3.12}$$

$$D_{log}(X,Y) = \sum_K (log\big(1 + f_n(k,X)\big) - \log(1 + f_n(k,Y)))^2 \tag{3.13}$$

Variance-stabilized (VS) (3.12) and Log transformed (LT) Euclidean distances (3.13) of compared experiments was performed to make direct comparison of frequency vectors. Let X and Y be two compared fingerprints; represents frequency of the k-mer k in X is given by $f_n(k,X)$, while $f_n(k,Y)$ refers to frequency of the same k-mer in Y. The score ranges between 0 and 1, 0 refers to the most similar experiments, while the score closing to 1 represents decreasing similarity.

$$similarity = cos(\theta) = \frac{X \cdot Y}{||X||||Y||} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{3.14}$$

In addition to this, fingerprints generated with LSA method were compared with Cosine distance. The obtained score (3.14) ranges between 0 and 1; if vectors are exactly same vectors their distance becomes 1, while they have no similarity the score becomes 0.

Kullback-Leibler (KL) divergence was used compare probability distributions, generated by LDA model, of experiments. KL divergence has been commonly used in data mining and pattern recognition [91]. It is not a symmetric metric and generates a non-negative distance value which takes the value of 1 if the compared objects are exactly same.

$$D_{KL}(P||Q) = \sum_{i=1}^{n} p_i log\frac{p_i}{q_i} \tag{3.15}$$

$P = (p_1, p_2, \ldots, p_n)$, $Q = (q_1, q_2, \ldots, q_n)$ for $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} q_i = 1$

Let P and Q be two probability distributions, KL divergence between them is shown by $D_{KL}(P||Q)$ (3.15). Both $D_{KL}(P||Q)$ and the average value of $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$ were calculated to observe that how the retrieval results are influenced by those different approaches, since the KL divergence is not a symmetric metric.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.16}$$

In the nullomer analysis, Jaccard and Jensen Shannon (JS) divergence were used. Jaccard coefficient was used for simple nullomer analysis, while JS divergence was used for 1-order nullomers analysis. Jaccard index between absent k-mer vectors, named A and B, of the experiments to be compared is calculated as given in the formula (3.16). Those vectors are binary vectors in which 1 represents an absent k-mer, 0 corresponds a k-mer occurs in the corresponding experiment. The ratio of number of common absent k-mers to the number of all k-mers gives the Jaccard index.

$$JS(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{3.17}$$

$$where\ M = \frac{1}{2}P + \frac{1}{2}Q$$

$$JS(P||Q) = -\sum Mlog(M) + \frac{1}{2}\sum Plog(P) + \frac{1}{2}\sum Qlog(Q) \tag{3.18}$$

$$JS(P||Q) = H(M) - \frac{1}{2}H(P) - \frac{1}{2}H(Q)$$

$$where\, H(P) = -\sum_{i=1}^{n} p_i \log(p_i)$$

Furthermore, Jensen Shannon (JS) divergence (3.17), (3.18) was used in the analysis of 1-order nullomers. The JS divergence is symmetrized and smoothed version of the KL divergence. Shannon entropy of the three distributions M, P and Q was used to calculate JS divergence. It actually measures the how these distributions are separable from each other.

Furthermore, there is an important point that needs attention when applying LDA model in the proposed retrieval system; the model generates an output named as “final.gamma“. It is a distribution matrix in which rows represents experiments; the columns represent the topics generated by the model. Produced topic distributions should be normalized such as the alpha parameter is subtracted from each row entry in the matrix then the row is renormalized to total value of them is 1. In this way, the distribution is a normal distribution that has equal mean, median and mode. Furthermore, various distributions (Poisson, Binomial etc.) based on the overall structure of the data can be used in finding KL divergence.

3.3 Results

This section consists of four sub-section such as; Data, Evaluation Criteria, Empirical Results and Implementation. The first sub-section gives information about whole metagenome sequencing samples dataset, second sub-section describes evaluation criteria of the proposed system and the third sub-section gives empirical results. The final sub-section describes implementation.

3.3.1 Data

In this study, a real human metagenomic dataset called Type 2 diabetes (T2D) [61] was used to evaluate the performance of the proposed system. The dataset contains human gut microbiota samples collected from 199 individuals, 100 of them are healthy people and 99 of them patients with type 2 diabetes. There are different phases in the dataset named phase I and phase II, phase II was selected, since its coverage is higher than the other phase type. The dataset size is about 1 terabyte and Illumina Genome Analyzer technology was used to get sequencing samples. The sequence data generated by this technology should be evaluated whether raw reads are in good or bad quality. A quality threshold is applied to eliminate nucleotides that have the quality value less than the threshold. The quality threshold is a widely used metric for assessing the accuracy of sequence read generation process. In this study, a quality threshold of 30 was applied to get base pairs that have good quality.

3.3.2 Evaluation criteria

In this chapter as in the first chapter of the study, the assessment of the system performance was performed by defined “ground truth”. Relevance definition between experiments is the first step in this process. In the data collection, positive samples are defined as patients with type 2 diabetes and negative samples are healthy people. If two samples are retrieved from the same class that is to say positive class, they are marked as relevant, otherwise they become irrelevant samples. The system aims at retrieving relevant samples which are patients with the same disease with the query.

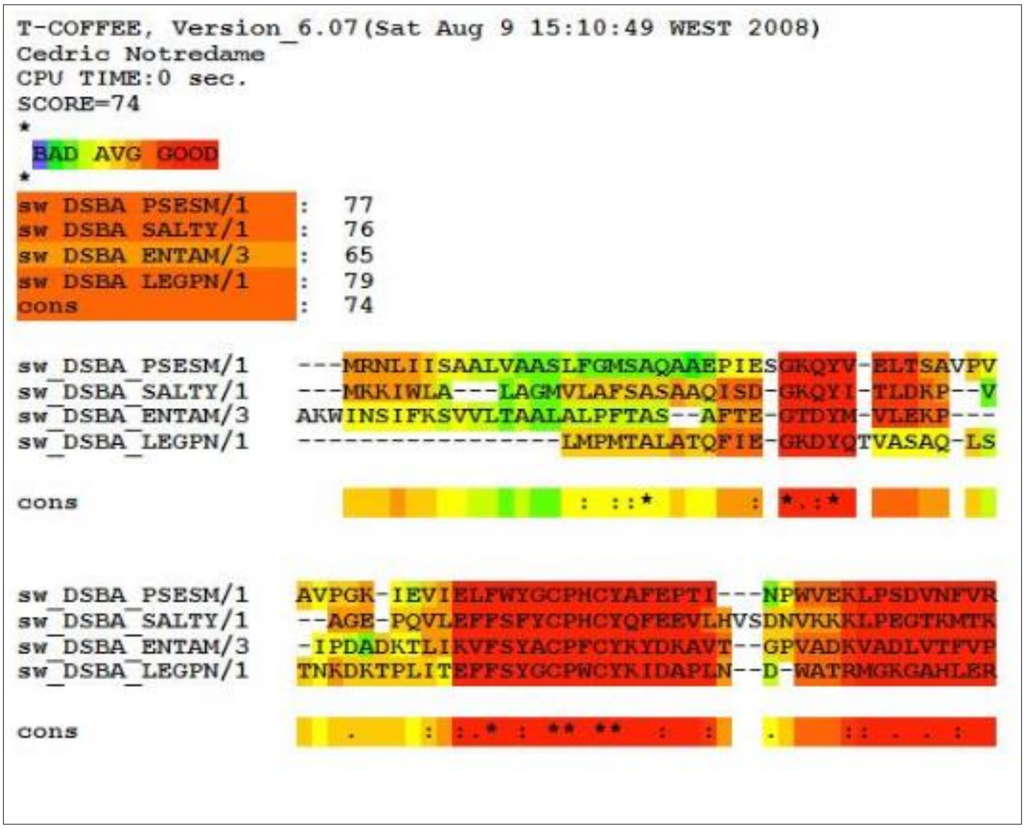
$$Precision\ (n;q) = \frac{number\ of\ samples\ relevant\ to\ q\ in\ n\ retrieved\ samples}{n} \quad (3.19)$$

$$MAP = \frac{1}{|Q|}\sum_{q\in Q} AveP(q) \quad (3.20)$$

$$AveP(q) = \frac{1}{m_q}\sum_{n\in R_q} Precision(n;q) \quad (3.21)$$

In order to test the system retrieval performance Mean average precision (MAP) (3.20) was used. It is a commonly used metric in information retrieval. The retrieval system produces a ranked list in ascending order based on obtained similarity scores for a query q. At the top of the list the most similar experiments are located. When calculating MAP score of the each query, precision (3.19) and average precision $AveP$ (3.21) are calculated as given below. $Precision$ is calculated using top n samples; $n \in \{1,2,\ldots.N\}$. In the formulas, Q represents the set of all queries, locations of relevant samples in the ranked list is given by R_q and m_q is the number of relevant samples to the query. Higher value of MAP indicates better retrieval performance.

