Journal of
Data Mining in Genomics & Proteomics

Maleki et al., J Data Mining Genomics Proteomics 2015, 6:2
DOI: 10.4172/2153-0602.1000171

Research article Open Access

Promoter Prediction in Bacterial DNA Sequences Using Expectation Maximization and Support Vector Machine Learning Approach

Ahmad Maleki[1*], Vahid Vaezinia[2] and Ayda Fekri[2]

[1]Department of Computer Engineering, Damavand Branch, Islamic Azad University, Damavand, Iran
[2]Iran University of medical sciences, Tehran, Iran

Abstract

Promoter is a part of the DNA sequence that comes before the gene and is key as a regulator of genes. Promoter prediction helps determine gene position and analyze gene expression. Hence, it is of great importance in the field of bioinformatics. In bioinformatics research, a number of machine learning approaches are applied to discover new meaningful knowledge from biological databases. In this study, two learning approaches, expectation maximization clustering and support vector machine classifier (EMSVM) are used to perform promoter detection. Expectation maximization (EM) algorithm is used to identify groups of samples that behave similarly and dissimilarly, such as the activity of promoters and non-promoters in the first stage, while the support vector machine (SVM) is used in the second stage to classify all the data into the correct class category. We have applied this method to datasets corresponding to σ24, σ32, σ38, σ70 promoters and its effectiveness was demonstrated on a range of different promoter regions. Furthermore, it was compared with other classification algorithms to indicate the appropriate performance of the proposed algorithm. Test results show that EMSVM performs better than other methods.

Keywords: Promoter; Expectation maximization (EM) clustering; Support vector machine (SVM) classifier; Machine Learning; *E. coli*

Introduction

DNA is the molecule that contains genes and genetic information. Genetic information determines the appearance and function of living organisms. In addition to genes, DNA strands contain other parts such as gene regulatory regions that are necessary for gene expression regulation and specify organisms' phenotypes. These regulations are parts of various stages of the optimal utilization of gene information. A promoter is a part of gene regulation region that is RNA polymerase connection position and the site of transcription initiation. In bacteria, core RNA polymerase contain five subunits: Beta, Beta′, Omega, and two Alpha subunits (α2ββ′ω). To compose holoenzyme and recognize promoters, core polymerase requires another subunit that is specialized for contacting promoter sequences and called the sigma factor (σ) [1-3]. *E. coli* RNA polymerase have seven σ factors which each bind to different groups of promoters: one primary sigma factor (σ70) which is the major sigma factor and recognize housekeeping gene, and six alternative sigma factors (σ19, σ24, σ28, σ32, σ38, σ54) which are used in particular conditions such as stress [4,5]. Promoter sequences include certain characteristics in order to be identified by the sigma subunit such as two short sequence elements in approximately 10 and 35 bp upstream from the transcription start site. For σ70 the two conserved sequences are TTGACA and TATAAT with 17 ± 1 base space between them [6]. That is the optimal template of promoter sequences and the nucleotides, and their interval could be different compared with this template in these two conserved sequences [7]. The rate of similarity between the promoter sequences and the template indicates the strength and function of the promoters [8]. Among the other elements that help RNA polymerase to recognize and connect to promoters are extended -10 and UP elements. The extended -10 is a weakly conserved sequence that has two nucleotides and is upstream of the -10 elements [9,10]. The UP element is located upstream of the -35 region, recognized by α- subunit. The presence of this element increases promoter strength [11,12]. Since promoters are located just before genes, the identification of their exact location is very important because it represents the location and the start site of the genes. Prediction in the presence of promoters in eukaryotes and prokaryotes by using different algorithms is one of the most widely used and important methods of finding genes [13-17].

The machine learning approaches are very important computational methods and efficacious tools in bioinformatics research. In classification, the task is to predict the outcome associated with a particular object given a feature vector describing that object. In clustering, objects are grouped together because they share certain properties [18]. In this work, two learning approaches for prediction of promoters in bacterial DNA sequences are used which are the expectation maximization method for clustering and the support vector machine method for classification. We call this combination expectation maximization support vector machine (EMSVM). Clustering is useful in bacterial DNA for separating promoter sequences from non-promoter sequences. Clustering provides significant advantages over classification techniques which help identify groups of data earlier that behave similarly or show similar characteristics. When compared with the support vector machine and another algorithm, EMSVM shows a significant improvement in promoter's prediction accuracy.

The rest of the paper is organized as follows. Section 2 illustrates datasets, reviews related work and describes the proposed method. Computational experiments and results are presented in Section 3. Finally, conclusions are drawn in Section 4.

***Corresponding author:** Ahmad Maleki, Department of Computer Engineering, Damavand Branch, Islamic Azad University, Damavand, Iran, Tel: 98 912 737 5680; E- mail: ahmadmaleki@live.com

Received June 04, 2014; **Accepted** June 30, 2015; **Published** July 08, 2015

Citation: Maleki A, Vaezinia V, Fekri A (2015) Promoter Prediction in Bacterial DNA Sequences Using Expectation Maximization and Support Vector Machine Learning Approach. J Data Mining Genomics Proteomics 6: 171. doi:10.4172/2153-0602.1000171

Copyright: © 2015 Maleki A, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Materials and Methods

Dataset

In this paper, we used the promoter regions in *E. coli* genome for the implementation of the proposed method [19]. For this purpose, the σ24, σ32, σ38, σ70 promoters were used which contain 520, 309, 217, 1907 instances respectively. All promoter regions contain 81 nucleotides between -60 to +21 bp, respectively, downstream and upstream from the transcription start site. Each of these datasets is located alongside the 2000 non-promoters (which are randomly extracted from the *E. coli* genome non-promoter areas).

