Informatics in Medicine Unlocked 13 (2018) 151–157

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Taxonomy metagenomic analysis for microbial sequences in three domains system via machine learning approaches

Heba M. Afify[a,*], Mohammed A. Al-Masni[b]

[a] Department of Bioelectronics Engineering, MTI University, Cairo, Egypt
[b] Department of Biomedical Engineering, Kyung Hee University, Yongin, South Korea

ARTICLE INFO

Keywords:
Bioinformatics
Genome sequences
Supper vector machine (SVM)
Deep belief network (DBN)
Taxonomy

ABSTRACT

The rapid advancements of using clinical microbiology and genome sequences encourage several taxonomic approaches, based upon both genome classification and bioinformatics surveys. This taxonomy arranges the tree of life among different organism databases and exploits the high similarly in biological information to access the best representation of genome sequences. However, there is still a challenge to find the entire hierarchy of this tree, due to the existence of the biodiversity databases, and the different classifiers, according to evolutionary or phylogenetic relationships. This paper presents the classification of three domains of microorganisms including Bacteria, Archaea, and Eukarya using two algorithms: the supper vector machine (SVM) and the deep belief network (DBN). The proposed approach utilized the alignment method and the code generation process as preprocessing steps on the EzBioCloud 16S rRNA database. In addition, this study accommodated the issue of choosing the proper reference sequence (RefSeq) and the appropriate code generation process of the genome sequences. Our results showed that the proposed method classifies the genome sequences with an overall classification accuracy of 99.99% and 99.93% for SVM and DBN classifiers using the standard RefSeq of each class, respectively. This paper enhanced the area of microbiological scientific classification through progress in using the character-based arrangement that will help in future evolutionary frameworks.

1. Introduction

In past years, Integrated Taxonomic Information System (ITIS) was an official tool for arranging different types of living organisms in order to establish the index of complete global species called the Catalogue of Life (COL) which is managed by the Federal Government of the United States [1]. There are many taxonomists who are dependent upon COL to distinguish between over a million species that are supported in the classification process. The first format of the whole genome is presented in 1995 [2]. The continually increasing genomes led to enormous databases with an emphasis on the demand of classification accuracy. The first bacterial and archaeal genomes were formed in 1995 and 1996, respectively [3]. Therefore, three domains of life were divided into Bacteria, Archaea, and Eukarya, which was utilized to construct the Integrated Microbial Genomes (IMG) database [4] to record all microbial information by full description of genes, genomes, and functions of each class in the database. The structural comparison of various microbial genomes of closely related species revealed a significant feature in the analysis of homologous genomes as introduced by the Comprehensive Microbial Resource (CMR) [5] and Microbial Genome Database (MBGD) [6].

In the tree of life, it was found that there were two kingdoms including the Prokaryota and Eukaryota which were based on the variations in 16S rRNA genes [7]. The Prokaryote kingdom consists of both Bacteria and Archaea due to the fact that this family has no nucleus, while the Archaea is more related to the Eukaryotes than the Bacteria. Recently, the 16S rRNA gene has been applied for correcting estimation of the phylogenetic relationships between microorganisms, especially for all types of bacteria [8]. The advantage of the 16S rRNA gene is focused on the large sequence (i.e., about 1500-bp length) that carried more genomic information. Thus, it used to achieve the taxonomy by useful characteristics and universal primers of the preserved parts [9]. Although the 16S rRNA gene has high quality, there are some constraints on handling the 16S rRNA genes, such as the technical and financial aspects as well as the deficiency of bioinformatics software tools for comparative analysis and sequences predication. Generally, the right category of organisms treated with rapid identification is the main objective for building the phylogenetic tree [10] and for collecting the pathogenic genes in one group under a single denomination [8]. Over the last few years, the combination of microbial laboratories and

* Corresponding author.
E-mail addresses: hebaaffify@yahoo.com (H.M. Afify), m.almasani@khu.ac.kr (M.A. Al-Masni).

https://doi.org/10.1016/j.imu.2018.05.004
Received 9 April 2018; Received in revised form 10 May 2018; Accepted 16 May 2018
Available online 17 May 2018

next-generation sequencing (NGS) technology has created the term of metagenomics, which guarantees the widespread analysis of novel bioinformatics tools [11].