Furthermore, multiple sequence alignment was used for assessing the topic distribution process of LDA model. Multiple sequence alignment (MSA) was applied to discover whether sequences in same topic have some similarity. In LDA model, firstly topics are generated then k-mer assignment to a topic is performed. Sequences in same topic are expected to similar to each other, while sequences in different topics are expected to dissimilar from each other. At this point, sequence alignment approaches provide valuable information about similarities between sequences. These approaches are efficient techniques to be applied in variety of bioinformatics analyses such as structure prediction, detecting structure/sequence similarity or phylogeny. There are two basic forms of those approaches; pairwise sequence alignment MSA. The former is used for aligning two sequences, while the latter is used for the alignment of three or more biological sequences of DNA, RNA or protein. The main goal of the MSA methods is to obtain maximal matching between characters of input sequences in terms of a scoring function. Over the past decade, plenty of MSA algorithms and programs have been developed to enhance alignment results. All these algorithms study on the same problem using various ways. Computational costs and alignment accuracy of the algorithms are the main issues in finding suitable algorithm for a specific sequence dataset. Until now, overall outcomes indicate that there is no perfect MSA method, because each method has its own some strengths and weakness based on the problem being addressed. In this study, a MSA method [92] which is an extension of a heuristic algorithm [93] was used to build multiple alignments between sequences both in the same and different topics. The method aims at optimizing the consistency between multiple alignments by combining the output of fifteen widely used MSA methods. To provide one consensus alignment it computes multiple alignments of the given sequences. The method generates a colored version of final alignment which indicates an agreement between all used MSA methods. A sample output of protein sequences alignment is given in Figure 3.5. Red regions show perfect agreement; blue regions refer to weak agreement, while green and yellow regions should be used with caution. As can be seen from the figure, each residue is colored separately with respect to the alignment of that specific residue. The score, called CORE (Consistency of the Overall Residue Evaluation) index, given in the top, which is the average consistency score indicates quality of the alignment. It ranges between 0 and 100 and sometimes may be scaled to the range between 0 and 1000. The higher the score, the more reliable the alignment is. A star (*) indicates an entirely conserved column in MSA, a gap is represented by indicator of (-).



```
T-COFFEE, Version_6.07(Sat Aug 9 15:10:49 WEST 2008)
Cedric Notredame
CPU TIME:0 sec.
SCORE=74
*
 BAD AVG GOOD
*
sw_DSBA_PSESM/1   :  77
sw_DSBA_SALTY/1   :  76
sw_DSBA_ENTAM/3   :  65
sw_DSBA_LEGPN/1   :  79
cons              :  74

sw_DSBA_PSESM/1   ---MRNLIISAALVAASLFGMSAQAAEPIESGKQYV-ELTSAVPV
sw_DSBA_SALTY/1   ---MKKIWLA---LAGMVLAFSASAAQISD-GKQYI-TLDKP--V
sw_DSBA_ENTAM/3   AKWINSIFKSVVLTAALALPFTAS--AFTE-GTDYM-VLEKP---
sw_DSBA_LEGPN/1   -----------------LMPMTALATQFIE-GKDYQTVASAQ-LS

cons                                : ::*      : *.:*


sw_DSBA_PSESM/1   AVPGK-IEVIELFWYGCPHCYAFEPTI---NPWVEKLPSDVNFVR
sw_DSBA_SALTY/1   --AGE-PQVLEFFSFYCPHCYQFEEVLHVSDNVKKKLPEGTKMTK
sw_DSBA_ENTAM/3   -IPDADKTLIKVFSYACPFCYKYDKAVT--GPVADKVADLVTFVP
sw_DSBA_LEGPN/1   TNKDKTPLITEFFSYGCPWCYKIDAPLN--D-WATRMGKGAHLER

cons                 .    : :.* : ** **  :  :   .    :: . . :
```

Figure 3.5 Sample output of used MSA method[6]

Besides aligning k-mers in same and different topics, motif discovery and motif comparison processes were used to evaluate the topic generation process of the LDA model. Motif-based sequence analysis methods have been widely used for sequence comparison, because set of sequences are assumed to have common sequence pattern if they are similar sequences. A sequence motif, so-called fixed length pattern or conserved area, is defined as a part of DNA or protein sequence which is in a specific structure. A motif in DNA represents a protein-binding site, while a motif in protein represents a basic unit of protein folding. Those motifs, which have structural and biological significance, can be used to observe evolutionary and functional relationships between sequences. They are seen as candidates for functionally important sites. Identifying and characterization of such motifs play an important role to understand the structure of cellular processes, such as mechanisms of diseases, in the molecular biology. As given in Figure 3.6, firstly unaligned sequences are taken as input to a motif discovery algorithm and motifs are discovered among given sequences. Then the discovered motifs are

[6] Quoted from: http://www.tcoffee.org/Projects/mcoffee/#COMMANDS. (01.08.2018)

searched through a known database that consists of known motifs. The searching process is performed by a motif comparison algorithm to find similarity between the query motif and motif collection. Finally, aligned motifs are obtained. In this manner, the relevance between the query motif and known motifs can be discovered.

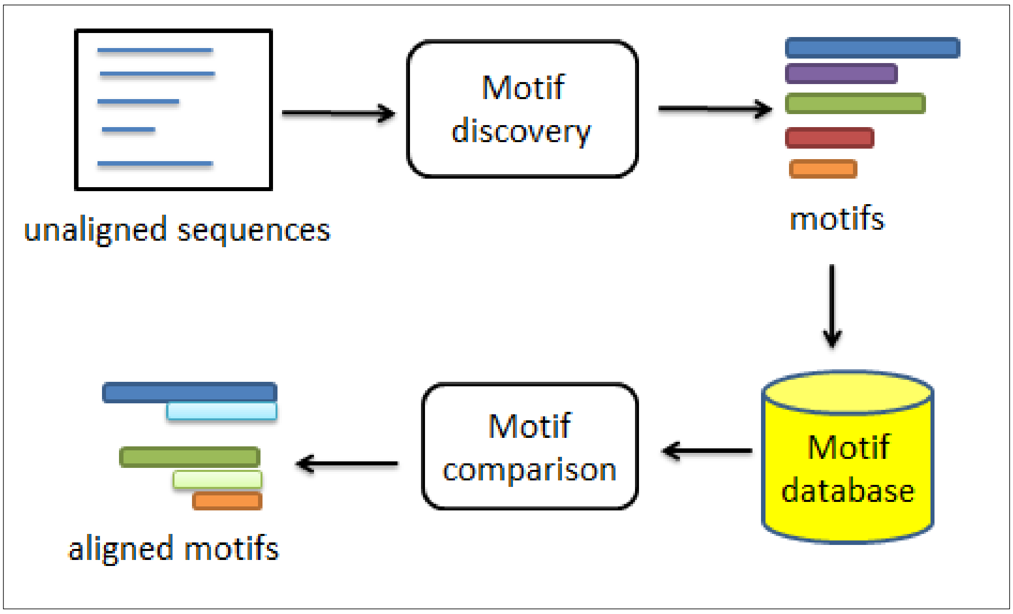
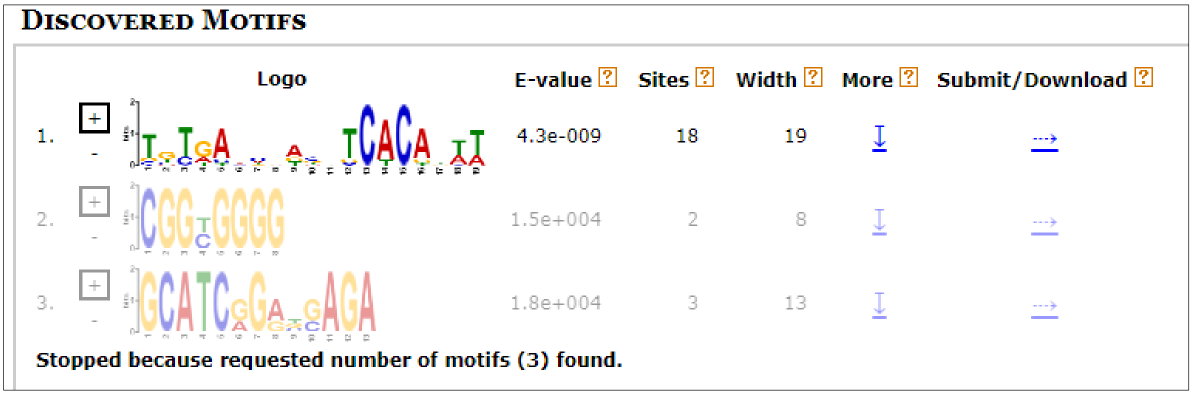


Figure 3.6 General view of the motif discovery and motif comparison processes

Motif discovery aims at finding short similar sequences that occur repeatedly in as many as sequences. Motif discovery problem has been solved by different manners until now. In this study, an algorithm [94], which uses an expectation maximization technique, was used to find motifs among k-mers in topics generated by the LDA model. The algorithm gives results of discovered motifs with a sequence logo, e-value, sites, and width information as given in Figure 3.7. The sequence logo is the graphical representation of displaying discovered motifs. For DNA sequences different colors are used such as red, blue, orange and green represents nucleic acids A, C, G and T respectively. The height of the each character in the logo depends on its relative frequency at the given position. In addition to this, y-axis represents the amount of information measured in bits. The second output is e-value which represents the statistical significance of the motif. The e-value is calculated using log-likelihood ratio, width and sites information of the motif. It is an estimate of the expected number of motifs with the given log-likelihood ratio (or higher), and with the same width and site count, that one would find in a similarly sized set of random sequences. The method ranks motifs based on their e-value, motifs with a low e-value, which has the most statistically significant, are given at the top of the list, while the motifs, which have e-values greater than 0.05, are displayed partially transparent. The other output parameters are sites and width; the former represents a number of sites contributing to the structuring of the motifs, the latter corresponds to the width of the motif.



DISCOVERED MOTIFS

		Logo	E-value ?	Sites ?	Width ?	More ?	Submit/Download ?
1.	⊞ -		4.3e-009	18	19	↧	⇢
2.	⊞ -		1.5e+004	2	8	↧	⇢
3.	⊞ -		1.8e+004	3	13	↧	⇢

Stopped because requested number of motifs (3) found.

Figure 3.7 Sample output of the used motif discovery algorithm

Furthermore, a method [95] was used for searching for similar motifs with the discovered motifs through the database of known motifs. The algorithm searches a query motif within a database and reports a ranked list of motifs according to statistical significance score between the query and the target motif. The result also contains an optimal alignment of two given motifs. The ranked lists of motifs are transcription factors (TF); each of them has a model, data source and TF family information. TFs are proteins that bind a specific DNA sequence to regulate gene expression. Transcription process, which contains basic information to make a protein, is defined as copying DNA sequence of a gene into RNA molecule. TFs are the key points for performing logic operations of information to decide whether to express a gene.

3.3.3 Empirical results

The proposed framework was evaluated using the dataset described in Section 3.3.1. There are 199 metagenomic experiments, 99 of them are positive samples, the others are negative samples. Retrieving relevant samples is mainly aimed regarding to the relevance definition. To this end, the first process is k-mer frequency calculation for k values between 2 and 13. After that, direct comparison of the frequency vectors was performed using LT and VS Euclidean distances. MAP scores of these distances based on different k values are depicted in Figure 3.8. The obtained scores indicate that LT and VS distances have similar performances in retrieving relevant samples. It is clearly seen that there is an obvious increment in MAP scores with the increasing values of k, because greater value of k value helps to better represent experiment content in the feature space. Table view of the obtained results was given in Table A.1.