Furthermore, the dataset applied in previous research (*E. coli*-2) was used in order to evaluate the performance of the proposed algorithm. This dataset consists of 106 instances, 53 of which are positive instances (promoters). The negative instances were generated from larger DNA sequences that are believed to contain no promoters. An instance consists of a sequence of 57 nucleotides (fifty nucleotides prior to the beginning of the gene along with seven nucleotides in the beginning of the gene)[20].

Related work

The need to provide promoter prediction in *E. coli* has motivated the research community to deal with the problem of the detection of promoters. Burden et al. suggested a series of time delay neural networks (TDNNs) to model multiple promoter elements. They demonstrate greatly improved accuracy when distance to gene start site is incorporated into the models. However, the number and type of model elements was fixed and the TDNNs are typically time consuming to train [15]. Li et al. proposed a position correlation scoring matrix (PCSM) algorithm by using conservative hexamer segments for predicting whether s70 promoter is present [20]. Gordon et al. suggested a position weight matrices (PWM) method for promoter prediction using a variant of the mismatch string kernel. The SVM approach was more accurate than the PWM approach but the highest accuracy was obtained with a model that combined scores from the combination of SVM, PWMs and the gene start to TSS distance [21]. Rhodius et al. checked the ability of different methods to predict promoter strength and the sequence properties that distinguish between active and weak promoters. They discovered that a combination of selected modules is moderately predictive of promoter strength and that imposing minimal motif scores distinguished active from weak promoters [8].

Using *E. coli*-2 dataset, different methods were proposed by other researchers for more accurate predictions of promoters [22]. Towel et al. suggested using KBANN (Knowledge-Based Neural Network) for the prediction of promoters. KBANN is a hybrid approach mixing neural networks and domain knowledge. KBANN was designed specifically to show how adding domain knowledge could improve the performance of a neural net learning algorithm. This approach was compared with a decision tree induced with an ID3 algorithm, a multilayer perceptron neural network, a clustering algorithm k-NN and the technique known as O'Neill. Tavares et al. applied several machine learning methods and the hidden Markov model (HMM) to the construction of classifiers for detection of promoters in the DNA of *E. coli* [23]. Ramos-Pollán et al. described a machine learning framework for medical classification that is scaled for grid computing. They used the Biomed TK software to train artificial neural networks (ANNs) with different methods in order to use them for the classification of medical datasets [24].

Expectation maximization support vector machine (EMSVM) learning approach

Learning approaches provide high detection rate in identifying genes. The EMSVM learning approach is formed by combining clustering and classification techniques. The expectation maximization clustering technique is used as a pre-classification component for grouping similar data at an earlier stage. For the second step of the clustering, the data will be classified by the category of genes using support vector machine classifier. Thus, data which are misclassified during the first stage will be classified according to their category in the second stage.

The expectation maximization algorithm was first introduced in 1958 by Hartley et al. and developed in 1977 by Dempster et al. The EM algorithm is used to find the maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. One requirement of the EM parameter learning procedure is that initial values for the parameters be specified. With each iteration, the EM algorithm will try to find parameters that improve fit to the data by maximizing the log likelihood function; a measure of model's fit with each iteration. In other words, the EM algorithm is an algorithm for probability based clustering [25,26]. The probability distribution of (X, Z) can be written as L (θ; X, Z) = p(X, Z|θ). Thus, it must be maximized with the likelihood function:

$$L(\theta;X)=p(X|\theta)=\sum_{Z}P(X,Z|\theta) \quad (1)$$

Suppose that Z is the missing data and x is the observed data and the stepwise approach of EM algorithm requires parameter $\theta^{(t)}$ and the searcher's next step parameter $\theta^{(t+1)}$. These steps are also classified into the expectation (E) step and the maximization (M) step.

Expectation stage:

$$Q(\theta|\theta^{(t)})=E_{Z|X,\theta^{(t)}}\left[\log L(\theta;X,Z)\right] \quad (2)$$

Maximization stage:

$$\theta^{(t+1)}=\arg\theta\max Q\left(\theta|\theta^{(t)}\right) \quad (3)$$

In practical usage, we initialize $\theta^{(0)}$ with any other proper values (or vectors) and iteratively calculate $\theta^{(t)}$ to a desirable range of approximation level.