Several attempts for classification of the metagenomics data are suffered from the lack of a good classifier. Wang et al. [12] applied the naïve Bayesian algorithm to classify bacterial 16S rRNA sequences. The classification accuracy was improved in cases of low similarity rate within existing sequences. Wu et al. [13] explained AMPHORA2 as an analysis tool for metagenomic databases containing bacterial and archaeal sequences, based upon increasing the time of the phylotyping process using a sequence alignment masking algorithm. The Integrated Microbial Genomes with Microbiome Samples (IMG/M) is another platform for metagenome datasets [14]. It is used to provide effective projects by the realization of new evolutionary relationships between genes, defining the boundaries among microbial diversity, and detection of alternative genes.

Moving forward, the metagenomic bioinformatics are continuously developed by the Efficient Database framework for comparative Genome Analyses using BLAST score Ratios (EDGAR) [15] which uses limited features to determine the relationships between old and new genomes for updating the phylogeny. It was based on identifying the similarity and difference rates in sequenced microbial genomes. The redevelopment of this software is designed by EDGAR 2.0 [16] which used high-level features and statistical analyses to avoid the phylogenetic mistakes among microbial organisms. On the other hand, OrthologeDB [17] was implemented for Bacteria and Archaea databases that is based upon statistical algorithms, ortholog features, and separation of any inadequate phylogenetic factors.

Yu et al. [18] discussed several bioinformatics programs for differentiation between phylogenetic approaches and ranking of microbial organisms. Recently, EzBiocloud [19] is represented as a comprehensive view of metagenomics taxonomy and storage genomic attributes through analytical and visualization features, with respect to 16S rRNA and completely sequenced genomes. The sequences with a high degree of similarity and phylogenetic context were considered as a strength of the EzBiocloud database for metagenomic analysis. Thus the EzBiocloud database was addressed as a genomic-related classification problem. The 16S rDNA sequence serves as a useful tool in the classification of the metagenomic database, and achieved the good accuracy for human microbiota sequences [20]. Marsh et al. developed a classification model of microbial metagenomic datasets using a sequence clustering method [21]. Generally, studies of the microbial metagenomic database were required to assist the bioinformatics sector in computer-based classification and computer-aided diagnosis of different sequences. The bioinformatics analysis of Bacterial pathogens is also used for developing the host cell RNA-sequencing experiments that were used in laboratory services only [22]. Therefore, bioinformatics analysis of Bacterial pathogens is performed for diagnostic and therapeutic goals.

In this paper, we developed a metagenomic classification model to distinguish between Archaea, Bacteria, and Eukarya sequences using the aligned sequences of the EzBioCloud database. All the sequences were converted into numerical values using code generation based genomic characters and a variable reference sequence (RefSeq). This study was evaluated by two classifiers that encompassed traditional machine learning such as the supper vector machine (SVM) [23] and a deep learning model such as the deep belief network (DBN) [24]. The contribution of our work is based on microbiological classification under the MATLAB platform version R2015a using the EzBioCloud database.

2. Materials and methods

In this paper, the proposed approach is divided into three steps: sequences alignment, code generation, and taxonomic analysis as shown in Fig. 1. First, the sequence alignment method [25] is used to generate many more similarities among the sequences by providing some gaps (−) according to similarity, replacement, insertion, and deletion operations of characters in the genomic sequences. Secondly, the code generation process is applied on the aligned sequences by converting the aligned characters to numerical values. The choice of RefSeq for the classification process is a critical problem. Therefore, we investigated two approaches of selecting the standard RefSeq. The first approach selects the three standard RefSeq of Archaea, Bacteria, and Eukarya classes in which each RefSeq is corresponding to its class. However, in the second approach, only one RefSeq from any class is selected for all classes. Finally, two classifiers are applied to the output sequences of the code generation step.

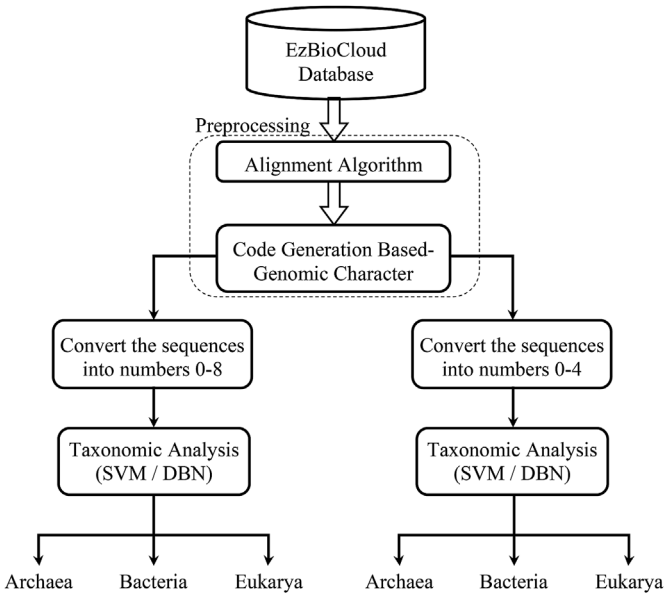

Fig. 1. Overall workflow for classification of Archaea, Bacteria, and Eukarya sequences in EzBioCloud database.

