

Duygu Dede Şener<sup>1</sup> / Daniele Santoni<sup>2</sup> / Giovanni Felici<sup>2</sup> / Hasan Oğul<sup>1</sup>

# A Content-Based Retrieval Framework for Whole Metagenome Sequencing Samples

<sup>1</sup> Başkent University, Faculty of Engineering, Computer Engineering Department, Ankara, Turkey, E-mail: ddede@baskent.edu.tr

<sup>2</sup> Institute of Systems Analysis and Computer Science “A. Ruberti”, National Research Council, Rome, Italy

## Abstract:

Finding similarities and differences between metagenomic samples within large repositories has been rather a significant issue for researchers. Over the recent years, content-based retrieval has been suggested by various studies from different perspectives. In this study, a content-based retrieval framework for identifying relevant metagenomic samples is developed. The framework consists of feature extraction, selection methods and similarity measures for whole metagenome sequencing samples. Performance of the developed framework was evaluated on given samples. A ground truth was used to evaluate the system performance such that if the system retrieves patients with the same disease, -called positive samples-, they are labeled as relevant samples otherwise irrelevant. The experimental results show that relevant experiments can be detected by using different fingerprinting approaches. We observed that Latent Semantic Analysis (LSA) Method is a promising fingerprinting approach for representing metagenomic samples and finding relevance among them. Source codes and executable files are available at [www.baskent.edu.tr/~hogul/WMS\\_retrieval.rar](http://www.baskent.edu.tr/~hogul/WMS_retrieval.rar).

**Keywords:** Whole-metagenome, sequence retrieval, sequence similarity, k-mer, Latent Semantic Analysis, Latent Dirichlet Allocation, Topic Model

**DOI:** 10.1515/jib-2017-0067


**Received:** October 31, 2017; **Revised:** March 26, 2018; **Accepted:** April 11, 2018

## 1 Introduction

Metagenomes, so-called random community genomes, are used to study microbial communities from various habitats [1]. Metagenomics was defined as “the application of modern genomics technique without the need for isolation and lab cultivation of individual species” by Chen and Pachter [2]. It is the study of genomes recovered from environmental samples and it is one of the developing areas of the molecular biology. Significant information about microbial communities is obtained from metagenomic data; analyzing metagenomic data has thus acquired valuable research interest over the past decades. Targeted studies include phylogenetic profiling with a lower cost; others use whole metagenomes analysis to get more information about metagenomics. Whole metagenome shotgun sequencing (WGS) has provided more complete information to analyze a large volume of data [3]. It also guides the development of new analysis methods for extracting knowledge from metagenomes and for finding relationships between complex sequences.

Finding similarities and differences between metagenomic samples within large repositories has been rather a significant issue for researchers. In recent years, content-based retrieval has been suggested by various studies from different perspectives. Sequence alignment methods are frequently used to identify organisms in whole metagenome sequencing datasets. Those methods provide similarity results by comparing sequence reads and reference sequence sets. Huson et al. [4] developed a tool, so-called MEGAN, to explore the taxonomical content of metagenomics sequence datasets. Wang et al. [5] developed a software package called ribosomal database project (RDP) in which the classifier is able to rapidly classify sequences without aligning them. Liu et al. [6] developed a novel distance-based learning method for multiclass classification and feature selection using metagenomic count data. The method performance was evaluated using different samples. Phenotype-associated taxonomic identification and simultaneous class prediction can be performed by the developed model. Su et al. [7] proposed a novel method, named MetaStorms, for organization and searching for samples. They use taxonomical annotations and phylogenetic structure of the samples to design their search engine based system. The proposed method was successfully applied in database generation and in retrieving similarity between samples.

Duygu Dede Şener is the corresponding author.

 ©2018, Duygu Dede Şener et al., published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

All methods mentioned above need basic existing annotations, known taxa or metabolic information to be applied, but there are some organisms which are unknown or uncultivable. Due to the fact that these unexplored organisms cannot be studied by the referred studies, new reference-free approaches have become a basic need in metagenomics. In comparison with the methods which use alignment against reference sequence sets or known databases of whole genome, analysis techniques based on raw read content provide a wider point of view [8]. Algorithms based on k-mers (substrings of length  $k$ , called n-grams) generated using raw reads have been recently suggested in various studies. In these works, unlike reference based sequence analysis methods, similarity between samples is calculated by comparison of k-mer frequencies. Maillot et al. [9] first developed an algorithm, called Compareads, based on raw read content of the samples for comparing two samples. The similarity between samples is computed using common k-mers and the proposed approach was successful in finding similarities and differences among experiments. Although it is stated that the method is much faster than traditional comparison methods, such as BLAST, in retrieving in very large datasets, it is computationally too expensive to store all k-mer information of the metagenomic samples. Seth et al. [10], instead of storing all k-mers, developed a distributed string mining framework to obtain informative substrings which can be of any length. The authors proposed a data-driven feature extraction and selection method to be applied in retrieving similar samples from metagenomic dataset. In order to detect features of metagenomic data, two classes of samples are compared directly in some studies. Also, many of them make an analysis using a great number of features without a feature selection step [11], [12], [13]. Beside these studies, Qin et al. [14] performed quantification and association testing for almost 5 million predefined genes. Weitschek et al. [15] proposed an alignment-free method using k-mer representation for sequence comparison. In this study, it is shown that the proposed method is a useful technique for reads comparison. Weitschek et al. [16] also get promising classification results of bacterial genomes by combining feature representation and logic data mining algorithms. Dubinkina et al. [8] developed a method based on short k-mers to find pairwise dissimilarity between metagenome samples. In their study, they compared results of reference based methods and proposed k-mer based method on simulated and real metagenome datasets. It was concluded that k-mer spectrum is more efficient than the reference based methods for comparing metagenomic samples and extracting valuable feature information.