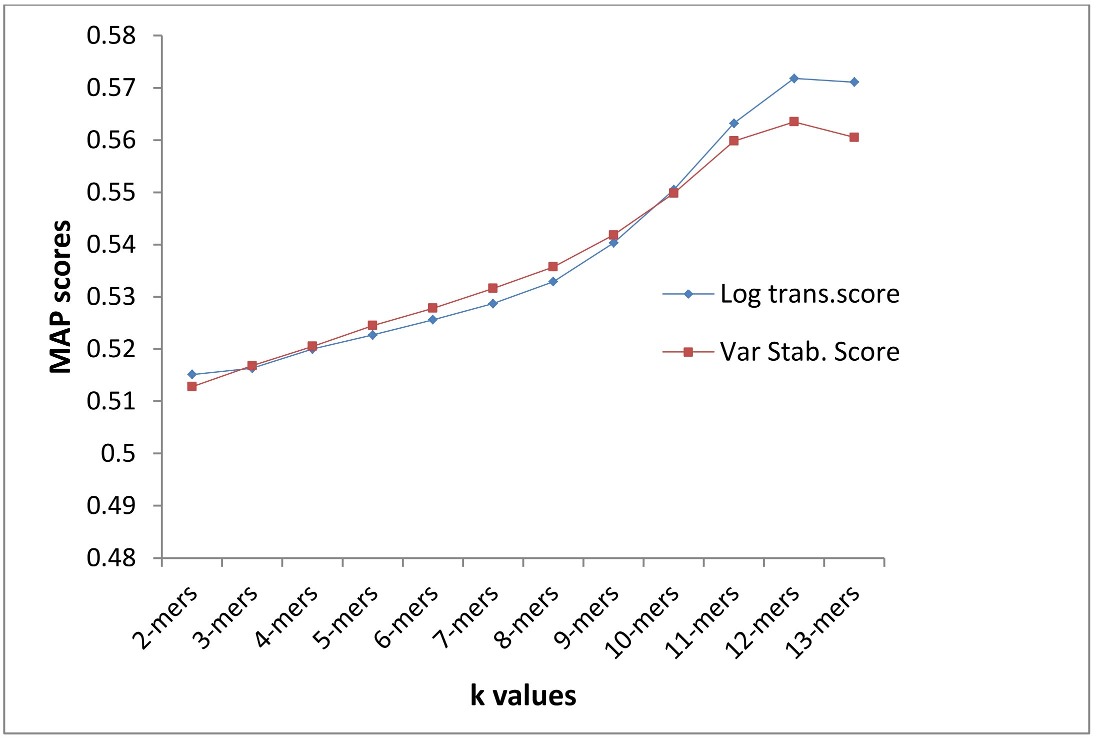


Figure 3.8 MAP scores of the Log transformed and Variance-stabilized Euclidean distances

In addition to the direct comparison of frequency vectors, two different fingerprinting approaches were performed in this study. LSA retrieval performance is given in Figure 3.9. MAP scores were computed for several values of d parameter, number of reduced dimensions, such as 10, 15, 20, 25 and 30. MAP scores were not computed for d >10, because 2-mer vector size is equal to 10. It is clearly seen that the best retrieval performance was achieved at d=10 for 2-mers among all k-mers. For 2-mers, it is much better to perform retrieval process with all k-mers, though it is not possible to use all k-mers for high k values, due to excessive increment in vector size. Furthermore, there is an exponential increment in k-mer vector size for k>6, so feature selection methods were performed to decrease the computational cost. As can be seen from the figure, the fingerprinting method performs well in general with the parameter d=15 for all k values, since the highest average MAP was observed at this value. Table view of the results was given in Table A.2.

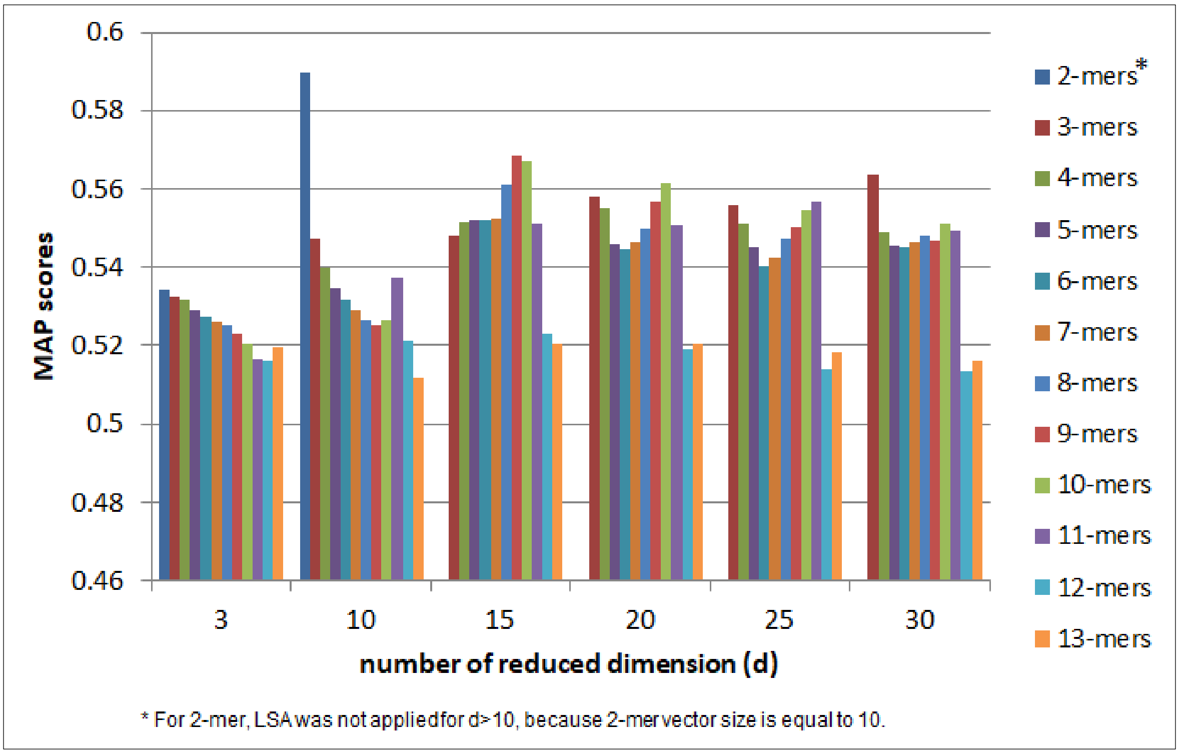


Figure 3.9 MAP scores of LSA fingerprint extraction method for different d values

LDA method was applied as the second fingerprinting approach in the proposed framework. There are some parameters, alpha (α), number of topics (k) and iteration number $(iter)$, to be defined before application of the model. In this study, these parameters were set as experimentally, since there is no known rule about the parameter selection. The MAP score which is greater than the score of direct comparison was selected as the final performance score for each k-mer. Moreover, the excessive growth of k-mer vector size for great k values has become the feature selection methods the basic need for efficient retrieval process. To this end, tf-idf, CAE, and combinatorial approach were performed to decrease the computational cost for k>6. When applying tf-idf method for 12-mers and 13-mers, a modified version of tf-idf approach was performed. Due to the fact that experiments have distinct k-mers for greater value of k, size of vocabulary used in LDA model has been growing extremely. Actually, the size of vocabulary is not expected to be too large, it is considered that it should be proportional to the k-mer vector size. In order to avoid this increase, the approach of “select only the terms which occur in maximum number of documents” was performed for the k-mer selection of 12-mers and 13-mers. Firstly, tf-idf scores of k-mers were calculated, then k-mers were sorted in descending order based on the obtained scores. After that, occurring number of k-mers among all experiments in the corpus was calculated and k-mers occurred in the maximum number of experiments were selected. Thus, a vocabulary with a considerable size was obtained to perform LDA model efficiently. LDA retrieval performance was given in Figure 3.10, beside that detailed information of the model parameters, feature selection methods, numbers of k-mers were given in Table A.3. The highest MAP scores of each run for each k value were given in bold in the table. As can be seen from the results, there is an obvious increment in the performance of LDA for k values between 2 and 9, but it has not achieved in finding relevant samples for k>9. This case can be explained that selecting informative k-mers by used feature selection methods has not become successful in the vector space model.

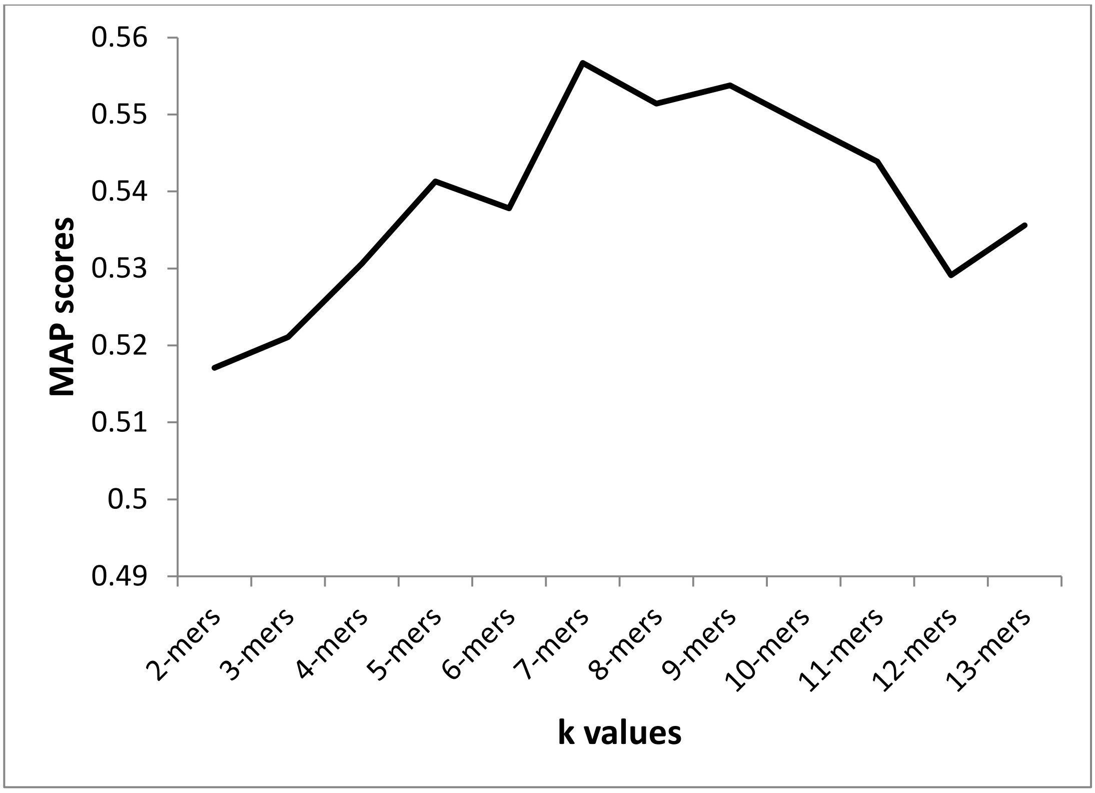


Figure 3.10 MAP scores of LDA fingerprint extraction method for different k values