2.1. Database preparation

We utilized the latest well-known EzBioCloud 16S rRNA database [19] to classify the Archaea, Bacteria, and Eukarya sequences. The EzBioCloud contains two formats, QIIME and MOTHUR pipelines. In this study, we used the MOTHUR pipeline since the sequences exist in aligned form.

The EzBioCloud database contains 63,240 sequences categorized into three classes which are Archaea, Bacteria, and Eukarya. In order to avoid the over-fitting of the recognizer networks, we have divided the data into three independent datasets; training, validation and test datasets. The training dataset is used to learn the deep learning model, while the validation dataset is utilized to evaluate and update the parameters of the model. Furthermore, the test dataset is used to evaluate the final model performance [26,27]. In this work, the training dataset contains 60% of all data, while the validation and test datasets consist of 10% and 30%, respectively. Table 1 summarizes the distribution of these datasets that are based on three classes.

Table 1
Distribution of the training, validation, and test datasets for Archaea, Bacteria, and Eukarya sequences.

Dataset	Percentage (%)	Archaea	Bacteria	Eukarya
Training	60	1665	35,495	784
Validation	10	277	5915	130
Test	30	832	17,747	392
Total	**100**	**2774**	**59,157**	**1306**

2.2. Alignment algorithm

The pairwise of DNA sequence alignment is achieved by using the global alignment algorithm [28] in which it performed with a similarity of 98.7% for bacterial taxonomy in EzTaxon of the EzBioCloud database. The alignment method is performed by adding gaps (−) into sequences in order to generate high similarities among the homologous sequences. The similarity is computed as follows,

$$\text{Similarity}\,(\%) = 100 \times \frac{Match}{Match + Mismatch}, \quad (1)$$

where *Match* represents the amount of characters in the sequences that are corresponding to each other. The alignment method is required as a preprocessing step, and it occurs among all sequences for each class to obtain the same length in all sequences.

2.3. Code generation

The code generation step is used for a numerical transaction that is based on conversion of the aligned sequences into a coded form. The choice of code translation plays an important role in the performance of classification results, which studies the relationship between a target sequence and its RefSeq in corresponding bases. We suggested two types of code generation to construct our model. In the first type, the aligned sequences are easily encoded as A = 1, C = 2, G = 3, T = 4, '-' = 0 which indicated to five bits encoding (0–4 bits). The second type is based on nine bits encoding (0–8 bits) which related to four operations of aligned sequences [29] as shown in Table 2. The list of op-codes vectors consists of a numerical form according to our assumptions. Then, we obtained the op-codes vectors from aligned sequences that represented the input element of the proposed classifiers.

According to EzBioCloud database, Fig. 2 elucidated the samples of datasets after code generation (0–8) of Archaea, Bacteria, and Eukarya sequences.

2.4. Taxonomic analysis

In this paper, the taxonomic analysis is used for classification and microbial recognition of the encoded sequences from the code generation step. We have separately applied two classifiers to the microbial sequences, which are SVM as a classifier of the traditional machine learning and DBN as a deep learning classifier.

We have selected SVM due to advances of the classifier stability and its high performance in bioinformatics challenges, especially for genomic homology-based classification [30]. SVM is considered as a supervised learning model in which the SVM model is built based on a given labeled training dataset [31,32]. In fact, SVM is based on the principle of the decision planes (i.e., hyperplanes) in which each plane enables to separate two classes according to their features distribution. Multiclass classification with SVM is achieved by a common technique called one-versus-rest or one-versus-all. During the training phase, SVM tried to find the proper hyperplane by minimizing the empirical error and maximizing the margins. The largest margin between classes implied better a hyperplane of the SVM. In this study, we have used a linear kernel function with the sequential minimal optimization (SMO) method to find the separating hyperplane.