In this study, we propose a reference-free and content-based framework for retrieval of whole metagenome sequencing samples. It consists of a feature extraction and selection methods and appropriate similarity measures for finding similarities between metagenomic samples. A metagenome sample is taken as a query and relevant samples from the data collection are listed by order of similarity. The performances of the developed framework are evaluated on a real dataset. A ground truth is used to evaluate the system performance such that if the system retrieves patient with the same disease, called positive sample, they are labeled as relevant samples otherwise irrelevant. Our main focus is to retrieve relevant metagenomic samples.

Our contribution proposes a novel approach to extract fingerprints and is based on two text mining methods, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), which have not been used before in finding similarities between metagenomic experiments. We observed that LSA Method is a promising fingerprinting approach for representing metagenomic samples and to find relevance among them.

## 2 Methods

Given a query experiment, we propose a computational framework that retrieves samples from the metagenomic sample repository that are relevant to the query experiment. The framework consists of k-mer extraction, k-mer selection, fingerprint extraction methods and appropriate similarity measures for calculating similarities between experiments (Figure 1).

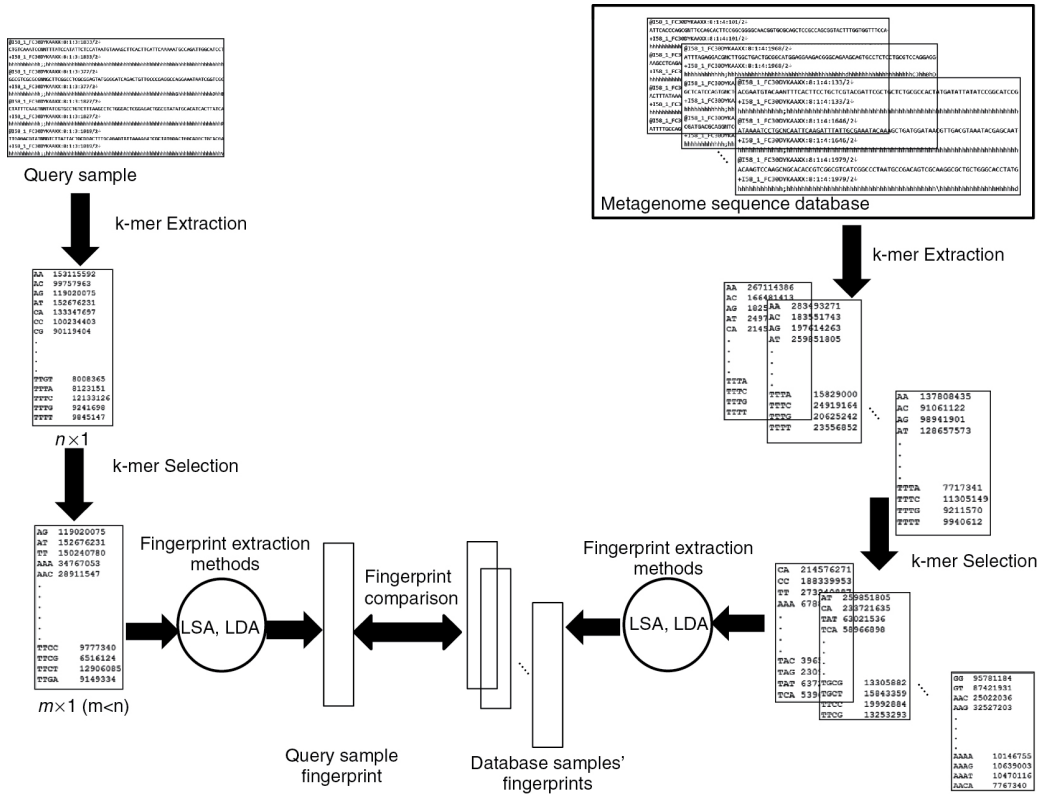


Figure 1: An overview of proposed retrieval framework.