The support vector machine is emerging as a popular technique in machine learning [27]. This approach is a classification technique and is based on neural network technology [28]. It is a parametric statistical linear classifier that performs a nonlinear mapping of the input space to a new feature space to which a linear machine can be applied. SVM constructs a hyper plane separating the positive examples from negative ones in the new space representation. To avoid over fitting, SVM chooses the optimal separating hyper plane that maximizes the margin in the feature space [29]. The margin is defined as the minimal distance between the hyper plane and the training examples. The selected data points that support the hyper plane are called support vectors. A smaller number of support vectors reflect a better generalization for linearly separable problems. SVM employs a maximum margin hyper plane for separating examples belonging to two different classes [25]. A support vector machine is a useful method for data classification and regression analysis. A classification task generally involves training and testing data which consist of some data instances. Each instance (promoters, non-promoters) in the training set contains one attribute class (+, -) and several attributes (a, c, g, t). The aim of SVM is to produce a model which predicts the attribute class of data instances in the testing set

which are only given the attributes.

Given some training data:

$\{(x_i, y_i) | x_i \in R^p, y_i \in \{-1,1\}\}$ where i=1,2,3,…,n

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=0}^{n} \xi_i \qquad c > 0 \tag{4}$$

Subject to:

$$y_i \left(w.x_i - b \right) > 1 - \xi_i \qquad \xi_i \geq 0 \tag{5}$$

Here x_i is a feature vector, training vectors x are mapped into a higher (high or infinite) dimensional space by the function. Therefore, SVM finds a linear separating hyper plane with the maximal margin in this higher dimensional space. w is normal to the hyper plane, C is the penalty parameter of the error term and ξ_i is slack variables that measure the degree of prediction error of x_i for a given hyper plane. The parameter b/||W|| determines the offset of the hyper plane from the origin along the normal vector w. If the space separating the two target classes is not linear, the points can be transformed into another high dimensional feature space. If the transformation to the high dimensional space is $\varphi()$. Then:

$$K(X_i, X_j) = \varphi(X_i)\varphi(X_j) \tag{6}$$

K is called the kernel function. Kernel function has a good performance if the support vectors that are calculated by using the corresponding transformation are few and the classification of the test data is successful. With a suitable kernel, SVM can separate in the feature space the data point that in the original input space was non-separable. Two common kernels functions include:

Gaussian radial basis function kernel:

$$K(X_i, X_j) = e^{-\gamma \|X_i - X_j\|^2} \tag{7}$$

The polynomial kernel:

$$K(X_i, X_j) = (X_i.X_j + 1)^p \tag{8}$$

In the proposed automated diagnostic system, we experimented with both the Gaussian radial basis function kernel and the polynomial kernel.

The dataset related to the sigma 70 is used in Figure 1, which perfectly displays the process of the EMSVM algorithm. Initially, 3907 instances (1907 promoter and 2000 non-promoter) were classified into eight clusters used by the EM algorithm. The numbers are real numbers defined over the interval [0 10] in these clusters. The clustered data would be applied in the SVM algorithm in order to make predictions.