Recently, the Deep belief network (DBN) classifier is widely used in bioinformatics fields such as splice junction prediction [33], protein expression [34] and it is also conducted in studies for discriminating between different genomic classes. Moreover, we also used the deep learning model of DBN to evaluate the microbial recognition and compare its results with the SVM algorithm. Generally, DBN is a composition of the unsupervised restricted Boltzmann machine (RBM) which is considered as a generative stochastic network with only connections between the visible and hidden nodes [35,36]. After the unsupervised propagation of RBM is completed, the back-propagation with supervised learning is achieved to fine-tune the parameters of the network. Thus, DBN has the capability to extract the prominent attributes form the input sequences. As shown in Fig. 3, our DBN architecture consists of an input layer with a visible node of m = 7682 which represented the input sequence features. The Hidden layers involved four layers with a different number of nodes of n = 1,000, o = 400, p = 15, and q = 8. The output layer with three nodes represented a number of classes (i.e., Archaea, Bacteria, and Eukarya).

Table 2
Operation of code generation of aligned sequences (0–8 bits).

Corresponding Base	Operation	Op-codes
The same	Similarity	“0”
A ← → T	Replacement	“1”
G ← → C		
A ← → G		“2”
C ← → T		
A ← → C		“3”
G ← → T		
A or T or G or C → ‘-’	Deletion	“4”
‘-’ → G	Insertion	“5”
‘-’ → A		“6”
‘-’ → C		“7”
‘-’ → T		“8”

3. Results

The EzBioCloud dataset has initially partitioned to 60% for training set, 10% for validation set and 30% for test set in order to estimate the fulfillment of the proposed approach. The proposed study is based on taxonomic analysis for all three classes which are Archaea, Bacteria, and Eukarya sequences. The SVM and DBN classifiers are utilized to improve the phylogenetic domains of metagenomic data. We have implemented two types of RefSeq and code generation to select the suitable process for this database. In the first RefSeq type selection, we used the three standard Cambridge references of the three classes (i.e., one sequence from each class as a RefSeq for all sequences in same class). While in the second RefSeq type, we selected only one RefSeq from any class for all sequences in all classes. For the code generation step, we applied the conversion of the aligned sequences into a numerical vector of five bits encoding (0–4 bits) or nine bits encoding (0–8 bits). The performance of this work is evaluated using the measures of a confusion matrix and the overall classification accuracy.

In terms of using the standard Cambridge references, results showed that the code generation (0–8) for both SVM and DBN classifiers provides better classification performance than utilizing the code generation (0–4) as shown in Tables 3 and 4.

The results show that code generation (0–8) is a good choice to strongly analyze the EzBioCloud database and distinguish the differences between the three classes. The SVM and DBN classifiers achieved overall accuracies of 99.99% and 99.93%, respectively. The results interpreted that both SVM and DBN classifiers yielded high classification accuracies in all three classes. Additionally, it is noted that the accuracy of Bacteria sequences in this database has the best recording in classification process, which is allowed to offer a reliable description of the Bacteria sequences based on their high similarity. This is due to fact that the number of Bacteria sequences in the database is much higher than the other classes.

In terms of using the encoding process (0–8 bits), Tables 5–7 fully addressed the modification of the RefSeqs regarding the selection of only one sequence as a reference for all classes. In the case of one sequence of Bacteria being set as a reference for all classes, the accuracy of Archaea sequences for SVM and DBN classifiers provided the highest accuracy of 100% when compared to other sequences as shown in Table 5. Results show that the performance of Archaea sequences are close to the RefSeq from Bacteria sequence, while the performance of