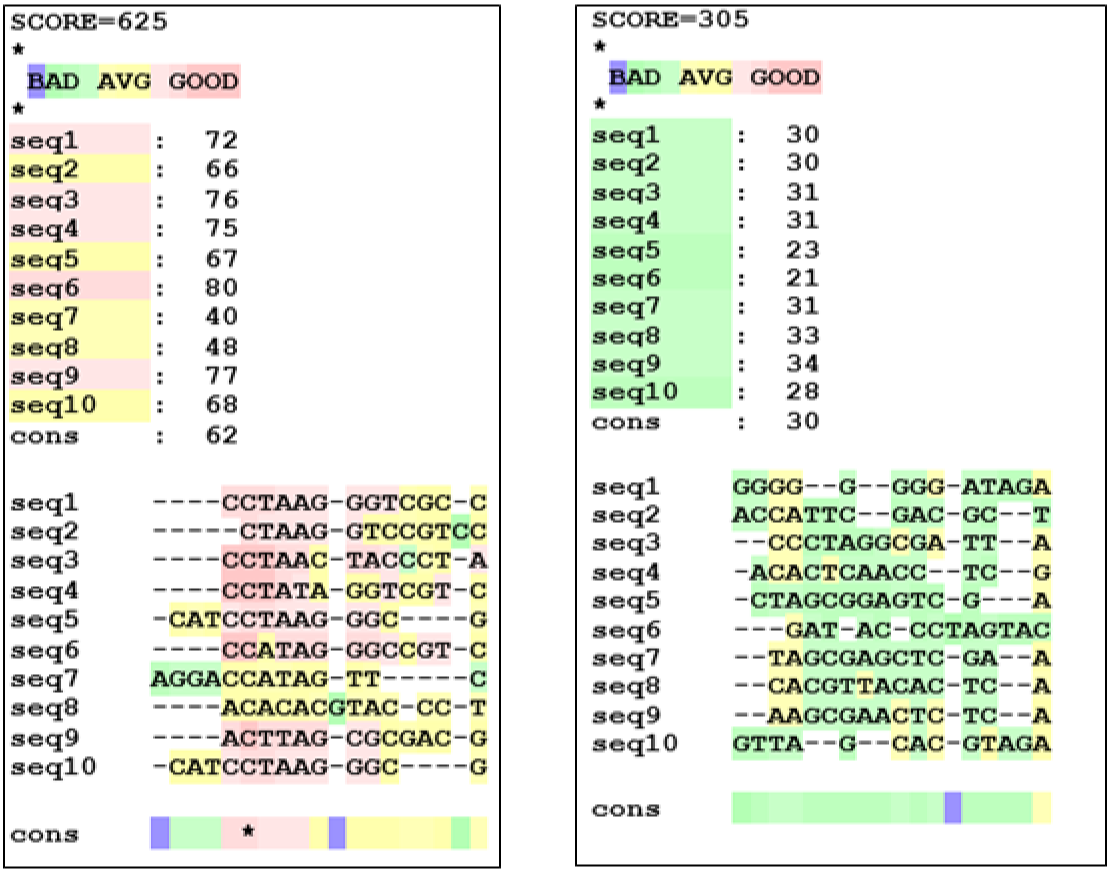
Moreover, to evaluate the robustness of the LDA model with respect to the different runs, the model was run ten times for 7-mers. After getting MAP scores for each run, a standard deviation of obtained MAP scores, the value of 0.0022, was calculated. The value shows that MAP scores are close to each other. In other words, the LDA model is a robust model in retrieving similar sequence samples. In addition, the combinatorial feature selection method was only applied for 7-mers to test the method. Best retrieval performance for 7-mers was achieved with this FS method.

As stated previously, LDA model assigns each k-mer to the one of the obtained topics. Topic level-distributions of 13-mers were used for evaluation of this assignment. Because of high k value stores more information than other k-mers, k=13 was selected for the evaluation of the model. Table 3.1 depicts the 13-mer lists for five obtained topics. K-mers in same topic are expected to have some biological similarities. Hence, MSA algorithm was performed for both the sequences in topic-1 and sequences from different topics given in the Table 3.1. Figure 3.11 gives MSA results which consist of two different cases; alignments of sequences from same topic (topic-1) are given in the first part (a) and alignments of sequences from different topics are given in the second part. To build multiple alignments for these cases, a MSA tool described in the Section 3.3.2 was used. In the figure, colored MSA’s with CORE (Consistency of the Overall Residue Evaluation) index scores are given. The topmost score is the average consistency score for each sequence. It is predicted that alignment score of sequences in same topic should be more consistent than results of sequences from different topics. According to the result, a larger consistency score (625) for the first case is obtained than the other case (305). This means that sequences in same topic are closer to each other than the sequences in different topic. Furthermore, numbers of red regions, which show a perfect agreement between the used methods, in topic-1 are higher than the regions in the different topics.

Table 3.1 Top ten ranked 13-mers for first five generated topics by LDA model

	topic-1	**topic-2**	**topic-3**	**topic-4**	**topic-5**
seq1	CCTAAGGGTCGCC	CCCTAGGAGCAGA	CCTAGCATCCCAG	CTAGCGGCTATAG	CCTAACTACCCTA
seq2	CTAAGGTCCGTCC	ATCTATCCCCCCC	GACCTCACACGTA	CAACCTAGCCGTC	CCCTAGGCGATTA
seq3	CCTAACTACCCTA	GATCCTAACCAGC	AGACTTAGGACCC	AGAGATGTGTCCC	AGTCAACCCCGAG
seq4	CCTATAGGTCGTC	CACGCGATGTGTA	ACCCTAGCCCGAA	CCGCACTAGGCAC	ACGAGACCTCTTA
seq5	CATCCTAAGGGCG	AGTCCGTCGCTAG	AGTTGGGTACCCG	AGTAACCGACTAA	AGGACCATAGTTC
seq6	CCATAGGGCCGTC	CTAGCGGAGTCGA	CTAGCGTGGCAAG	ACCAGCTAGGGCT	CTATAGTTGTACA
seq7	AGGACCATAGTTC	GACGTCTCAGTTA	CCTAATGAGGGAC	CCCCCTTAACCCC	AGTCTCGCGAGCA
seq8	ACACACGTACCCT	AACACTACACGTA	ACACTCAACCTCG	ACTTGAGTCTCTA	ACTTAGCGCGACG
seq9	ACTTAGCGCGACG	CCTAGTCAGCAGG	TAGGACCCACATA	GAACCCCTACTGA	CGGATAGCTAGAA
seq10	CATCCTAAGGGCG	CACGTTAGTTGGA	CAGCCCTAGTTCG	AGTTGTACGACTA	CTAAGGGTTAAAC

A motif was discovered among the sequences occurred in topic-1. It is a 9-base long sequence and its logo is given in Figure 3.12. The motif was discovered by e-value of 1.5E-006 which means that it is a statistically significant motif. Besides this, it is assumed that motifs with small e-values (e.g. less than 0.001) are very unlikely to be random sequence artifacts. The obtained e-value is very smaller than the specified threshold, the discovered motif does not occur in randomly among sequences. Moreover, as can be seen from the given logo, C and T bases generally occur in the first two positions in the discovered motif. Motif locations based on each sequence is given in the appendix in the Figure A.1. The figure demonstrates the motif site locations. The position and strength of the motif were represented by individual blocks, while the significance of the site is depicted by the height of each block. The height is calculated to be proportional to the negative logarithm of the p-value of the site, truncated at the height for a p-value of 1E-10.



```
SCORE=625
*
 BAD  AVG  GOOD
*
seq1      :   72
seq2      :   66
seq3      :   76
seq4      :   75
seq5      :   67
seq6      :   80
seq7      :   40
seq8      :   48
seq9      :   77
seq10     :   68
cons      :   62

seq1      ----CCTAAG-GGTCGC-C
seq2      -----CTAAG-GTCCGTCC
seq3      ----CCTAAC-TACCCT-A
seq4      ----CCTATA-GGTCGT-C
seq5      -CATCCTAAG-GGC----G
seq6      ----CCATAG-GGCCGT-C
seq7      AGGACCATAG-TT-----C
seq8      ----ACACACGTAC-CC-T
seq9      ----ACTTAG-CGCGAC-G
seq10     -CATCCTAAG-GGC----G

cons             *
```

(a)

```
SCORE=305
*
 BAD  AVG  GOOD
*
seq1      :   30
seq2      :   30
seq3      :   31
seq4      :   31
seq5      :   23
seq6      :   21
seq7      :   31
seq8      :   33
seq9      :   34
seq10     :   28
cons      :   30

seq1      GGGG--G--GGG-ATAGA
seq2      ACCATTC--GAC-GC--T
seq3      --CCCTAGGCGA-TT--A
seq4      -ACACTCAACC--TC--G
seq5      -CTAGCGGAGTC-G---A
seq6      ---GAT-AC-CCTAGTAC
seq7      --TAGCGAGCTC-GA--A
seq8      --CACGTTACAC-TC--A
seq9      --AAGCGAACTC-TC--A
seq10     GTTA--G--CAC-GTAGA

cons
```

(b)

Figure 3.11 (a) MSA result of 13-mers in topic-1 and (b) MSA result of 13-mers in different topics

To find similar motifs with the discovered motif in known motif databases, motif comparison method was performed. After having selected Human DNA database, searching for similar motifs was applied for the discovered motif. According to the optimal alignment results, most significant matches with the target motif were generated by the method. The obtained motifs are transcription factors which are given in Table 3.2. Each motif is given with its model name and transcription factor name.