Computational Experiments and Results

In this paper, the EMSVM algorithm was implemented in Waikato

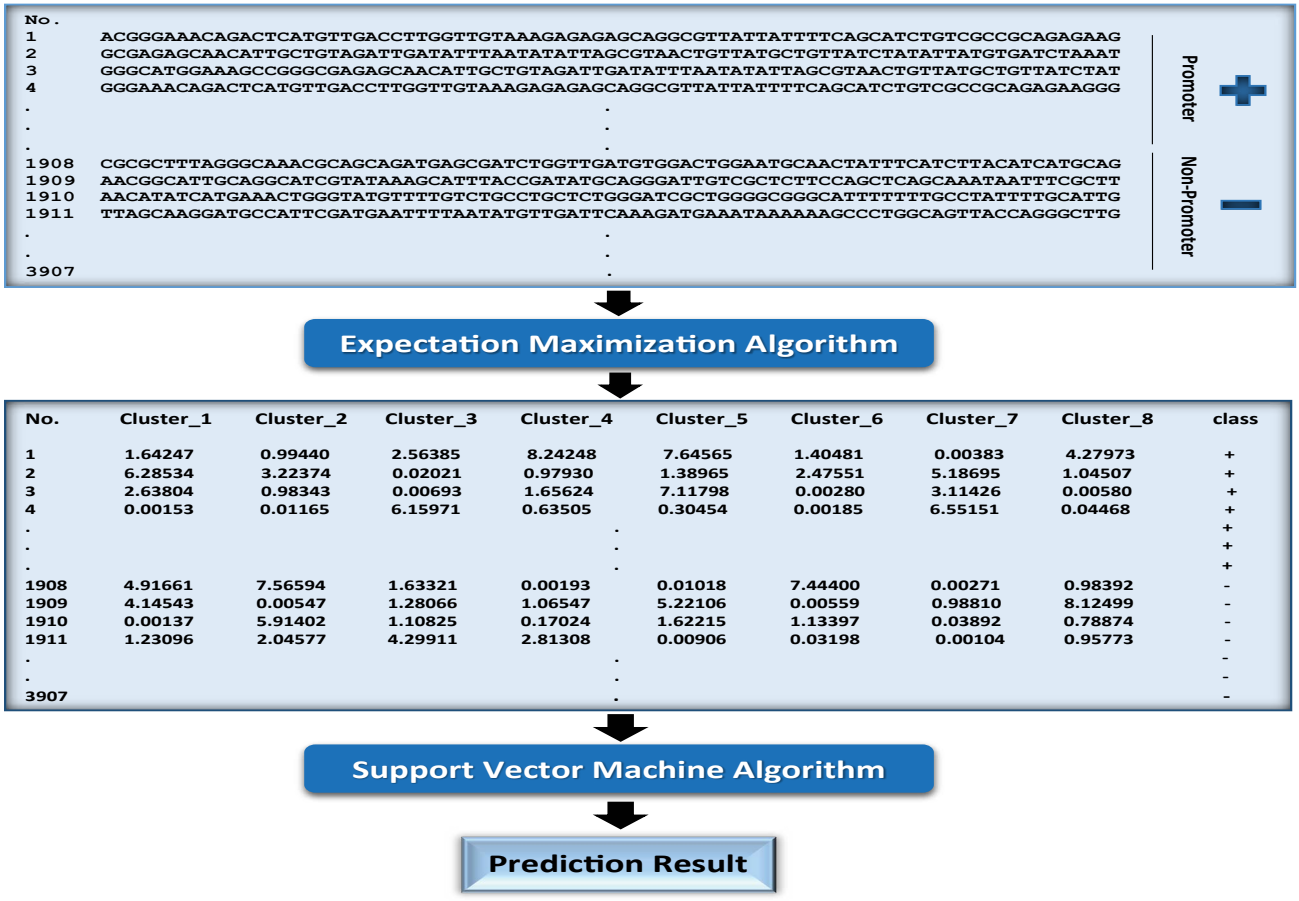

Figure 1: Process of the EMSVM algorithm using two stages.

environment for knowledge analysis (Weka). Weka contains tools for developing new machine learning schemes [18]. It can be used for pre-processing, classification and clustering [30].

Evaluation measures

Next, the performance of the method for calculating the sensitivity (SE), specificity (SP), accuracy (ACC) and the Matthews correlation coefficient (MCC) of the prediction will be evaluated. These parameters can be calculated using equations 9-12, where TP is positive instances (promoter sequences) classified as positive, TN is negative instances (non-promoter sequences) classified as negative, FP is negative instances (non-promoter sequences) classified a positive FN is positive instances (promoter sequences) classified as negative.

$$\mathrm{SE}(\text{sensitivity}) = \frac{TP}{TP + FN} \tag{9}$$

$$\mathrm{SP}(\text{specificity}) = \frac{TN}{TN + FP} \tag{10}$$

$$\mathrm{ACC}(\text{accuracy}) = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$\mathrm{MCC} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{12}$$

Results and discussions

In this step, the SVM and EMSVM algorithms would be compared and then the efficacy of the proposed methods would be studied on various areas of the σ24, σ32, σ38, σ70 promoters. At the end, the new algorithm would be compared to other algorithms by using the *E. coli*-2 dataset with the aim of appropriately demonstrating its performance.

The performance of all the classifiers was evaluated using a standard 10-fold cross-validation. In the 10-fold cross-validation, the dataset was partitioned into 10 subsets. Each subset had an equal ratio of promoter and non-promoter fragments. Each classifier was trained 10 times, each time using nine subsets for training, while keeping the 10th subset for testing. In this way, 10 models were generated during the cross-validation. The final prediction performance was obtained by averaging the results obtained from each model.

Performance comparison of SVM and EMSVM: The sigma 70 dataset (1907 promoter and 2000 non-promoter) was applied in order to compare the performance of the proposed method with the SVM algorithm. The experimental results for a single classifier support vector machine and EMSVM are summarized in Table 1, which represents measurements in terms of SE, SP, ACC, and MCC In order to develop an SVM model, we experimented with both polynomial and Gaussian radial base functions, as they are presented in (7) and (8), with three penalty parameters (C). The clustering techniques used as a pre-classification component for grouping similar data by classes in the earlier stage helps EMSVM produce a better result compared to the SVM classifier. For the EMSVM, the best accuracy rate belongs to the polynomial kernel with values of C = 1, P = 1 and the Gaussian radial basis function kernel with values of C = 1 and γ = 0.01

Compare influenced EMSVM on different areas of promoters: The plot of the distribution of nucleotides generally indicates an uneven and unequal distribution in the promoter regions which helps them to be identified. According to the lack of uniform distribution of a, c, g, t nucleotides in the promoter regions, the effectiveness of the EMSVM algorithm has been evaluated in this section. For this purpose, the two graphs are presented in Figure 2, which are related to the distribution of nucleotides in both the non-promoter region and a random region of the *E. coli* genome. The random distribution in these two graphs indicates the absence of clearly defined region which could be identified by the RNA polymerase subunits. It should be noted that this could reduce the possibility of detecting the region as a promoter. Figure 3 respectively displays the distribution plots of σ24, σ32, σ38, and σ70 promoters' nucleotides. The sequences have been classified into five intervals (-60 to +21, -60 to +1, -60 to +10, -40 to +21, -50 to +21) and the prediction accuracy of each region was measured by using an EMSVM algorithm (Table 2). The predicted results in each interval have a remarkable correlation with the distribution of the nucleotides in any promoter. The polynomial kernel with values C = 1, P = 1 was applied in all the predictions. There is no significant change in the results after the exclusion of the initial and final parts of the sigma 24 promoters, due to the distribution of specific nucleotides between -40 to +1 (it should be noted that there is no certain distribution in the close intervals) except for the -50 to +20 interval in which the exclusion of the final ten nucleotides would reduce the prediction accuracy. This decrease in accuracy could be improved by eliminating the -40 to -50 intervals (the percentage of the changes in the distribution of nucleotides would be significantly low). The sigma 32 contains a specific distribution of nucleotides in the intermediate interval, the same as sigma 24. Since there are certain areas with a minimum distribution