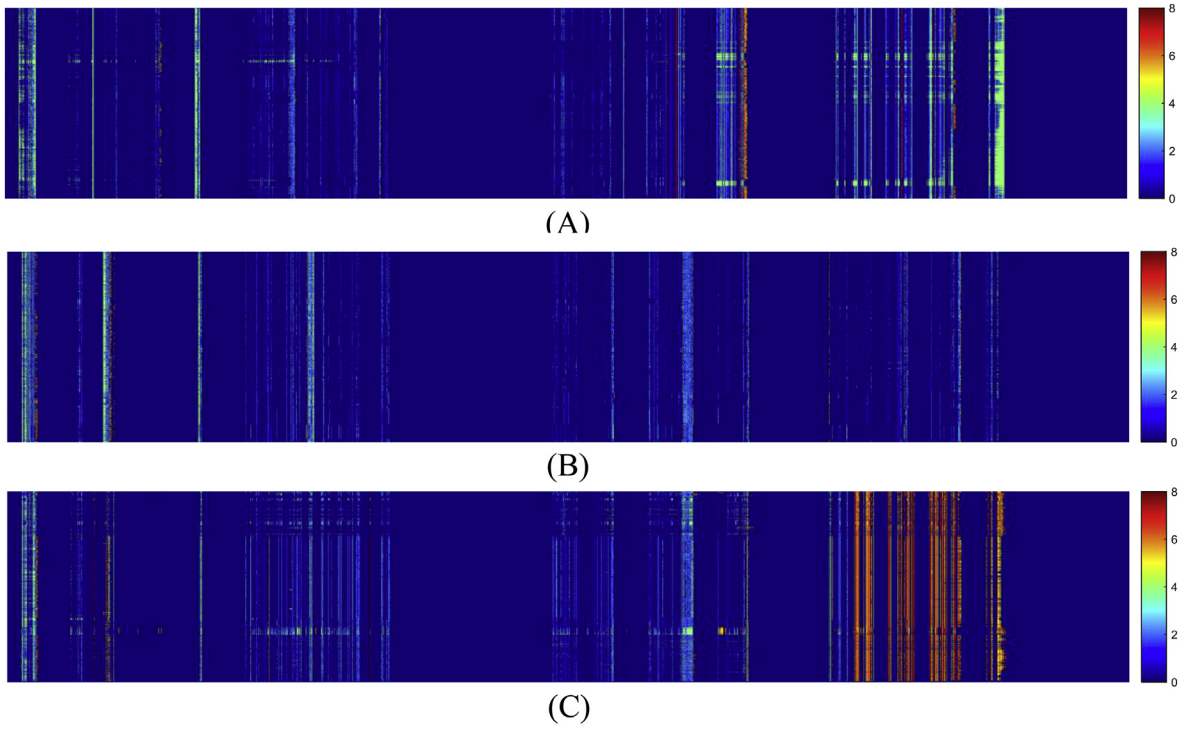

Fig. 2. Samples of code generation (0–8) of three sequences (a) Archaea, (b) Bacteria, and (c) Eukarya.

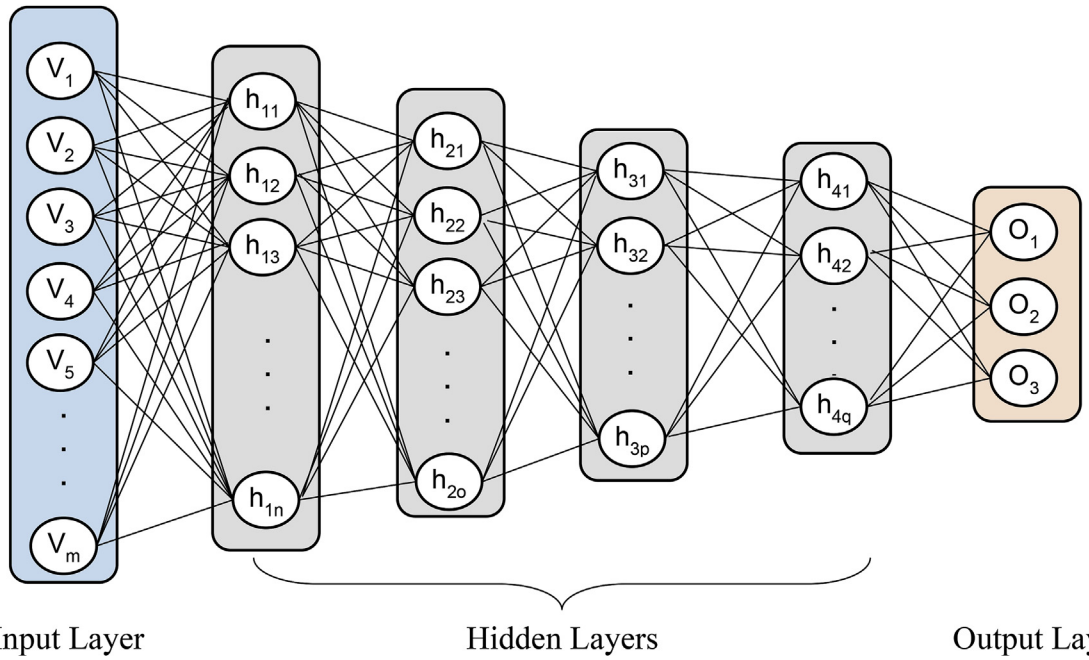

Fig. 3. The DBN architecture consists of a visible input layer with m nodes, four hidden layers with n, o, p, and q nodes, respectively, and an output layer with three nodes which represent the number of classes.