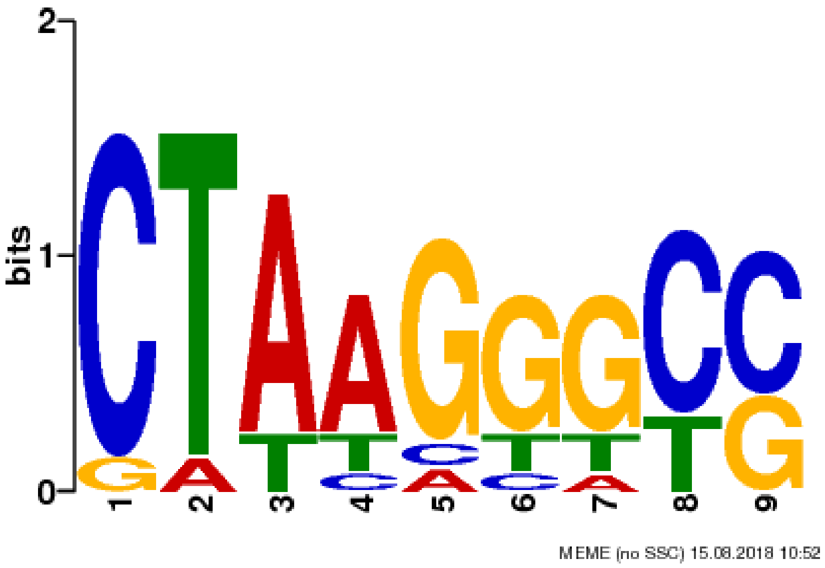


Figure 3.12 Sequence logo of the discovered motif

Table 3.2 Transcription factor list of the discovered motif

Model	**Transcription Factor**
ISL1_HUMAN.H11MO.0.A	ISL1
HXB4_HUMAN.H11MO.0.B	HOXB4
SOX9_HUMAN.H11MO.0.B	SOX9
ARNT_HUMAN.H11MO.0.B	ARNT
ETV1_HUMAN.H11MO.0.A	ETV1

In this study, LSA and LDA were used to obtain fingerprints of the experiments. Retrieval performance of those methods was compared to the direct comparison of the frequency vectors. The comparative results of these methods and direct comparison were given in Figure 3.13. In direct comparison, Log score and Var score Euclidean distance performances were given separately. The obtained results depicts that LSA has achieved in detecting similar metagenomic experiments for k<11. In addition to this, LDA has close performance with LSA method for k values between 5 and 8. It is also clearly observed that direct comparison of frequency vectors has become more successful rather than fingerprinting techniques for k>11. It should be noted that direct comparison k-mer frequency vectors rely on only time and space factors, though a fingerprinting approach needs a proper feature selection method in addition to these factors. Thus, a new research interest should be addressed has been emerged that is out of the scope of this study. According to the experimental results, LSA and LDA methods can be used efficiently in transforming experiment content in the feature space and they have promising results in detecting relevance information within samples for small k values.

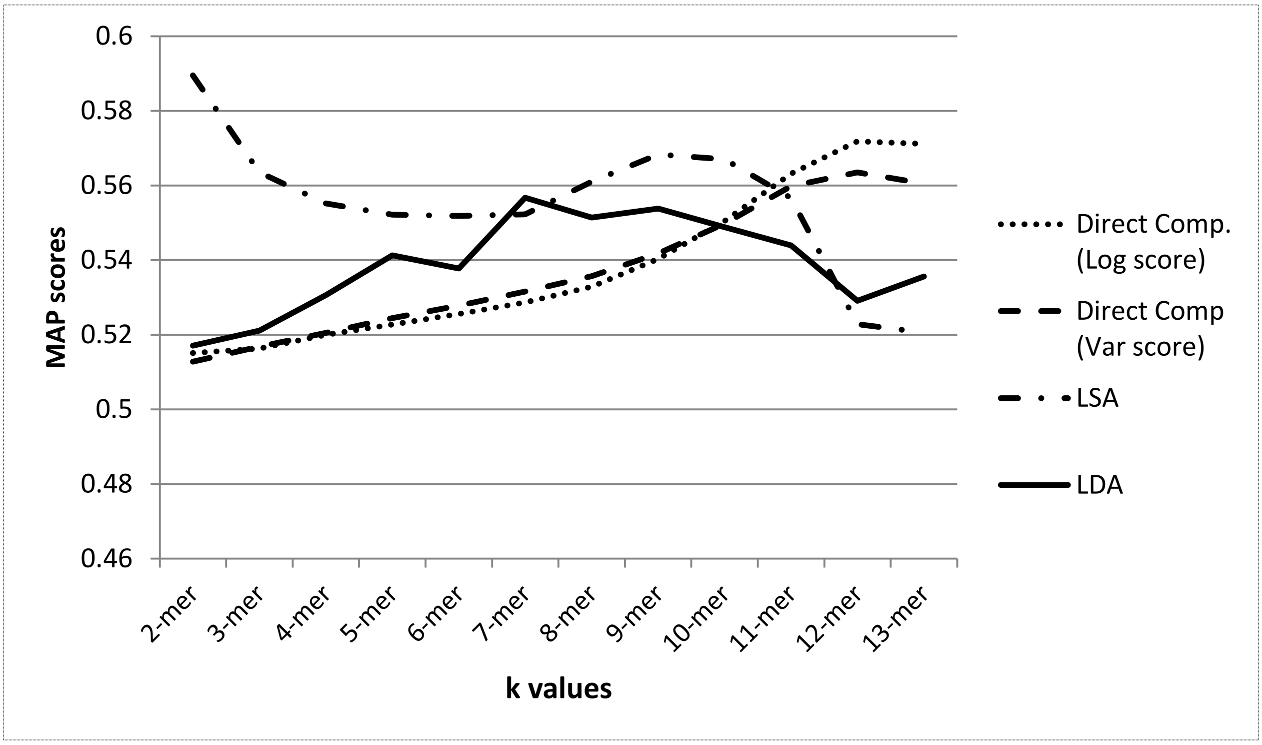


Figure 3.13 Comparative results of LSA and LDA fingerprint extraction methods with direct comparison by using Log score and Var score

In this study, the dataset from the study of Seth et al. [70] was used to test the proposed system. Seth et al. [70] has extracted great k values (30-mers) to detect similarities between metagenomics samples, while in this study the maximum k value is 13. However, any selection process was not applied, a lower score than the study of Seth et al. was achieved by the direct comparison considering 12-mers.

Nullomer analysis involves two parts; simple nullomer analysis and 1-order nullomer analysis. In order to perform nullomer analysis, firstly absent k-mers were discovered for each experiment and binary vectors transformation was performed. In these vectors an element of the vector takes the value of 1 if the related k-mer is absent, otherwise it takes the value of 0 for the corresponding experiment. The dataset has absent k-mers for k values greater than or equal to 11. Distances between binary vectors of experiments were calculated with Jaccard coefficient. As showed in Table 3.3, Jaccard scores are not good because, experiments share limited absent k-mers. In addition to this, first-order nullomer analysis was performed. It is note that nullomers appears at length of 11, while first-order nullomers appear at length of 14. For first-order nullomer, two different coefficient called Jaccard and Jensen Shannon divergence were used. According to the results given in Table 3.4, there is a little difference between MAP scores of two coefficients. First-order nullomers could not provide efficient retrieval performance as well as simple nullomers.

Table 3.3 MAP scores of Jaccard coefficient for absent k-mers

Absent k-mers	**MAP**
11-mers	0.4799
12-mers	0.5067
13-mers	0.5030

Table 3.4 MAP scores of Jaccard and JS coefficient of the 1-order nullomers

	Coefficient	**MAP**
14-mers	Jaccard	0.4949
	JS	0.4829

As stated previously, fingerprinting approach LSA outperforms the other methods for k values between 2 and 10 in retrieving relevant experiments from the data collection. Wilcoxon Signed Rank and Paired t-test were performed to observe whether differences between direct comparison and fingerprinting approaches were statistically significant. P-value of 8.99E-07 and 2.39E-07 between LSA and direct comparison using Wilcoxon Signed Rank and Paired t-test were obtained respectively. P-value of 1.26E-04 and 1.02E-04 were calculated in comparing results of LSA and LDA. It is clearly seen that obtained p-values are below the threshold of 0.05. That is to say, these values support that LSA has become more successful in detecting similarities rather than other methods. The tests were also used to discover the statistical significance between retrieval by fingerprint technique and retrieval by random and p-value of 8.83E-07 and 2.07E-08 were obtained. It can be concluded that fingerprinting technique based on MAP scores is statistically significant. To this end, the success of retrieval by fingerprinting approaches has statistical significance as against the retrieval by random.

3.4 Implementation

The system implementation was done using C++, R and MATLAB and it was tested on Windows platform. Boost 1.64 and zlib were used as external libraries. Some supplementary files (executable files, documentation and test data files) are available in the link: www.baskent.edu.tr/~hogul/WMS_retrieval.rar. The proposed framework can be used for any dataset.

4. CONCLUSION

Over the recent years, a massive volume of genomic data has been accumulated mainly in public repositories. Rapid increase of such data raises an important need for efficient data analysis approaches that should be addressed by researchers. Due to the fact that current database applications provide meta-data based search which has limited searching options for retrieval of genomic data, content-based search approach has become an alternative solution recently. In this thesis study, developing content-based retrieval frameworks using different data types and perspectives was aimed to retrieve relevant experiments from genomic databases. The study contains two main chapters which are time-series experiment retrieval and whole-metagenome sequencing sample retrieval. The retrieval frameworks aim at designing and development of targeted sub-models, creation of suitable comparing mechanisms, evaluating the developed models with real datasets.

The first chapter, Time-Series Experiment Retrieval, to the best our knowledge, is the first study that builds fingerprints of experiments using all time-series expression profiles for comparing experiments. In fingerprint extraction, four different methods such as Differentially Expression Profile-based, Transition Model-based, Time Warping and Lyapunov Exponent methods were used. According to the experimental results, Time Warping method is a promising fingerprinting approach in detecting relevance between time-course experiments. The system performance was evaluated based on ROC scores. It can be observed that the proposed system has become successful in retrieving relevant experiments with high ROC scores. The results also point that using whole time-course experiments as a query and obtaining fingerprints with all expression values is an efficient approach for detecting relevance between experiments. Moreover, the system retrieval performance was assessed with an indirect evaluation based on gene-sets and direct evaluation based on manual annotations of experiments. Thus, discovered relevance between compared experiments with fingerprinting techniques was verified by these assessment approaches. In addition to this, adapting the proposed framework for large data collections that consists of a variety of organisms and platforms will be our future work.

In the second chapter, called Whole Metagenome Sequencing Sample Retrieval, a framework using novel fingerprinting approaches, LSA and LDA, was proposed. The proposed system contains k-mer extraction, selection, fingerprint extraction and comparison processes. To extract fingerprints, performing the application of data mining algorithms is the novelty of the study in this field. According to the experimental results, LSA is an efficient fingerprinting technique to represent the experiment content and detecting relevance information between compared experiments. It is also observed that LSA method outperforms LDA and direct comparison of frequency vectors for k<11, though the slight decrement has been seen in the performance at increasing value of k. Besides this, direct comparison has better retrieval performance rather than fingerprinting methods for k>10. Computational cost is the main challenge of this study, that is to say LDA fingerprinting technique took almost 1 week to work especially for k>10. In addition to this, there are some issues such as; the value of k affects precision and efficiency of results directly and direct comparison method for high k values is not reliable. Therefore, feature selection algorithms were applied for high k values to overcome specified issues.