## 2.1 k-mer Extraction

The first step is the extraction of k-mers, (with  $k$ , ranging from 2 to 13), from experiments in order to represent the samples in a feature space. The frequency of each k-mer is computed by counting occurrences of the considered k-mer and dividing the value by total number of k-mers in the experiment. In this process reverse complement of k-mers was considered, as which strand of DNA is sequenced and reads' strand direction are unknown. Therefore, we take both a read and its reverse complement in extracting k-mers from a sequence read. Finally for each couple of reverse complementary k-mers, only the one that firstly occurs, according to lexicographically ordering, is selected and stored.

## 2.2 k-mer Selection

In this study, we used two feature selection (FS) techniques to reduce dimension of feature vectors for high  $k$  (>7) values; Term Frequency Inverse Document Frequency ( $tf-idf$ ) scores and Correlation Attribute Evaluation (CAE).

### 2.2.1 Selecting Features with $tf-idf$ Scores

For each experiment  $tf-idf$  scores are computed for k-mer selection. By this model, we aim to weight k-mers in each experiment [17].

NOO = number of occurrences of  $r$  in experiment  $e$

TNK = total number of k-mers in experiment  $e$

TNE = total number of experiments

NOE = number of experiments which  $r$  occur

$$tfidf_{r,e} = tf_{r,e} * idf_r \tag{1}$$

$$tf_{r,e} = \frac{NOO}{\overline{TNK}} \quad (2)$$

$$idf_r = \log_{10} \frac{TNE}{NOE} \quad (3)$$

To adapt this methodology to our system, each k-mer corresponds to a term; each experiment corresponds to a document in the collection. We calculated *tf-idf* scores of k-mers with regard to related experiment by the formulas given above. In the formula (1),  $tfidf_{r,e}$  is the product of two terms;  $tf_{r,e}$  (2) and  $idf_r$  (3). k-mer frequency ( $tf_{r,e}$ ) represents how frequently a k-mer ( $r$ ) occurs in an experiment ( $e$ ), while inverse document frequency ( $idf_r$ ) measures how important a k-mer is within the dataset collection. After having calculated *tf-idf* scores, we sort k-mers by descending orders according to their *tf-idf* scores and then we select the first  $N$  of them.

### 2.2.2 Correlation Attribute Evaluation (CAE)

In this technique Pearson correlation between attribute and class are calculated to evaluate the worth of attributes. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is obtained via a weighted average. In our retrieval model, we use CAE technique for ranking features in descending order by their correlation between class attribute. Then, we select the first  $N$  of them according to a cut-off value experimentally determined. For each  $k$  ( $>7$ ) value, we performed several runs to observe how the changing number of selected features affect the retrieval performance. When the  $k$  value increases, number of selected features ( $N$ ) increases as well. Best retrieval performances were selected among all runs.

## 2.3 Fingerprint Extraction Methods

After feature extraction and selection processes, fingerprint extraction process is used to represent each metagenomic experiment in the feature space. We aim to study patterns of k-mer features for each experiment by using different models. LSA and a Probabilistic Topic Model, called LDA, were used as fingerprint extraction methods for our retrieval system. We adopted the terms used in text mining such as a k-mers represents text words or terms, metagenomics experiments or samples represent documents, while experiment collection represents the corpus.

### 2.3.1 Latent Semantic Analysis

LSA is a widely used mathematical technique in text mining which aims to detect relationships between documents and terms they contain by using linear algebra techniques. It was first introduced by Dumais et al. [18] and Deerwester et al. [19] as a dimension reduction technique used for information retrieval applications. LSA is composed of four main processes;

- Construction of the term-document matrix, in which each cell of the matrix stores the frequency of the term occurring in the corresponding document.
- Transformation of term-document matrix where elements of the term document matrix are translated to account how important a word is for a document in a collection instead of using raw term frequencies.
- Dimension reduction where singular value decomposition (SVD) is applied on the matrix to get latent structure model. By this process,  $x$  largest singular values are obtained and the others are set as 0.
- Retrieval in Reduced space: after having represented documents and terms reduced dimension, similarities between them are calculated.

In our model, k-mer experiment matrix corresponds to the term document matrix in which each entry represents k-mer frequency for corresponding experiment. LSA has an important parameter, called  $d$ , which is the number of dimensions used in the reduced space. Performance of our retrieval system is evaluated with different values of  $d$ , because there is no strict rule for setting parameter  $d$ .

### 2.3.2 Latent Dirichlet Allocation

Probabilistic Topic Models are used to get semantic information from a set of documents. The main goal in these approaches is discovering topics which represent classes of documents. They are represented by probability distributions over the vocabulary. In our study, we used LDA as a Probabilistic Topic Model to extract fingerprint for each experiment in our corpus. LDA introduced by Blei et al. [20] is used as a generative probabilistic model for a data collection called *corpus*. It is also an effective unsupervised machine learning algorithm.