Eukarya sequences is less than that of the Bacteria sequence. The reference from Bacteria sequences is more adapted to the classification of Archaea sequences and the reference from Archaea sequences is also more adapted to the classification of Bacteria sequences by using both classifiers. Thereby, the results confirmed the concept of a phylogenetic tree that indicated substantial relationships between Archaea sequences and Bacteria sequences but slight relationships between Eukarya sequences and Bacteria sequences, shown in Tables 5–7 On the other hand, Table 7 illustrated the results in terms of experimental accuracies for using the SVM classifier in case of selection of the Eukarya sequence as a reference for all classes.

As a summary, the best accuracy is achieved by the SVM classifier with code generation (0–8) and standard RefSeq for each class in which it has slightly improvement compared to the DBN classifier as shown in Table 3. In terms of training time computation, the SVM classifier required less time for the classification process based on code generation (0–8) and standard RefSeq compared to DBN classifier as shown in Table 8.

Table 3
Confusion matrix and overall accuracies of the proposed SVM method using the standard RefSeq for each class.

With Code Generation (0–4)				With Code Generation (0–8)			
Confusion Matrix				Confusion Matrix			
	Archaea	Bacteria	Eukarya		Archaea	Bacteria	Eukarya
Archaea	831	0	1	Archaea	831	1	0
	99.88%	0%	0.12%		99.88%	0.12%	0%
Bacteria	0	17,744	3	Bacteria	0	17,747	0
	0%	99.98%	0.02%		0%	100%	0%
Eukarya	29	4	359	Eukarya	0	1	391
	7.40%	1.02%	91.58%		0%	0.26%	99.74%
Classification Accuracy (%)				**Classification Accuracy (%)**			
Accuracy of Archaea			99.88	Accuracy of Archaea			99.88
Accuracy of Bacteria			99.98	Accuracy of Bacteria			100
Accuracy of Eukarya			91.58	Accuracy of Eukarya			99.74
Overall Accuracy			**99.80**	**Overall Accuracy**			**99.99**

Table 4
Confusion matrix and overall accuracies of the proposed DBN method using the standard RefSeq for each class.

With Code Generation (0–4)				With Code Generation (0–8)			
Confusion Matrix				Confusion Matrix			
	Archaea	Bacteria	Eukarya		Archaea	Bacteria	Eukarya
Archaea	0	832	0	Archaea	831	1	0
	0%	100%	0%		99.88%	0.12%	0%
Bacteria	0	17,747	0	Bacteria	0	17,747	0
	0%	100%	0%		0%	100%	0%
Eukarya	0	392	0	Eukarya	2	10	380
	0%	100%	0%		0.51%	2.55%	96.94%
Classification Accuracy (%)				**Classification Accuracy (%)**			
Accuracy of Archaea			0	Accuracy of Archaea			99.88
Accuracy of Bacteria			100	Accuracy of Bacteria			100
Accuracy of Eukarya			0	Accuracy of Eukarya			96.94
Overall Accuracy			**93.54**	**Overall Accuracy**			**99.93**

Table 5
Confusion matrix and overall accuracies of the proposed SVM and DBN methods using code generation (0–8) and one RefSeq of Bacteria for all classes.

SVM Classifier				DBN Classifier			
With Code Generation (0–8)				With Code Generation (0–8)			
Confusion Matrix				Confusion Matrix			
	Archaea	Bacteria	Eukarya		Archaea	Bacteria	Eukarya
Archaea	832	0	0	Archaea	832	0	0
	100%	0%	0%		100%	0%	0%
Bacteria	0	17,744	3	Bacteria	0	17,744	3
	0%	99.98%	0.017%		0%	99.98%	0.017%
Eukarya	24	4	364	Eukarya	51	9	332
	6.12%	1.02%	92.86%		13.00%	2.31%	84.69%
Classification Accuracy (%)				**Classification Accuracy (%)**			
Accuracy of Archaea			100	Accuracy of Archaea			100
Accuracy of Bacteria			99.98	Accuracy of Bacteria			99.98
Accuracy of Eukarya			92.86	Accuracy of Eukarya			84.69
Overall Accuracy			**99.84**	**Overall Accuracy**			**99.67**

4. Discussion

The accurate classification of the clinical microbiology is an extremely complex and challenging process, due to the growing rate of genomic sequencing in the metagenomic database [37]. The diversity of genomic structures between microbial sequences is useful to recognize the different diseases and create groups of similar sequences. Therefore, it is imperative to appropriate and reorder the microbial groupings, keeping in mind the similarity between genomic sequences. It is realized that machine learning strategies are utilized for programmed

Table 6
Confusion matrix and overall accuracies of the proposed SVM and DBN methods using code generation (0–8) and one RefSeq of Archaea for all classes.