		SVM		EMSVM	
		p=1	γ=0.01	p=1	γ=0.01
C=1	SE	0.849	0.865	0.923	0.916
	SP	0.857	0.888	0.937	0.942
	ACC(%)	85.28	87.63	93.01	92.93
	MCC	0.705	0.753	0.860	0.859
C=2	SE	0.847	0.872	0.926	0.918
	SP	0.856	0.888	0.933	0.941
	ACC(%)	85.15	87.99	92.93	92.93
	MCC	0.703	0.760	0.859	0.859
C=3	SE	0.848	0.873	0.931	0.919
	SP	0.856	0.892	0.924	0.939
	ACC(%)	85.20	88.22	92.75	92.91
	MCC	0.704	0.764	0.855	0.858

Table 1: Results of the SVM and EMSVM algorithms using alternative kernel functions and penalty parameters C.

		-60+21	-60+1	-40+21	-60+10	-50+21
Sigma24	SE	0.867	0.871	0.887	0.871	0.84
	SP	0.979	0.976	0.978	0.977	0.973
	ACC(%)	95.59	95.43	95.87	95.51	94.56
	MCC	0.863	0.859	0.873	0.861	0.831
Sigma32	SE	0.709	0.751	0.803	0.819	0.819
	SP	0.973	0.978	0.978	0.982	0.977
	ACC(%)	93.76	94.75	95.4	95.97	95.53
	MCC	0.719	0.765	0.798	0.822	0.805
Sigma38	SE	0.705	0.645	0.756	0.682	0.774
	SP	0.974	0.976	0.98	0.974	0.981
	ACC(%)	94.76	94.31	95.76	94.49	96.07
	MCC	0.697	0.66	0.754	0.678	0.773
Sigma70	SE	0.923	0.909	0.927	0.912	0.921
	SP	0.937	0.934	0.933	0.929	0.933
	ACC(%)	93.01	92.14	93.01	92.06	92.73
	MCC	0.86	0.843	0.86	0.841	0.855

Table 2: Compare influenced EMSVM on different areas of promoters.

rate around the intermediate interval, the exclusion of these regions might totally increase the prediction accuracy. There is a little region with a certain distribution in the +7 to +13 intervals which reduces the rate of improvement in the results, in terms of the exclusion of the first 20 nucleotides compared with the first 10 nucleotides. There is no specific distribution of nucleotides in the -40 to -60 intervals in sigma 38, thus the results would be improved by the exclusion of this region. However, the exclusion of the little specific region in the +1 to +20 intervals could lead to a reduction in the ability to predict. The graph which is related to the sigma 70 variation indicates the non-normal distribution in a wider range compared with the other graphs so the predicted results would be changed or decreased by the exclusion of region in this sequence.

Performance comparison EMSVM and another algorithm: Table 3 gives detailed results of EMSVM for predicting promoter and non-promoter sequences by using the *E. coli*-2 dataset. The evaluation was performed in the form of a comprehensive comparison with previous studies [22-24]. The Matthews correlation coefficient (MCC), accuracy (ACC), sensitivity (SE) and the specificity (SP) values of the proposed method are equal to 0.98, 0.99, 0.98 and 1, respectively. ROC is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied [31]. Figure 4 shows a ROC space containing EMSVM and different algorithms that are used [23]. The x and y axes of the ROC space represent the false positive rate (FPR, equal to 1−specificity) and the true positive rate (TPR, the same as sensitivity), respectively.

A perfect classifier would be represented by the point (0, 1), which corresponds to the maximum specificity and sensitivity, the perfect classification. EMSVM has a better place in the ROC environment (0, 0.98) compared to other algorithms. This is because expectation maximization clustering technique that used as a pre-classification component in the first stage groups similar data respectively and instances which were misclassified during the first stage of clustering were classified correctly by support vector machine classifier in the second stage.

Conclusions

In summary, we have developed an accurate prediction method for detecting promoters in bacterial DNA sequences. This approach is called EMSVM and it is evaluated using σ24, σ32, σ38, σ70 datasets. EMSVM comprises two stages. In the first stage, data are clustered using the expectation maximization (EM), while in the second stage the data obtained from the previous stage are classified using the support vector machine (SVM). In this manuscript, a comparison was made between the EMSVM and another algorithm. The implementation results on the various *E. coli* datasets demonstrate that the proposed method has the best performance in the area of promoter prediction compared to the other methods proposed. The main innovation in this research is the combination of clustering and a classification algorithm which has considerable influence on increasing the accuracy of predictions; the second innovation is the proposed algorithm for the different intervals in the promoter sequence which lead to reasonable results consistent with the structure of the promoter. This paper presents some recommendations which will help the results of this article to be further developed by future research studies: the proposed algorithm should be