Terms in the model are described below [20];

- Vocabulary is defined as a vector  $\{1, \dots, V\}$  which consists of different items named words. Words are represented by unit-basis vectors such that  $v^{\text{th}}$  word in the vocabulary is shown as  $w^v = 1$  and  $w^u = 0$  for  $u \neq v$ .
- Each document consists of  $N$  words given as  $\mathbf{w} = \{w_1, w_2, \dots, w_N\}$  in which  $w_n$  is the  $n^{\text{th}}$  word in the document.
- Corpus  $D$  is defined by  $M$  documents, such that  $D = \{w_1, w_2, \dots, w_M\}$

We can describe the LDA model as follows: there are metagenomic samples or experiments and  $T$  topics in our collection in which  $w$  represents a k-mer. A sample  $d$  consists of  $K$  k-mers which is defined as  $d = \{w_1, w_2, \dots, w_K\}$ . A topic is a distribution over the k-mers of the samples.

Each sample in the collection is defined using the probability distribution given in the formula (4).

$$P(w_i) = \sum_{j=1}^T P(w_i|z = z_j)P(z = z_j) \quad (4)$$

Probability of a k-mer  $w_i$  in a given document is described as  $P(w_i)$ , 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  $P(w_i|z = z_j)$ .

The workflow of the LDA model applied in our study is described in Figure 2. All metagenomic samples in our collection are decomposed into different length k-mers corresponding to the words in the LDA model. Then the LDA method is applied with different number of topics and finally each sample in the collection is represented by topic distributions generated by the model. As shown in Figure 2, the LDA model is fitted in the workflow by Gibbs Sampling [21] as described in [22] where Gibbs Sampling is used for determining the posterior probability of the latent variables. Griffiths and Steyvers [22] suggest a value of  $50/k$  for  $\alpha$  and 0.1 for  $\beta$ , where  $k$  represents the number of topics. In LDA model, the numbers of topics is not known and there is no strict rule for setting the number of topics, so it is experimentally determined in our study.

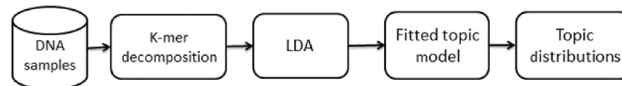


Figure 2: LDA process steps in our framework.

### 2.4 Fingerprint Comparison

Different distance metrics based on fingerprint extraction method are used to find similarity between experiments in our dataset. To find distances between metagenomic samples, we calculated the Variance-stabilized (VS) (5) and Log transformed (LT) Euclidean distances (6) between k-mer frequency vectors of the experiments. These distances are direct comparisons of frequency vectors of samples.

$$D_{\text{sqrt}}(X, Y) = \sum_K (\sqrt{f_n(k, X)} - \sqrt{f_n(k, Y)})^2 \quad (5)$$

$$D_{\text{log}}(X, Y) = \sum_K (\log(1 + f_n(k, X)) - \log(1 + f_n(k, Y)))^2 \quad (6)$$

Distance calculation is defined as follows:

Let  $X$  and  $Y$  be two fingerprint vectors of separate samples;  $f_n(k, X)$  represents frequency of  $k$ -mer  $k$  in sample  $X$  while  $f_n(k, Y)$  represents frequency of  $k$ -mer  $k$  in sample  $Y$ . For both distance metrics, the score is close to 0 for very similar experiments while it diverges from 0 when similarity decreases between them.

In addition to this, Cosine distance was used to find distance between fingerprint vectors obtained from LSA method.

$$\text{similarity} = \cos(\theta) = \frac{X \cdot Y}{\|X\| \|Y\|} = \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n X_i^2} \sqrt{\sum_{i=1}^n Y_i^2}} \quad (7)$$

Cosine score (7) ranges between 0 and 1, where 1 means that vectors are in same orientation, 0 means that they have a similarity of 0.  $X$  and  $Y$  are two fingerprint vectors of separate samples.

Kullback-Leibler (KL) divergence was used to measure the difference between two probability distributions. We used KL divergence to find differences between topic distributions of our samples after the application of the LDA method. It is a popularly used technique in statistics and pattern recognition [23].

$$D_{\text{KL}}(p(x) \| q(x)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (8)$$

Let  $p(x)$  and  $q(x)$  be two probability distribution; the KL divergence (8),  $D_{\text{KL}}(p(x) \| q(x))$ , measures the distance between them; it is a non-symmetric and non-negative measure and it is 0 if the two distributions are identical. Since it is a non-symmetric measurement, we calculated both  $D_{\text{KL}}(p(x) \| q(x))$  and the average value of  $D_{\text{KL}}(p(x) \| q(x))$  and  $D_{\text{KL}}(q(x) \| p(x))$  to observe how these two measurements affect results of finding similarities between experiments.

LDA-C implementation produces the “gamma” matrix which has number of experiments represented in rows and topics in columns. To get expected posterior probability of each topic  $\alpha$  value is subtracted from each row element and we renormalized the row to sum to one. Therefore, the distribution becomes a normal distribution that has equal mean, median and mode. In addition to this, different distributions (Poisson, Binomial, etc.) depending on the data structure can be used in finding KL divergence.