SVM Classifier				DBN Classifier			
With Code Generation (0–8)				With Code Generation (0–8)			
Confusion Matrix				Confusion Matrix			
	Archaea	Bacteria	Eukarya		Archaea	Bacteria	Eukarya
Archaea	832	0	0	Archaea	832	0	0
	100%	0%	0%		100%	0%	0%
Bacteria	0	17,743	4	Bacteria	0	17,747	0
	0%	99.98%	0.023%		0%	100%	0%
Eukarya	25	5	362	Eukarya	49	343	0
	6.38%	1.28%	92.35%		12.5%	87.5%	0%
Classification Accuracy (%)				**Classification Accuracy (%)**			
Accuracy of Archaea			100	Accuracy of Archaea			100
Accuracy of Bacteria			99.98	Accuracy of Bacteria			100
Accuracy of Eukarya			92.35	Accuracy of Eukarya			0
Overall Accuracy			**99.82**	**Overall Accuracy**			**97.93**

Table 7
Confusion matrix and overall accuracies of the proposed SVM and DBN methods using code generation (0–8) and one RefSeq of Eukarya for all classes.

SVM Classifier				DBN Classifier			
With Code Generation (0–8)				With Code Generation (0–8)			
Confusion Matrix				Confusion Matrix			
	Archaea	Bacteria	Eukarya		Archaea	Bacteria	Eukarya
Archaea	832	0	0	Archaea	0	832	0
	100%	0%	0%		0%	100%	0%
Bacteria	0	17,745	2	Bacteria	0	17,747	0
	0%	99.99%	0.011%		0%	100%	0%
Eukarya	26	6	360	Eukarya	0	392	0
	6.63%	1.53%	91.84%		0%	100%	0%
Classification Accuracy (%)				**Classification Accuracy (%)**			
Accuracy of Archaea			100	Accuracy of Archaea			0
Accuracy of Bacteria			99.99	Accuracy of Bacteria			100
Accuracy of Eukarya			91.84	Accuracy of Eukarya			0
Overall Accuracy			**99.82**	**Overall Accuracy**			**93.55**

Table 8
Training time of the proposed SVM and DBN methods using code generation (0–8) and standard RefSeq for each class.

	SVM Classifier	DBN Classifier
Training Time (Seconds)	1495	82,609

characterization to decrease the execution time of grouping and to avoid the errors in the classifier [38].

The majority of research has been concentrated on the digital image analysis of Bacterial species [39,40]; however, a bioinformatics viewpoint for microbial sequences currently suffers from a shortage of scientific research studies.

In this paper, the exploitation of the EzBioCloud database is revisited in the evolutionary tree and the taxonomic search tools, as well as developed genomic representation of 16S rRNA gene and genome sequences. The proposed approach included a preprocessing phase with an alignment algorithm and code generation, and a classification phase with SVM and DBN algorithms. This approach is designed to differentiate between various types of Archaea, Bacteria, and Eukarya sequences by the SVM and DBN classifiers. The compatible RefSeq and encoding step are controlled in the quality of classification. According to the promising results, the factor of accuracy and speed are combined in the SVM classifier to gain a higher level of classification by using an encoding step (0–8) and standard RefSeq for each class. For selection of one reference for all classes, the Bacteria RefSeq for all database via SVM classifier outperformed the Archaea and Eukarya RefSeqs. The validation of our approach is demonstrated by supporting the perspective of physiological relationships that are based upon machine learning techniques. This work encouraged the integration of the bioinformatics field and clinical microbiology, which will be prominent in the future research studies.

5. Conclusion

The effectiveness of this proposed model is obtained by the bioinformatics analysis that depended on the classification of microorganism's database. The machine learning approaches, namely SVM and DBN, were successfully adapted to classify the multi-classes in the EzBioCloud database. The strategy of selecting the proper RefSeq with code generation process was investigated. The proposed study concludes that the better choice was the standard reference for each class with an encoding process of (0–8). This model perceived the thought of contrast zones among the genomic sequences and took into account updating of the evolutionary relationships in the tree of life. Finally,

this approach optimizes the classification methods of microbial sequences that are related to the three-domains system.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Roskov Y., Kunze T., Paglinawan L., Orrell T., Nicolson D., Culham A., Bailly N., Kirk P., Bourgoin T., Baillargeon G., Hernandez F., De Wever A. Species 2000 & ITIS Catalogue of life; 2013 annual checklist. Species 2000. Reading, UK.

[2] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.