The proposed framework indicates the adaptability of text mining techniques in extracting fingerprints of metagenomic experiments. The obtained results clearly showed that used fingerprinting approaches have encouraging results to represent experiments in a feature space for finding similarities between them. Furthermore the study has two biological contributions. The first one is that if two samples are assumed to be relevant, they should have similarity between their experiment content. The experimental results have confirmed this idea by means of fingerprinting approaches and similarity metrics. In addition to this, LDA model presents the second contribution which is that sequences in same group have evolutionary relationships such as having similar biological functions or sharing common ancestor. As a conclusion, the results guide us to a new motivation for the developed system to give more efficient retrieval results with high k values.

According to the observations of experimental results, biological relevance between experiments can be detected without being dependent on any user-defined textual annotation. As against traditional meta-data-based retrieval techniques, the proposed models can provide more intelligent search strategies. Empirical results lead to researchers to apply the proposed models in current database searching implementations. Moreover, the fingerprinting approaches and similarity metrics for different types of genomic data presented in this study is expected to provide a new perspective for future implementations. Finally, the proposed retrieval frameworks can be used in a laboratory environment, so biological knowledge extracted from experiments can be used in building new hypothesis.

REFERENCES

[1] OĞUL, Hasan, Content-Based Retrieval of Microarray Experiments, Pattern Recognition in Computational Molecular Biology: Techniques and Approaches. ELLOUMI, M. John Wiley & Sons, p.315–334, 2015.

[2] BARRETT, Tanya and EDGAR, Ron., Mining Microarray Data at NCBI's Gene Expression Omnibus (GEO), Gene Mapping, Discovery, and Expression, New Jersey: Humana Press, p.175–190, 2006.

[3] PARKINSON, H. et al. , ArrayExpress-A public database of microarray experiments and gene expression profiles, Nucleic Acids Research, Vol.35, no.SUPPL.1, p.747–750., 2007.

[4] BENSON, Dennis et al., GenBank, Nucleic acids research, Vol.28, no.1, p.15–8., 2000.

[5] MEYER, Folker et al.,The Metagenomics RAST Server: A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes, Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches, Vol.8, p.325–331, 2011.

[6] SHIER, Rosie, Mathematics Learning Support Centre Statistics: Paired t-test. 2004.

[7] WHITLEY, Elise and BALL, Jonathan, Statistics review 6: Nonparametric methods. Critical Care, Vol.6, no.6, p.509–513, 2002.

[8] IMAM, Akeyede, USMAN, Mohammed and CHIAWA, Moses Abanyam, On Consistency and Limitation of paired t-test, Sign and Wilcoxon Sign Rank Test, IOSR Journal of Mathematics, Vol.10, no.1, 2014.

[9] FARINA, Lorenzo, SANTIS, Alberto De, MORELLI, Giorgio and RUBERTI, Ida, Dynamic Measure of Genes Co-regulation, Vol.1, no.1, p.10–17, 2007.

[10] ERNST, Jason, NAU, Gerard J. and BAR-JOSEPH, Ziv. ,Clustering short time series gene expression data. Bioinformatics. ,Vol. 21, no. SUPPL. 1, p. 159–168. ,2005.

[11] WANG, X-D, QI, Y.-X., JIANG, Z.-L., Reconstruction of transcriptional network from microarray data using combined mutual information and network-assisted regression, IET systems biology, Vol. 5, no. 2, p.95–102, 2011.

[12] QIAN, Liwei, ZHENG, Haoran, ZHOU, Hong, QIN, Ruibin and LI, Jinlong. , Classification of Time Series Gene Expression in Clinical Studies via Integration of Biological Network PLoS ONE, Vol.8, no.3, 2013.

[13] LIN, T.-h., KAMINSKI, N. and BAR-JOSEPH, Z., Alignment and classification of time series gene expression in clinical studies, Vol. 24, no.13, p.i147–i155.

[14] MARAZIOTIS, Ioannis A., DRAGOMIR, Andrei, BEZERIANOS, Anastasios, Gene networks reconstruction and time-series prediction from microarray data using recurrent neural fuzzy networks, IET systems biology, Vol.1, no.1, p.41–50, 2007.

[15] BANSAL, Mukesh and DI BERNARDO, Diego, Moment-closure approximations for mass-action models, IET systems biology, Vol.2, no.2, p.64–72, 2008.

[16] WANG, Yaqun, XU, Meng, WANG, Zhong, TAO, Ming, ZHU, Junjia, WANG, Li, LI, Runze, BERCELI, Scott A. and WU, Rongling, How to cluster gene expression dynamics in response to environmental signals, Briefings in Bioinformatics, Vol.13, no.2, p.162–174, 2012.

[17] LIU, Zhenqiu, HSIAO, William, CANTAREL, Brandi L., DRÁBEK, Elliott Franco and FRASER-LIGGETT, Claire, Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data, Bioinformatics. ,Vol. 27, no.23, p.3242–3249, 2011.

[18] HUNTER, Lawrence et al., GEST: a gene expression search tool based on a novel Bayesian similarity metric, Bioinformatics (Oxford, England). ,Vol. 17 Suppl 1, p. S115-22. ,2001.

[19] HORTON, Paul B, KISELEVA, Larisa and FUJIBUCHI, Wataru, RaPiDS: an algorithm for rapid expression profile database search, Genome informatics, International Conference on Genome Informatics, Vol.17, no.2, p.67–76, 2006.

[20] FUJIBUCHI, Wataru, KISELEVA, Larisa, TANIGUCHI, Takeaki, HARADA, Hajime and HORTON, Paul, CellMontage: Similar expression profile search server, Bioinformatics, Vol.23, no.22, p.3103–3104, 2007.

[21] CHEN, Rong, MALLELWAR, Rohan, THOSAR, Ajit, VENKATASUBRAHMANYAM, Shivkumar and BUTTE, Atul J., GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed, BMC bioinformatics, Vol.9, p.548, 2008.

[22] HIBBS, Matthew A., HESS, David C., MYERS, Chad L., HUTTENHOWER, Curtis, LI, Kai and TROYANSKAYA, Olga G., Exploring the functional landscape of gene expression: Directed search of large microarray compendia, Bioinformatics. ,Vol. 23, no.20, p.2692–2699, 2007.

[23] ENGREITZ, Jesse M., CHEN, Rong, MORGAN, Alexander A., DUDLEY, Joel T., MALLELWAR, Rohan and BUTTE, Atul J., Profilechaser: Searching microarray repositories based on genome-wide patterns of differential expression. Bioinformatics, Vol. 27, no.23, p.3317–3318, 2011.

[24] ENGREITZ, Jesse M, DAIGLE, Bernie J, MARSHALL, Jonathan J and ALTMAN, Russ B, Independent component analysis : Mining microarray data for fundamental human gene expression modules, Journal of Biomedical Informatics, Vol.43, no.6, p.932–944, 2010.

[25] BELL, Francis and SACAN, Ahmet, Content based searching of gene expression databases using binary fingerprints of differential expression profiles, 7th International Symposium on Health Informatics and Bioinformatics., p.107–113, 2012.

[26] CALDAS, José, GEHLENBORG, Nils, FAISAL, Ali, BRAZMA, Alvis and KASKI, Samuel, Probabilistic retrieval and visualization of biologically relevant microarray experiments. Bioinformatics, Vol.25, no.12, p.145–153, 2009.

[27] SUTHRAM, Silpa, DUDLEY, Joel T., CHIANG, Annie P., CHEN, Rong, HASTIE, Trevor J. and BUTTE, Atul J., Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets, PLoS Computational Biology, Vol.6, no.2, p.1–10, 2010.

[28] GEORGII, Elisabeth, SALOJÄRVI, Jarkko, BROSCHÉ, Mikael, KANGASJÄRVI, Jaakko and KASKI, Samuel, Targeted retrieval of gene expression measurements using regulatory models, Bioinformatics, Vol.28, no.18, p.2349–2356, 2012.

[29] AÇICI, Koray, TERZI, Yunus Kasim and OĞUL, Hasan, Retrieving relevant experiments: The case of microRNA microarrays, BioSystems, Vol.134, p.71–78, , 2015.

[30] HAYRAN, Ahmet, OGUL, Hasan and OZKOC, Esma, Content-based search on time-series microarray databases, Proceedings-International Workshop on Database and Expert Systems Applications, DEXA, p.89–93, 2014.

[31] SCHENA, Mark et al., Parallel human genome analysis: microarray-based expression monitoring of 1000 genes, Proceedings of the National Academy of Sciences, Vol.93, no.20, p.10614–10619,1996.

[32] NAU, Gerard J., Human macrophage activation programs induced by bacterial pathogens, Proceedings of the National Academy of Sciences, Vol.99, no.3, p.1503–1508, 2002.

[33] ZHU, Gefeng et al., Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth, Nature, Vol.406, no.6791, p.90–94, 2000.

[34] DEAN, Nema and RAFTERY, Adrian E., Normal uniform mixture differential gene expression detection for cDNA microarrays, BMC Bioinformatics, Vol.6, p.1–15, 2005.

[35] SAHOO, Debashis, DILL, David L., TIBSHIRANI, Rob and PLEVRITIS, Sylvia K. , Extracting binary signals from microarray time-course data, Nucleic Acids Research , Vol.35, no.11, p.3705–3712, 2007.

[36] SAKOE, Hiroaki and CHIBA, Seibi, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.26, no.1, p.43–49,1978.

[37] VELİCHKO, V.M, ZAGORUYKO, N.G., Automatic recognition of 200 words, International Journal of Man-Machine Studies, Vol. 2, no.3, p.223–234,1970.

[38] EFRAT, Alon, FAN, Quanfu and VENKATASUBRAMANIAN, Suresh, Curve matching, time warping, and light fields: New algorithms for computing similarity between curves, Journal of Mathematical Imaging and Vision, Vol.27, no.3, p.203–216., 2007.

[39] TAPPERT, Charles C., SUEN, Ching Y. and WAKAHARA, Toru, The state of the art in online handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.12, no.8, p.787–808,1990.

[40] GU, Jie and JIN, Xiaomin, A simple approximation for dynamic time warping search in large time series database, International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin p. 841–842, 2006.

[41] MULLER, Meinard, Dtw-based motion comparison and retrieval, Information Retrieval for Music and Motion, p.211–226, 2007.

[42] AACH, John and CHURCH, George M., Aligning gene expression time series with time warping algorithms, Bioinformatics, Vol.17, no.6, p.495–508, 2001.