## 2.5 Data

We evaluated our retrieval system performance on a real, publicly available dataset named T2D Phase II dataset [14]. The dataset consists of 199 metagenomic samples in which there are 100 healthy people and 99 patients with type II diabetes. The dataset includes phase I and phase II data, we chose the latter one since it has higher coverage. The dataset is generated with Illumina Genome Analyzer technology and its size is about 1 terabyte. We used raw read data and applied a quality threshold of 30.

## 2.6 Evaluation Criteria

Actual relevance between compared experiments-referred to as “ground truth” – is used to evaluate our retrieval system. Entities retrieved by the system are considered relevant or not based on this ground-truth. Retrieving patients with the same disease is the ground truth for relevance: two samples are considered relevant if they are from the same class (disease class). In the dataset, a positive sample represents a patient with a disease; a negative sample represents a healthy person. For a given positive query sample, the system should retrieve relevant samples, that is patient with the same disease as the query. This case is labeled as relevant, while all other cases are labeled as irrelevant retrieving.

Mean average precision (MAP) is used in order to evaluate retrieval performance of the system. MAP is a widely used technique in information retrieval for measuring retrieval performance. For query  $q$ , system generates a ranked list of retrieved samples in ascending order with regard to similarity scores of them. Most similar samples with the query are listed on the top, others are listed below. Precision is calculated as follows with using top  $n$  samples;  $n \in \{1, 2, \dots, N\}$ ;

$$\text{Precision}(n; q) = \frac{\text{number of samples relevant to } q \text{ in } n \text{ retrieved samples}}{n} \quad (9)$$

$$\text{MAP} = \frac{1}{|M|} \sum_{q \in M} \text{AveP}(q) \quad (10)$$

$$\text{AveP}(q) = \frac{1}{k_q} \sum_{n \in L_q} \text{Precision}(n; q) \quad (11)$$

In order to establish that the LDA model can generate accurate topic distributions between k-mers, we made multiple sequence alignment to find similarities of sequences grouped in the same topic. In our model, user-defined number of topics are generated and k-mers are assigned with a probability distribution to the one of the topics. It is expected that similar sequences in the same topic have some biological similarities. Sequence alignment approaches are powerful techniques for many biological problems such as detecting the similarities between sequences. We used a fast and scalable multiple sequence alignment method from the study of Sievers et al. [24]. The method basically uses insertion of gaps in the sequences to maximize the overall similarity. To show the degree of matching a similarity measure called the percent identity matrix is generated by the method [25].

## 2.7 Implementation

The framework is implemented using C++, R and MATLAB and tested on Windows platform. We used Boost 1.64 and zlib as external libraries. Executable files, documentation and test data files can be downloaded from the link: [www.baskent.edu.tr/~hogul/WMS\\_retrieval.rar](http://www.baskent.edu.tr/~hogul/WMS_retrieval.rar). The user can run the framework with user specific parameters and datasets using provided executable files.

## 3 Results

We tested the proposed retrieval system on a real dataset which consists of 199 metagenomic samples, 99 of which are patients with type II diabetes so-called positive samples, while the others are healthy ones. The proposed system is expected to retrieve samples that are relevant w.r.t. to the definition above, i.e. positive samples.

We computed k-mer frequency values of samples for each  $k$  ranging from 2 to 13. After representing each sample with a frequency vector, similarity scores of experiments were calculated and comparison of similarity metrics was performed for related  $k$  value. MAP scores related to k-mer frequencies, ranging from 2 to 13, computed by LT and VS Euclidean distances are shown in Figure 3. Result values highlight that LT and VS distances have comparable performances. As expected it can be observed that MAP scores increase at increasing values of  $k$ , since higher  $k$  provides more information.

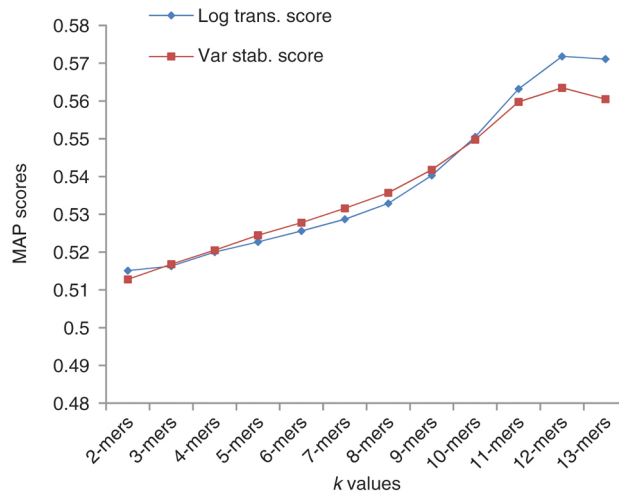
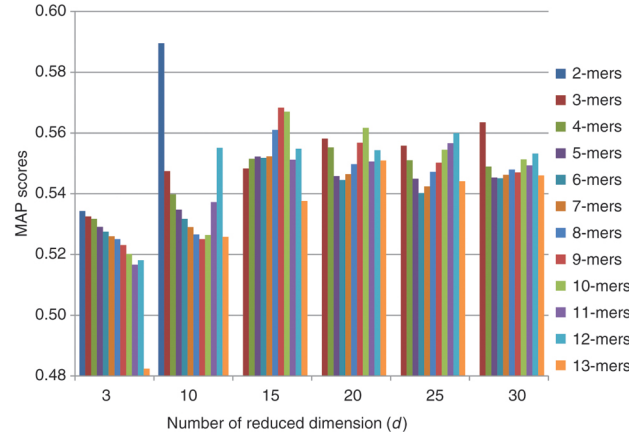