[3] Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 2008;36:6688–719.

[4] Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Anderson I, Lykidis A, Mavromatis K, Ivanova NN, Kyrpides NC. The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res 2009;38:382–90.

[5] Peterson J, Umayam L, Dickinson T, Hickey E, White O. The comprehensive microbial resource. Nucleic Acids Res 2001;29:123–5.

[6] Uchiyama I. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res 2003;31:58–62.

[7] Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 1990;87:4576–9.

[8] Clarridge 3rd JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 2004;17:840–62.

[9] Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S, Nadarajan R. Brodie eoin l, lynch susan V. Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. PLoS One 2015;10:e0117617.

[10] Letunic I, Bork P. Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 2007;23:127–8.

[11] Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inf Exp 2012;2:1–12.

[12] Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbial 2007;73:5261–7.

[13] Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 2012;28:1033–4.

[14] Chen IMA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Kyrpides NC. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res 2017;45:507–16.

[15] Blom J, Albaum SP, Doppmeier D, Pühler A, Vorhölter FJ, Zakrzewski M, Goesmann A. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinf 2009;10:1–14.

[16] Blom J, Kreis J, Spänig S, Juhre T, Bertelli C, Ernst C, Goesmann A. EDGAR 2.0: an enhanced software platform for comparative gene content analyses. Nucleic Acids Res 2016;44:22–8.

[17] Whiteside MD, Winsor GL, Laird MR, Brinkman FS. OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. Nucleic Acids Res 2013;41:366–76.

[18] Yu J, Blom J, Glaeser SP, Jaenicke S, Juhre T, Rupp O, Schwengers O, Spänig S, Goesmann A. A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies. J Biotechnol 2017;261:2–9.

[19] Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, Chun J. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Syst Evol Microbiol 2017;67:1613–7.

[20] Tanaseichuk O, Borneman J, Jiang T. Phylogeny-based classification of microbial communities Oxford Academic Bioinformatics 2014;30:449–56.

[21] Yooseph S, Li W, Sutton G. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. BMC Bioinf 2008;9:182.

[22] Marsh JW, Hayward RJ, Shetty AC, Mahurkar A, Humphrys MS, Myers GSA. Bioinformatic analysis Search for other works by this author on: oxford Academic PubMed Google Scholar of bacteria and host cell dual RNA-sequencing experiments. Briefings Bioinf 2017.

[23] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.

[24] Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput 2006;18:1527–54.

[25] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. JMolBiol 1990;215:403–10.

[26] Christof A, Tanel P, Leopold P, Oliver S. Deep learning for computational biology. Mol Syst Biol 2016;12:878.

[27] Trevor H, Robert T, Jerome F. The elements of statistical learning: data mining, inference, and prediction. second ed. Verlag: Springer; 2009.

[28] Myers EW, Miller W. Optimal alignments in linear space. Bioinformatics 1988;4:11–7.

[29] Afify H, Islam M. Abdel Wahed M. DNA Lossless Compression Algorithm based on similarity of genomic sequences database. Int J Comput Sci Inf Technol 2011;3:145–54.

[30] Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 2013;30:472–9.

[31] Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. Singapore: World Scientific; 2002.

[32] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. first ed. Cambridge University Press; 2000.

[33] Lee T, Yoon S. Boosted categorical restricted Boltzmann machine for computational prediction of splice junctions. International conference on machine learning, vol. 37. 2015. p. 2483–92.

[34] Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, Zeng J. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res 2015;44:e32.

[35] Salakhutdinov R, Larochelle H. Efficient learning of deep Boltzmann machines. Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010. p. 693–700.

[36] Al-Antari MA, Al-Masn MA, Park SU, Park JH, Metwally MK, Kadah YM, Han SM, Kim TS. An automatic computer-aided diagnosis system for breast cancer in digital mammograms via deep belief network. J Med Biol Eng 2017:1–14.

[37] Bourbeau PP, Ledeboer NA. Automation in clinical microbiology. J clin microbial 2013;51:1658–65.

[38] Zieliński B, Plichta A, Misztal K, Spurek P, Brzychczy-Wøoch M, Ochońska D. Deep learning approach to bacterial colony classification. PLoS One 2017;12:1–14.

[39] Bruyne K, Slabbinck B, Waegeman W, Vauterin P, De Baets B, Vandamme P. Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. Syst Appl Microbiol 2011;34:20–9.

[40] Trattner S, Greenspan H, Tepper G, Abboud S. Automatic identification of bacterial types using statistical imaging methods. IEEE Trans Med Imag 2004;23:807–20.