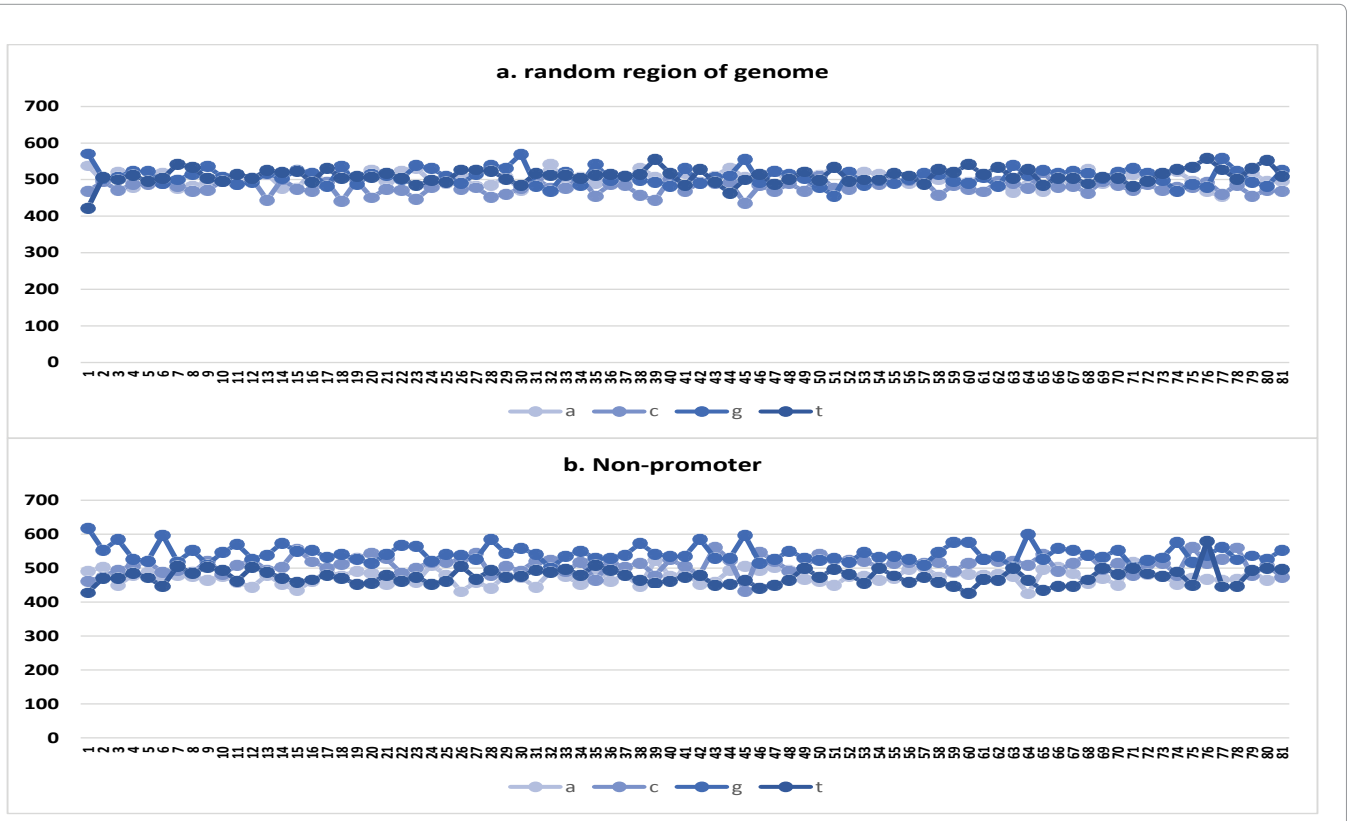

Figure 2: The plot of the distribution: a) 160 thousands nucleotides of *E. coli* genome (recognized as a random region, includes the promoter and non-promoter, divided into 81 equal intervals) b) 2000 non-promoters.