[43] KRUSKAL, J.B and LIBERMAN, M., The symmetric time-warping problem: from continuous to discrete, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, p.125–161, 1999

[44] MCCUE, Leigh S. and TROESCH, Armin W., Use of Lyapunov exponents to predict chaotic vessel motions, Fluid Mechanics and its Applications, Vol.97, p.415–432, 2011.

[45] ECKMANN, J.P. and RUELLE, David, Ergodic theory of chaos and strange attractors, In The Theory of Chaotic Attractors, p.273–312, 1985.

[46] FARMER, J.D and SIDOROWICH, J.J., Predicting chaotic time series, Physical review letters, Vol.59, no.8, p.845,1987.

[47] GRASSBERGER, Peter and PROCACCIA, Itamar, Characterization of strange attractors, Physical review letters, Vol.50, no.5, p.346,1983.

[48] ROSENSTEIN, Michael T., COLLINS, James J., DE LUCA, Carlo J., A practical method for calculating largest Lyapunov exponents from small data sets, Physica D: Nonlinear Phenomena,Vol.65, no.1–2, p.117–134,1993.

[49] BORIAH, Shyam, CHANDOLA, Varun, KUMAR, Vipin, Similarity measures for categorical data: A comparative evaluation, In: Proceedings of the 2008 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, p.243-254, 2008.

[50] ROGERS, David J. and TANIMOTO, Taffee T., A Computer Program for Classifying Plants, Science, Vol.132, no.3434, p.1115-1118, 1960.

[51] LOURENÇO, Fernando, LOBO, Victor and BAÇÃO, Fernando., Binary-based similarity measures for categorical data and their application in Self- Organizing Maps, p.1-18, 2004.

[52] SUBRAMANIAN, Aravind et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences, vol.102.43, p.15545-15550, 2005.

[53] FAWCETT, Tom, An introduction to ROC analysis, Pattern Recognition Letters. ,Vol.27, no.8, p.861-874, 2006.

[54] KO, Jae-Heung, KIM, Won-Chan and HAN, Kyung-Hwan, Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in Arabidopsis, The Plant Journal, Vol.60, no. 4, p. 649–665, 2009.

[55] VANNESTE, Steffen et al., Cell Cycle Progression in the Pericycle Is Not Sufficient for SOLITARY ROOT/IAA14-Mediated Lateral Root Initiation in Arabidopsis thaliana, The Plant cell, Vol.17, no.11, p.3035-3050, 2005.

[56] DELKER, Carolin et al., Natural Variation of Transcriptional Auxin Response Networks in Arabidopsis thaliana, The Plant cell, Vol.19, no.5, p.1665-168, 2008.

[57] OP DEN CAMP, Roel G.L et al., Rapid induction of distinct stress responses after the release of singlet oxygen in Arabidopsis Plant Cell., Vol.15.10, p.2320-2332., 2013.

[58] YI, Xin, DU, Zhou and SU, Zhen, PlantGSEA: a gene set enrichment analysis

toolkit for plant community, Nucleic acids research, Vol.41.w1, p.98-103, 2013.

[59] QIN, Junjie et al., A human gut microbial gene catalogue established by metagenomic sequencing, Nature, Vol.464, no.7285, p.59-65, 2010.

[60] HUTTENHOWER, Curtis et al., Structure, function and diversity of the healthy human microbiome, Nature, Vol. 486, no. 7402, p. 207-214, 2012.

[61] QIN, Junjie et al., A metagenome-wide association study of gut microbiota in type 2 diabetes, Nature, Vol.490, no.7418, p.55-60, 2012.

[62] http://data.imicrobe.us/.

[63] HUSON, Daniel et al., MEGAN analysis of metagenome data, Gennome Res. ,Vol.17, p.377-386, 2007.

[64] WANG, Qiong, GARRITY, George M., TIEDJE, James M. and COLE, James R. , Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, Applied and Environmental Microbiology, Vol.73, no.16, p.5261-5267, 2007.

[65] SU, Xiaoquan, XU, Jian and NING, Kang, Meta-storms: Efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data, Bioinformatics, Vol.28, no.19, p.2493-2501, 2012.

[66] MAILLET, Nicolas, LEMAITRE, Claire, CHIKHI, Rayan, LAVENIER, Dominique and PETERLONGO, Pierre, Compareads: comparing huge metagenomic experiments. BMC Bioinformatics, Vol.13, no.19, p. S10, 2012.

[67] WHITE, James Robert, NAGARAJAN, Niranjan and POP, Mihai., Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples, PLoS Computational Biology, Vol.5, no.4, 2009.

[68] PARKS, Donovan H. and BEIKO, Robert G., Identifying biologically relevant differences between metagenomic communities, Bioinformatics, Vol. 26, no. 6, p.715-721, 2010.

[69] SEGATA, Nicola, WALDRON, Levi, BALLARINI, Annalisa, NARASIMHAN, Vagheesh, JOUSSON, Olivier and HUTTENHOWER, Curtis., Metagenomic microbial community profiling using unique clade- specific marker genes. Nat Methods, Vol.9, no.8, p.811-814, 2013.

[70] SETH, Sohan, VÄLIMÄKI, Niko, KASKI, Samuel and HONKELA, Antti. ,Exploration and retrieval of whole-metagenome sequencing samples, Bioinformatics, Vol.30, no.17, p.2471-2479, 2014.

[71] WEITSCHEK, Emanuel, SANTONI, Daniele, FISCON, Giulia, COLA, Maria Cristina De, BERTOLAZZI, Paola and FELICI, Giovanni, Next generation sequencing reads comparison with an alignment-free distance, p.1-13, 2014.

[72] POLYCHRONOPOULOS, Dimitris, WEITSCHEK, Emanuel, DIMITRIEVA, Slavica, BUCHER, Philipp, FELICI, Giovanni and ALMIRANTIS, Yannis, Classification of selectively constrained DNA elements using feature vectors and rule-based classifiers. Genomics, Vol.104, no.2, p.79–86, 2014.

[73] DUBINKINA, Veronika B., ISCHENKO, Dmitry S., ULYANTSEV, Vladimir I.,

TYAKHT, Alexander V. and ALEXEEV, Dmitry G. , Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis, BMC Bioinformatics, Vol.17, no.1, p.1-11, 2016.

[74] SPARCK JONES, Karen, A Statistical Interpretation of Term Specificity and Its Application in Retrieval, A Journal of Documentation, Vol.28, no.1, p.11-21,1972.

[75] BERTOLAZZI, Paola et al., Integer programming models for feature selection: New extensions and a randomized solution algorithm, European Journal of Operational Research, Vol.250, no.2, p.389-399, 2016.

[76] WEITSCHEK, Emanuel, LO PRESTI, Alessandra, DROVANDI, Guido, FELICI, Giovanni, CICCOZZI, Massimo, CIOTTI, Marco and BERTOLAZZI, Paola., Human polyomaviruses identification by logic mining techniques, Virology Journal., Vol. 9, p.1-6, 2012.

[77] DUMAIS, Susan T. et al., Using latent semantic analysis to improve access to textual information. In : Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’88, p.281-285, 1988.

[78] DEERWESTER, Scott and DUMAIS, ST., Indexing by latent semantic analysis, Journal of the American society for information science, Vol.41, no.6, p.391,1990.

[79] DE LATHAUWER, Lieven, DE MOOR, Bart, VANDEWALLE and VANDEWALLE, Joos, A multilinear singular value decomposition, SIAM Journal on Matrix Analysis and Applications (SIMAX), Vol.21, no.4, p.1253-1278, 2000.

[80] HOFMANN, Thomas, Probabilistic latent semantic indexing, Annual ACM Conference on Research and Development in Information Retrieval, p. 50–57, ,1999.

[81] BLEI, David, Y.NG, Andrew and JORDAN, Michael., Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol.3, p.993-1022, 2003.

[82] CHEN, Xin, HU, Xiaohua, SHEN, Xiajiong and ROSEN, Gail, Probabilistic topic modeling for genomic data interpretation, Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on, p.149-152, 2010.

[83] CHEN, Xin, HE, Tingting, HU, Xiaohua, ZHOU, Yanhong, AN, Yuan and WU, Xindong. ,Estimating functional groups in human gut microbiome with probabilistic topic models. IEEE Transactions on Nanobioscience. ,Vol. 11, no. 3, p. 203–215. ,2012.

[84] CHEN, Xin, HU, Xiaohua, LIM, Tze Yee, SHEN, Xiajiong, PARK, E. K. and ROSEN, Gail L, Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.9, no.4, p.980-991, 2012.

[85] ENGREITZ, Jesse M, MORGAN, Alexander A, DUDLEY, Joel T, CHEN, Rong, THATHOO, Rahul, ALTMAN, Russ B and BUTTE, Atul J., Content-based microarray search using differential expression profiles, BMC Bioinformatics, 2010.

[86] LA ROSA, Massimo et al., Probabilistic topic modeling for the analysis and classification of genomic sequences, Bmc Bioinformatics, Vol. 16, no.6, p. 9, 2015.

[87] CASELLA, George and GEORGE, Edward, Explaining the Gibbs Sampler, The

American Statistician, vol.46, no.3, p.167-174, 1992.

[88] GRIFFITHS, Thomas L. and STEYVERS, Mark, Finding scientific topics. Proceedings of the National Academy of Sciences, Vol. 101, no. Supplement 1, p.5228–5235, 2004.

[89] VERGNI, Davide and SANTONI, Daniele, Nullomers and high order nullomers in genomic sequences. PLoS ONE, Vol.11, no.12, p.1-15, 2016.

[90] ACQUISTI, Claudia, POSTE, George, CURTISS, David and KUMAR, Sudhir, Nullomers: Really a matter of natural selection? PLoS ONE, Vol.2, no.10, p.2-4, 2007.

[91] JOYCE, JM, Kullback-Leibler divergence. International Encyclopedia of Statistical Science, p.720-722, 2011.

[92] WALLACE, Iain M., O'SULLIVAN, Orla, HIGGINS, Desmond G. and NOTREDAME, Cedric. ,M-Coffee: Combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Research, Vol.34, no.6, p.1692-1699, 2006.

[93] NOTREDAME, Cédric, HIGGINS, Desmond G. and HERINGA, Jaap. ,T-coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology, Vol. 302, no.1, p.205–217, 2000.

[94] BAILEY, Timothy L and ELKAN, Charles, Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers, Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, p. 28-36, 1994.

[95] GUPTA, Shobhit, STAMATOYANNOPOULOS, John A., BAILEY, Timothy L. and NOBLE, William Stafford, Quantifying similarity between motifs, Genome Biology, ,Vol.8, no.2, 2007.