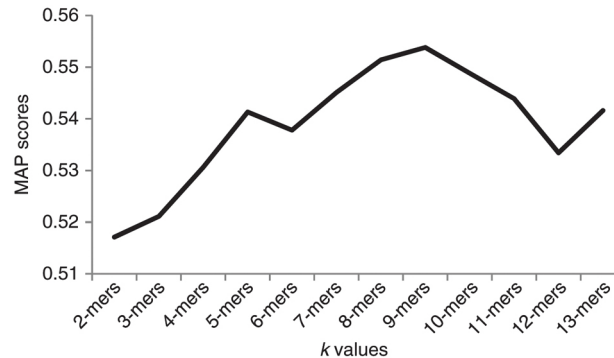
Figure 3: MAP scores of the Log transformed and Variance-stabilized Euclidean distances.

Results related to LSA fingerprint extraction method are shown in Figure 4. We simply computed MAP scores for several values of  $d$ , the LSA feature dimension parameter. For 2-mers, scores could not be calculated for values of  $d = 20$ ,  $d = 25$ , and  $d = 30$ , since 2-mer vector size is equal to 10 considering complementary reverse k-mer couples as single k-mer. In addition to this, we applied feature selection methods to reduce dimension of k-mer frequency vectors for  $k$  values greater than 11, because of exponential increase of vector dimensions. According to the results, LSA generally performs well with the parameter  $d = 15$  for each  $k$  value.



**Figure 4:** MAP scores of LSA fingerprint extraction method considering different  $d$  values.

The second fingerprint extraction method, called LDA, has some parameters to be defined such as number of topics ( $k$ ), alpha ( $\alpha$ ) and iteration number (iter) and there is no strict rule for setting those parameters. In order to observe how LDA model parameters affect retrieval performance, we applied the method with different parameters that were set experimentally for each  $k$  value. For each k-mer, best MAP scores, which are greater than direct comparison scores of the same k-mer, are taken as retrieval performance. Furthermore, when the total number of k-mers increases exponentially with increasing  $k$  values, a feature selection method is needed to lower the computational cost. For  $k$  values which are greater than 7, the feature selection methods CAE and TFIDF were thus used to reduce the dimension of k-mer frequency vectors of each experiment. According to the Figure 5, LDA has increasing retrieval performance with  $k$  between 2 and 9, but it does not have the same performance for  $k$  values equal to or greater than 10. This can be explained by the fact that feature selection methods failed to select informative k-mers.



**Figure 5:** MAP scores of LDA fingerprint extraction method for different  $k$  values.

In the LDA model, each k-mer is assigned to a topic generated by the system with a probability distribution. To assess the accuracy of this assignment, we used topic-level distributions of 8-mers. Table 1 represents the 8-mer lists for five generated topics. It is expected that k-mers in same topic should have some common functional roles from a biological point of view. In order to verify this, we applied multiple sequence alignment for the sequences of topic 1 given in Table 1. Percent identity matrix is generated to show the degree of matching between sequences. To visualize evolutionary relationships or pathway among the sequences, a phylogenetic tree based on evolutionary distances between sequences is built. According to the tree, given in Figure 6, there are some sequences such as seq7, seq5, seq1 and seq2 which have high similarity with other sequences which means that the model has achieved in grouping similar sequences in same latent topic. As shown from the figure, seq1 and seq2 are the closest sequences in the set sharing the same branch with seq5 and seq7.

**Table 1:** Top ten ranked 8-mers for latent topics generated by the LDA model.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Seq1	TCGGAAGA	AAAAAGAA	ATGAAGAA	ATGAAAAA	AAAAGAAA
Seq2	CGGAAGAG	AAAAAGGA	GATGCTGA	AAAAGAAG	CAATGGCA
Seq3	CCGATCTC	AATTTTTC	AAGAAGAA	TATCCGGA	CATCATCA
Seq4	AAAAGAAA	AAAAGAAA	AAAAGAAA	CGGAAGAA	GGCATCAA
Seq5	ATCGGAAG	GAAAAAGA	GATGGCAA	AAAGAAGA	CATTGCCA
Seq6	AGAAAGAA	AGAAAAAG	CATCATCG	AAGAAGAA	AAAAATAA
Seq7	GATCGGAA	TATGAAAA	GGCGATGA	CTTTTTC	ATGCCATA
Seq8	GAAAGAAA	CTTTTTC	TATCATCA	ATGAAAAA	ACAAGCAA
Seq9	GAAGGAAA	AAGAAAAA	GATGATGC	CCGAAAAA	CATCGACA
Seq10	AGAAGAAA	ATGAAAAA	AGGAAGAA	TGGATGAA	AACAAAAA