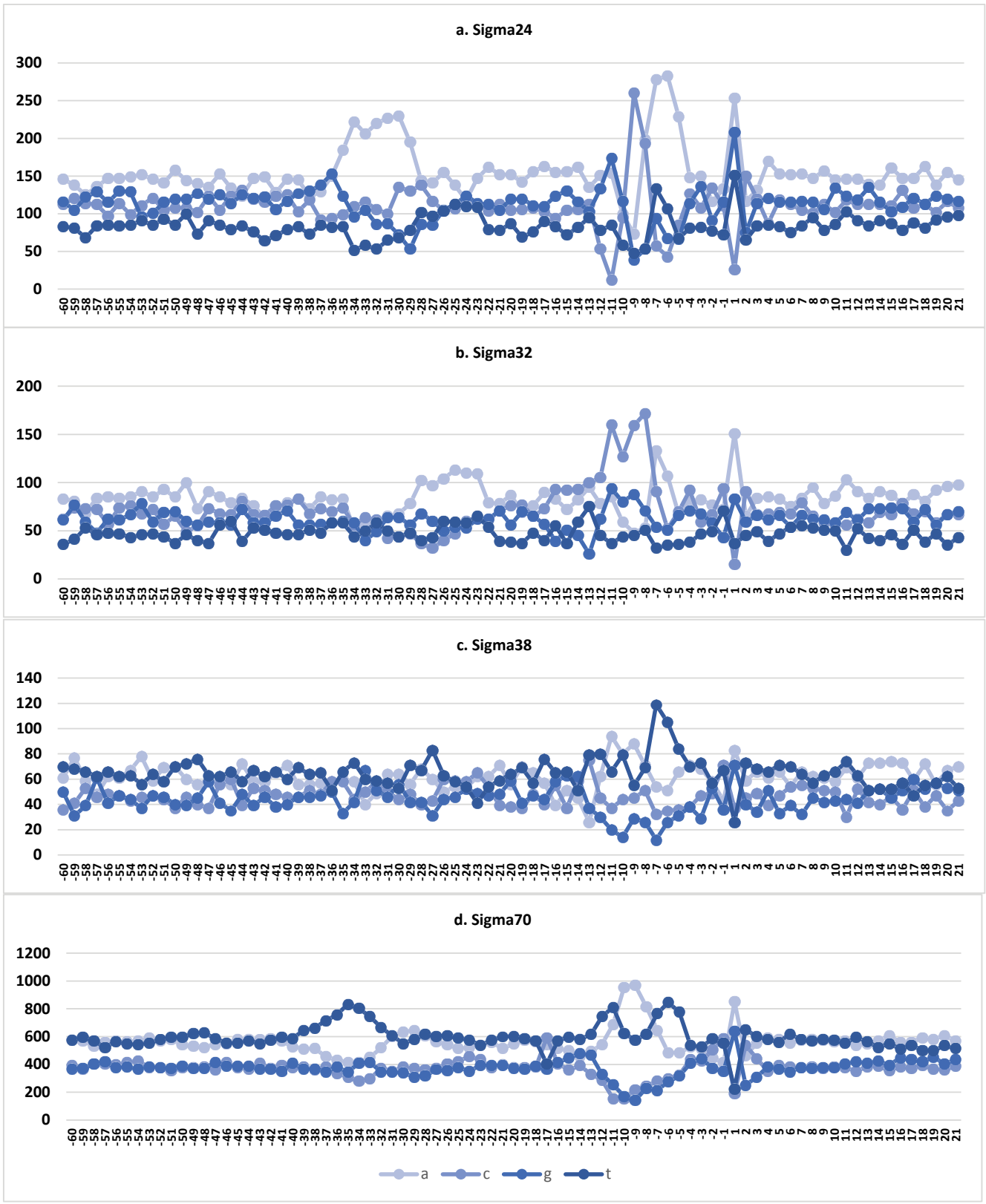

Figure 3: The plot of the distribution of nucleotides: a) 520 sigma 24 promoters, b) 309 sigma 32 promoters, c) 217 sigma 38 promoters, d) 1907 sigma 70 promoters.

Prediction method	Reference	tp	fn	fp	tn	MCC	ACC(%)
Expectation Maximization and Support Vector Machine(EMSVM)	This paper	52	1	0	53	0.98	99.05
Knowledge Based Neural Network(KBANN)	Geoffrey G. Towell et al. (1990)	-	-	-	-	-	96.22
Multilayer Perceptron		-	-	-	-	-	92.45
OtNeill's Method		-	-	-	-	-	88.68
K-Nearest Neighbors(k-NN)		-	-	-	-	-	87.74
Decision Tree (ID3)		-	-	-	-	-	82.08
Hidden Markov Model(HMM)	Leonardo G. Tavares et al .(2008)	50	3	5	48	0.850	92.45
Complement Class Naive Bayes(CNB)		49	4	3	50	0.868	93.40
Multilayer Perceptron Neural Network(MLP)		49	4	3	50	0.968	93.40
Support Vector Machine(SVM)		49	4	4	49	0.849	92.45
LogitBoost		47	6	5	48	0.793	89.62
NBTree		47	6	5	48	0.793	89.62
Lazy Bayesian Rules Classifier(LBR)		48	5	3	50	0.850	92.45
PART		44	9	11	42	0.623	81.13
ANN trained with backpropagation	Raúl Ramos-Pollán et al. (2012)	-	-	-	-	-	89.09
ANN trained with resilient propagation		-	-	-	-	-	94.36
ANN trained with simulated annealing		-	-	-	-	-	88.50
ANN trained with genetic algorithms		-	-	-	-	-	73.37

Table 3: Comparison between our method and other reported methods.

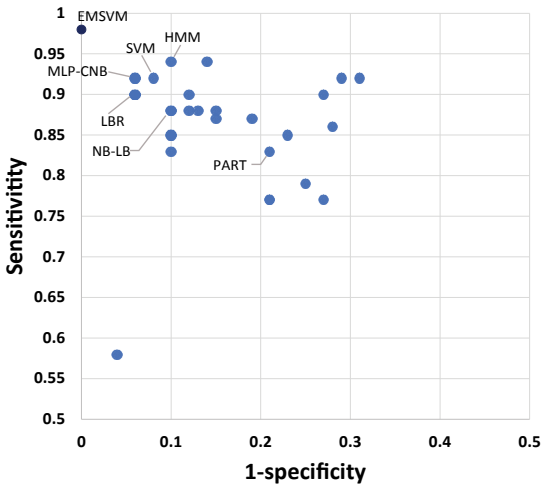

Figure 4: EMSVM algorithm in ROC space.

used on other datasets, in order to assess the efficiency of the algorithm compared to the other methods applied to the datasets considered. The other clustering algorithms and classification should be examined in order to compare their performance with the EMSVM. Our collected datasets should be applied. Furthermore, the divided intervals and also their results should be considered for further research in the field of *E. coli*.