APPENDIX

Table A.1 MAP scores of the of Log transformed (LS) and Variance-stabilized (VS) Euclidean distances

	LS score MAP	**VS score MAP**
2-mers	0.5151	0.5128
3-mers	0.5163	0.5168
4-mers	0.52	0.5205
5-mers	0.5227	0.5245
6-mers	0.5256	0.5278
7-mers	0.5287	0.5316
8-mers	0.5329	0.5357
9-mers	0.5403	0.5418
10-mers	0.5505	0.5498
11-mers	0.5632	0.5598
12-mers	0.5718	0.5635
13-mers	0.5711	0.5605

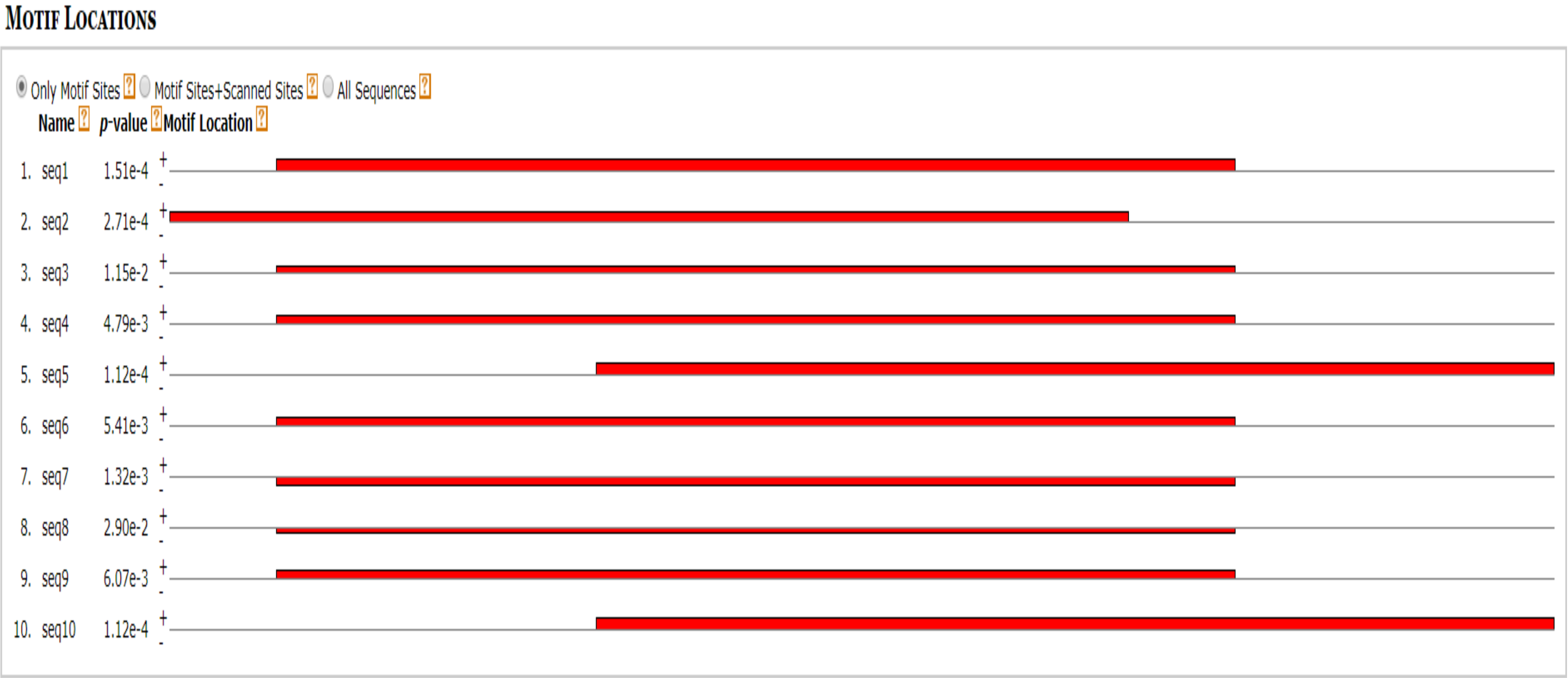
Table A.2 MAP scores of LSA fingerprint extraction method by using k-mer frequency values

	LSA/Cosine score MAP					
	d=3	**d=10**	**d=15**	**d=20**	**d=25**	**d=30**
2-mers	0.5343	**0.5895**				
3-mers	0.5325	0.5474	0.5483	0.5581	0.5558	**0.5635**
4-mers	0.5317	0.5398	0.5515	**0.5552**	0.551	0.5489
5-mers	0.5291	0.5347	**0.5522**	0.5458	0.5449	0.5453
6-mers	0.5275	0.5317	**0.5518**	0.5445	0.5402	0.5451
7-mers	0.526	0.529	**0.5523**	0.5464	0.5424	0.5462
8-mers	0.525	0.5266	**0.561**	0.5497	0.5472	0.5479
9-mers	0.5231	0.525	**0.5683**	0.5568	0.5502	0.547
10-mers	0.5202	0.5264	**0.567**	0.5617	0.5545	0.5513
11-mers	0.5166	0.5372	0.5512	0.5506	**0.5566**	0.5493
12-mers	0.5161	0.5213	0.5228	0.519	0.5138	0.5136
13-mers	0.5196	0.5117	0.5203	0.5206	0.5184	0.5162

Table A.3 MAP scores of LDA fingerprint extraction method for k values between 2 and 13 using different LDA model parameters

								KL divergence	
								MAP	
		FS method	orj nof term	reduced nof term	nof topic (k)	alpha (50/k)	iter	Normal	Symetric
2-mers	**run-1**	No	10	10	5	10	100	0.5123	0.5127
	run-2	No	10	10	5	10	300	**0.5171**	0.5158
	run-3	No	10	10	5	10	500	0.5136	0.5127
	run-4	No	10	10	7	7.14	300	0.5033	0.5037
3-mers	**run-1**	No	32	32	16	3.13	100	0.5161	0.5175
	run-2	No	32	32	16	3.13	300	0.5200	**0.5211**
4-mers	**run-1**	No	136	136	10	5	20	0.5124	0.5129
	run-2	No	136	136	50	1	50	0.5225	0.5219
	run-3	No	136	136	100	0.5	100	**0.5306**	0.5300
	run-4	No	136	136	100	0.5	1000	0.5236	0.5199
5-mers	**run-1**	No	512	512	20	2.5	100	0.5230	0.5227
	run-2	No	512	512	100	0.5	100	0.5348	0.5337
	run-3	No	512	512	100	0.5	250	**0.5413**	0.5400
6-mers	**run-1**	No	2080	2080	50	1	100	0.5342	0.5327
	run-2	No	2080	2080	100	0.5	100	**0.5378**	0.5374
7-mers	**run-1**	No	8192	8192	20	2.5	100	0.5250	0.5242
	run-2	No	8192	8192	100	0.5	100	0.5451	0.5448
	run-3	Yes	8192	705	100	0.5	100	0.5410	0.5389
	run-11	CAE(0.3415)	8192	200	20	2.5	100	0.5382	0.5344
	run-12	CAE(0.3661)	8192	50	20	2.5	100	0.5332	0.5310
	run-13	CAE(0.3939)	8192	10	20	2.5	100	0.5069	0.5066
	run-14	CAE(0.3997)	8192	5	20	2.5	100	0.5158	0.5155
	run-4	Comb. App.	8192	50	20	2.5	100	0.5441	0.5444
	run-5	Comb. App.	8192	50	100	0.5	100	0.5549	**0.5567**
	run-6	Comb. App.	8192	20	20	2.5	100	0.5452	0.5424
8-mers	**run-1**	No	32896	32896	20	2.5	30	0.5289	0.5249
	run-2	No	32896	32896	50	1	50	0.5289	0.5290
	run-3	No	32896	32896	100	0.5	100	**0.5514**	0.5510
	run-6	CAE(0.30)	32896	2630	100	0.5	100	0.5378	0.5355
	run-7	CAE(0.30)	32896	2630	100	0.5	350	0.5505	0.5430
	run-8	CAE(0.20)	32896	16200	100	0.5	100	0.5413	0.5394

	run-1	tf-idf	32896	2000	100	0.5	100	0.5012	0.5009
	run-2	tf-idf	32896	2000	200	0.25	150	0.4979	0.4984
9-mers	**run-1**	CAE(0.35)	131072	1521	100	0.5	100	0.5368	0.5335
	run-2	CAE(0.33)	131072	1521	100	0.5	250	0.5465	0.5374
	run-3	CAE(0.33)	131072	1521	100	0.5	350	0.5521	0.5414
	run-4	CAE(0.35)	131072	3507	100	0.5	155	0.5444	0.538
	run-5	CAE(0.35)	131072	3507	100	0.5	350	**0.5538**	0.5431
10-mers	**run-1**	CAE(0.37)	524801	1500	100	0.5	100	0.5447	0.5373
	run-2	CAE(0.37)	524801	1500	100	0.5	220	**0.5488**	0.5355
	run-3	CAE(0.37)	524801	1500	100	0.5	350	0.5424	0.5316
11-mers	**run-1**	CAE(0.38)	2097153	1501	100	0.5	250	**0.5439**	0.5257
	run-2	CAE(0.38)	2097153	1501	100	0.5	350	0.5263	0.5136
12-mers	**run-5**	tfidf/nof docs	8390656	1112	100	0.5	100	0.5258	0.5271
	run-6	tfidf/nof docs	8390656	1112	50	1	100	0.5256	**0.5291**
13-mers	**run-4**	tfidf/nof docs	33554355	1160	100	0.5	100	0.5282	0.5296
	run-5	tfidf/nof docs	33554355	1160	200	0.25	100	0.5300	0.5158
	run-6	tfidf/nof docs	33554355	1160	50	1	100	0.5360	0.5287
	run-7	tfidf/nof docs	33554355	2020	100	0.5	100	0.5354	0.5352
	run-8	tfidf/nof docs	33554355	2020	200	0.25	100	**0.5356**	0.5300



MOTIF LOCATIONS

○ Only Motif Sites ? ○ Motif Sites+Scanned Sites ? ○ All Sequences ?

	Name ?	*p*-value ?	Motif Location ?
1.	seq1	1.51e-4	
2.	seq2	2.71e-4	
3.	seq3	1.15e-2	
4.	seq4	4.79e-3	
5.	seq5	1.12e-4	
6.	seq6	5.41e-3	
7.	seq7	1.32e-3	
8.	seq8	2.90e-2	
9.	seq9	6.07e-3	
10.	seq10	1.12e-4	

Figure A.1 Motif locations of the discovered motif