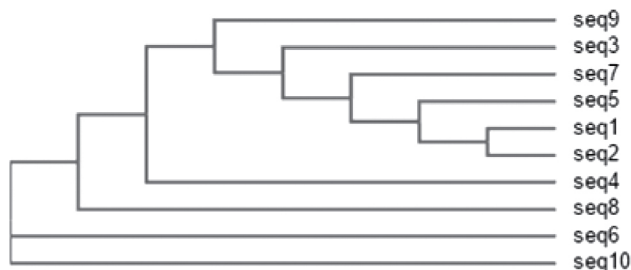


Figure 6: Phylogenetic tree of sequences in the topic 1.

In our study, we used two different fingerprint extraction methods called LSA and LDA. We compared the results of these methods with the reference method called direct comparison. Comparison process was performed through MAP scores of each method and related similarity metric. Figure 7 gives the comparative results of LSA and LDA methods with direct comparison using Log transformed and Variance-stabilized Euclidean distances. According to the figure, LSA method has been successful in retrieving relevant experiments for  $k$  values less than or equal to 10. LDA method performance is close to LSA for  $k$  values between 5 and 8. For higher  $k$  values, the direct comparison method has better retrieval performance than fingerprinting methods. Direct comparison of feature vectors with huge dimension depends only on time and space factors, but representing these vectors by any of fingerprint extraction method also depends on an appropriate feature selection algorithm. Thus, it leads us to an additional challenge that is out of the scope of this study. The results obtained show that representing metagenomic samples with small  $k$ -mer frequency vectors and transforming the vectors by LSA and LDA methods can help to find similarities between samples with a good accuracy.

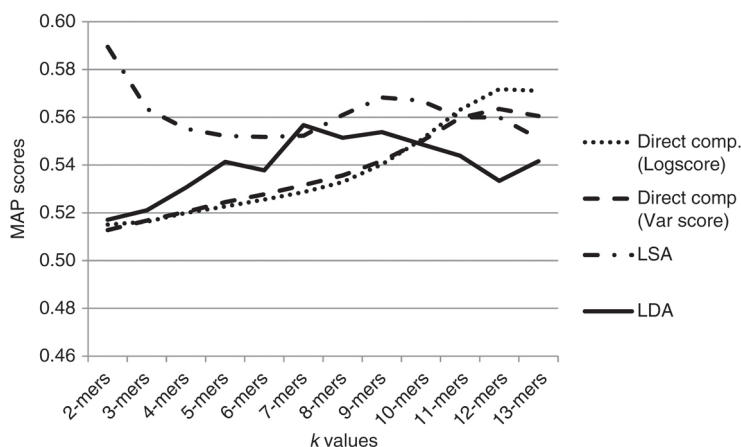


Figure 7: Comparative results of LSA and LDA fingerprint extraction methods with direct comparison by using Log trans. score and Var. stab score.

In order to assess whether differences in performances of fingerprint extraction and direct comparison methods were statistically significant two statistical tests were performed, non-parametric Wilcoxon Signed Rank and Paired  $t$ -test; the related  $p$ -values were computed. We obtained  $p$ -value of  $8.99E-07$  and  $2.39E-07$  between LSA and direct comparison with using Wilcoxon Signed Rank and Paired  $t$ -test, respectively. For comparison of LSA and LDA, we obtained a  $p$ -value of  $1.26E-04$  and  $1.02E-04$ . All  $p$  values are smaller than 0.05 which represents strong evidence that LSA outperforms the other methods in retrieving relevant experiments using

small  $k$  values. We also performed the same tests for observing the difference between retrieval by fingerprinting approach and retrieval by random. We get the  $p$ -value of  $8.83E-07$  and  $2.07E-08$  for Wilcoxon and paired  $t$ -test, respectively. We can conclude that using fingerprinting approach in terms of MAP scores is statistically significant.

## 4 Conclusion

Along with the rapid growth of sequencing technologies, the analysis and interpretation of metagenomic data have become a major challenge in bioinformatics. Finding similarities and differences between metagenomic samples within large repositories has been rather a significant issue for researchers. In this study, a reference-free framework based on novel fingerprinting approaches is proposed. The framework consists of feature extraction and selection and experiment comparison. Performance of the developed framework was evaluated by using given metagenomic experiments, called T2D.