References

1. Murakami KS, Masuda S, Campbell EA, Muzzin O, Darst SA (2002) Structural basis of transcription initiation: An RNA polymerase holoenzyme-DNA complex. Science 296: 1285-1290.
2. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16-23.
3. Young BA, Gruber TM, Gross CA (2002) Views of transcription initiation. Cell 109: 417-420.
4. Guha Thakurta D, Stormo GD (2001) Identifying target sites for cooperatively binding factors. Bioinformatics 17: 608-621.
5. Rhodius VA, Suh WC, Nonaka G, West J, Gross CA (2005) Conserved and variable functions of the σE stress response in related genomes. PLoS biology 4: e2.
6. Gruber TM, Gross CA (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. Annual Reviews in Microbiology 57: 441-466.
7. Shultzaberger RK, Chen Z, Lewis KA, Schneider TD (2007) Anatomy of *Escherichia coli* σ70 promoters. Nucleic acids research 35: 771-788.
8. Rhodius VA, Mutalik VK (2010) Predicting strength and function for promoters of the *Escherichia coli* alternative sigma factor, σE. Proceedings of the National Academy of Sciences 107: 2854-2859.
9. Campbell EA, Muzzin O, Chlenov M, Sun JL, Olson CA, et al. (2002) Structure of the bacterial RNA polymerase promoter specificity σ subunit. Molecular cell 9: 527-539.
10. Koo BM, Rhodius VA, Campbell EA, Gross CA (2009) Mutational analysis of *Escherichia coli* σ28 and its target promoters reveals recognition of a composite− 10 region, comprised of an 'extended− 10'motif and a core− 10 element. Molecular microbiology 72: 830-843.
11. Mutalik VK, Nonaka G, Ades SE, Rhodius VA, Gross CA (2009) Promoter strength properties of the complete sigma E regulon of *Escherichia coli* and *Salmonella enterica*. Journal of bacteriology 191: 7279-7287.
12. Rhodius VA, Mutalik VK, Gross CA (2012) Predicting the strength of UP-elements and full-length *E. coli* σE promoters. Nucleic acids research 40: 2907-2924.
13. Abeel T, Van de Peer Y, Saeys Y (2009) Toward a gold standard for promoter prediction evaluation. Bioinformatics 25: 313-320.
14. Bland C, Newsome AS, Markovets AA (2010) Promoter prediction in *E. coli* based on SIDD profiles and Artificial Neural Networks. BMC bioinformatics 11: S17.
15. Burden S, Lin Y-X, Zhang R (2005) Improving promoter prediction improving promoter prediction for the NNPP2, 2 algorithms: a case study using *Escherichia coli* DNA sequences. Bioinformatics 21: 601-607.
16. Demeler B, Zhou G (1991) Neural network optimization for *E. coli* promoter prediction. Nucleic acids research 19: 1593-1599.
17. Xie X, Wu S, Lam K-M, Yan H (2006) Promoter Explorer: an effective promoter identification method based on the AdaBoost algorithm. Bioinformatics 22: 2722-2728.
18. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479-2481.
19. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, et al. (2013) RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic acids research 41: D203-D213.

20. Li Q-Z, Lin H (2006) The recognition and prediction of σ 70 promoters in *Escherichia coli* K-12. Journal of theoretical biology 242: 135-141.

21. Gordon JJ, Towsey MW, Hogan JM, Mathews SA, Timms P (2006) Improved prediction of bacterial transcription starts sites. Bioinformatics 22: 142-148.

22. Towell GG, Shavlik JW, Noordewier M (1990) Refinement of approximate domain theories by knowledge-based neural networks. Proceedings of the eighth National conference on Artificial intelligence, Boston, MA: 861-866.

23. Tavares LG, Lopes HS, Lima CRE (2008) A comparative study of machine learning methods for detecting promoters in bacterial DNA sequences. Advanced Intelligent Computing Theories and Applications with Aspects of Artificial Intelligence: Springer 5227: 959-966.

24. Ramos-Pollán R, Guevara-López MÁ, Oliveira E (2012) A software framework for building biomedical machine learning classifiers through grid computing resources. Journal of medical systems 36: 2245-2257.

25. Campbell C (2014) Machine Learning Methodology in Bioinformatics. Springer Handbook of Bio-/Neuroinformatics: 185-206.

26. Dellaert F (2002) The expectation maximization algorithm.

27. Han J, Kamber M (2006) Data Mining, Southeast Asia Edition: Concepts and Techniques. (2ndedn) Morgan kaufmann, USA.

28. Vapnik VN, Vapnik V (1998) Statistical learning theory. (1stedn) Wiley, New York, USA.

29. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2: 121-167.

30. Sharmaa BR, Paula A (2013) Clustering Algorithms: Study and Performance Evaluation Using Weka Tool. International Journal of Current Engineering and Technology 3: 1094-1098.

31. Fawcett T (2006) An introduction to ROC analysis. Pattern recognition letters 27: 861-874.