Encoding metagenome sequencing experiments into their fingerprints and comparing the fingerprints to detect similarities among them is our main goal. To extract fingerprints, we applied and proposed the application of algorithms that are a novelty in this field. We used two different fingerprint extraction methods LSA and LDA and direct comparison of  $k$ -mer frequency vectors. It is observed that LSA is a promising fingerprinting approach for representing metagenomic samples to find relevance between them. LSA method outperforms the LDA and direct comparison method with short  $k$ -mers ( $k \leq 10$ ), but there is a smooth decrease in retrieval performance when  $k$  value increases. For greater  $k$ -mers ( $k > 10$ ), it can be stated that direct comparison becomes more successful in retrieving relevant experiments. The choice of  $k$  highly impacts on precision and efficiency of results, moreover, for the large value of  $k$  direct comparison is not reliable so that an effective feature selection step is needed.

There is only one study (Seth et al. [10]) that can be compared with our study in respect to achievement in finding relevant experiments and we used the same dataset of the study of Seth et al. [10]. To the best of our knowledge, there is no other study that uses this dataset for finding similarities between experiments. Seth et al. get big  $k$  values (30-mers) with a distributed system, while we get the maximum value of 13 for  $k$ . Regarding 12-mers, we obtained a lower score than the score of Seth et al. with a direct comparison using Log transformed Euclidean distance, but it was achieved without any selection process.

The study in its current form is a proof of concept to show the adaptability of text mining approaches as fingerprinting approaches for representing experiment content in a feature space. The results show that  $k$ -mer frequency representation by these methods provides a suitable and promising approach to efficiently compare metagenomic samples. Moreover, there are two biological contributions to the work. Having assumed that when two metagenomic samples are relevant, e.g. come from patients with the same disease, they also share similar experiment content. This assumption is verified by the obtained results using different fingerprinting approaches and similarity metrics. The second contribution provided by the LDA method is that sequences in the same group have evolutionary relationships such as having similar biological functions or sharing a common ancestor. These contributions encourage the researchers to implement new techniques for metagenomic data analysis, especially for whole database search.

Finally, retrieval of metagenomics samples with large  $k$ -mers leads us to new motivations in the field of bioinformatics. This motivation will be our future work.

## Acknowledgements

This study was funded by a bilateral project CNR-TUBITAK, Italy-Turkey 2016-2017 (Retrieval of whole-metagenome sequencing samples) and by the Scientific and Technological Research Council of Turkey (TUBITAK) under the Project 215E369.

**Conflict of interest statement:** Authors state no conflict of interest. All authors have read the journal's publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

## References

- [1] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. 2008;9:386.
- [2] Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol*. 2005;1:e24.
- [3] Jiang B, Song K, Ren J, Deng M, Sun F, Zhang X. Comparison of metagenomic samples using sequence signatures. *BMC Genomics*. 2012;13:730.
- [4] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res*. 2007;17:377–86.
- [5] Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*. 2007;73:5261–7.
- [6] Liu Z, Hsiao W, Cantarel BL, Drábek EF, Fraser-Liggett C. Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics*. 2011;27:3242–9.
- [7] Su X, Xu J, Ning K. Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. *Bioinformatics*. 2012;28:2493–501.
- [8] Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev, DG. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*. 2016;17:1.
- [9] Maillot N, Lemaître C, Chikhi R, Lavenier D, Peterlongo P. Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics*. 2012;13(Suppl. 19):S10.
- [10] Seth S, Välimäki N, Kaski S, Honkela A. Exploration and retrieval of whole-metagenome sequencing samples. *Bioinformatics*. 2014;30:2471–9.
- [11] White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*. 2009;5:e1000352.
- [12] Parks DH, Beiko RC. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*. 2010;26:715–21.
- [13] Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. *Genome Biol*. 2011;12:R60.
- [14] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*. 2012;490:55–60.
- [15] Weitschek E, Santoni D, Fison G, De Cola MC, Bertolazzi P, Felici G. Next generation sequencing reads comparison with an alignment-free distance. *BMC Res Notes*. 2014;7:869.
- [16] Weitschek E, Cunial F, Felici G. Classifying bacterial genomes with compact logic formulas on k-Mer frequencies. In: 25th International Workshop on Database and Expert Systems Applications (DEXA). IEEE; 2014, p. 69–73.
- [17] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manage*. 1988;24:513–23.
- [18] Dumais ST, Furnas GW, Landauer TK, Deenvester S. Using latent semantic analysis to improve information retrieval. In: Proceedings of CHI'88 Conference on Human Factors in Computing Systems. 1988; p. 281–85.
- [19] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J Assoc Inf Sci Technol*. 1990;41:391407.
- [20] Blei DM, Andrew Y, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993–1022.
- [21] Casella G, George EI. Explaining the Gibbs sampler. *Am Stat*. 1992;46:167–74.
- [22] Griffiths TL, Steyvers M. Finding scientific topics. *Proc Natl Acad Sci*. 2004;101(Suppl. 1):5228–35.
- [23] Joyce JM. Kullback-Leibler divergence. In: International Encyclopedia of Statistical Science. Berlin, Heidelberg: Springer; 2011, p. 720–2.
- [24] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
- [25] Petsko GA, Ringe D. Chapter 4: From Sequence to Function. Protein structure and function. United Kingdom: New Science Press; 2004.