Computational Prediction of Sigma-54 Promoters in Bacterial Genomes by Integrating Motif Finding and Machine Learning Strategies

Bingqiang Liu, Ling Han, Xiangrong Liu, Jichang Wu, and Qin Ma

Abstract—Sigma factor, as a unit of RNA polymerase holoenzyme, is a critical factor in the process of gene transcriptional regulation. It recognizes the specific DNA sites and brings the core enzyme of RNA polymerase to the upstream regions of target genes. Therefore, the prediction of the promoters for a particular sigma factor is essential for interpreting functional genomic data and observation. This paper develops a new method to predict sigma-54 promoters in bacterial genomes. The new method organically integrates motif finding and machine learning strategies to capture the intrinsic features of sigma-54 promoters. The experiments on *E. coli* benchmark test set show that our method has good capability to distinguish sigma-54 promoters from surrounding or randomly selected DNA sequences. The applications of the other three bacterial genomes indicate the potential robustness and applicative power of our method on a large number of bacterial genomes. The source code of our method can be freely downloaded at https://github.com/maqin2001/PromotePredictor.

Index Terms—Computational genomics, DNA motifs, gene transcription, machine learning

1 Introduction

Transcription is the first step of gene expression, leading genetic information in a cell to all kinds of biological functions [1]. It determines the activity of the majority of genes in a particular circumstance, which is initiated through interactions between RNA polymerase and a specific DNA sequence, often called promoter, together with one or more general transcription factors as trans-regulatory elements. The core enzyme of RNA polymerase contains five subunits (β, β', α^{I} and α^{II}, and ω) [2]. To bind to DNA sequences, RNA polymerase core is associated with a sigma (σ) factor to compose RNA polymerase holoenzyme [3], [4]. The σ–factor directs the core enzyme to transcript-specific genes by recognizing corresponding promoters, i.e., the σ–factor selects which genes will be transcribed. There are several kinds of σ–factors in bacterial species according to their molecular weights, and the number of σ–factors varies between species [5]. In the most well-studied model organism, *E. coli*, the most popular one is $\sigma - 70$, which has a molecular weight of $70\,\mathrm{kDa}$ and transcribes most genes in growing *E. coli* cells [6], [7]. Another important one in *E. coli* is σ-54, which plays essential regulatory roles in nitrogen metabolism and assimilation under nitrogen limiting conditions, as well as a variety of other cellular processes [8].

Obviously, determining the binding promoters for a specific σ-factor is essential for any further studies in gene regulation and functional genomics [9]. Since the experimental identification of such promoters is expensive and time-consuming, the computational prediction of σ-factor promoters becomes a vital bioinformatics problem [10]. The main strategies used to computational predict σ-factor binding sites include phylogenetic footprinting and motif finding. The former strategy relies on the assumption that the functional elements in the upstream region of coding genes may have higher conservation than surrounding nucleotides among variant species through evolutionary pressure [9]. Although phylogenetic footprinting can detect the potential binding sites, it is difficult to direct these binding sites to a specific σ-factor. The motif finding based methods usually analyze potential sequence specificity of the to-be-discovered σ-factor binding sites. For example, $\sigma - 70$ promoters have a canonical model, which contains a -35 hexamer, and $\mathrm{a} - 10$ hexamer, with consensus sequences TTGACA and TATAAT, respectively [11]. For $\sigma - 54$ promoters, the corresponding two elements are located around -12 and -24 regions from the transcription start sites of downstream genes. However, the flexibility of the DNA motif bound by the σ-factor is difficult to capture in an efficient way computationally. In addition, other known or unknown DNA elements in the promoter region, as well as DNA shape information [12], [13], also can help to predict promoters, and many factors can affect the prediction performance [14].

- B. Liu, L. Han, and J. Wu is with the School of Mathematics, Shandong University, Jinan, Shandong 250100, China.
E-mail: bingqiang@sdu.edu.cn, hlingly@163.com, jichangwu@126.com.
- X. Liu is with the Department of Computer Science, Xiamen University, Xiamen, Fujian 361005, China. E-mail: xrliu@xmu.edu.cn.
- Q. Ma is with the Department of Mathematics & Statistics and the Department of Agronomy, Horticulture & Plant Sciences, South Dakota State University, Brookings, South Dakota 57006.
E-mail: qin.ma@sdstate.edu.

Manuscript received 12 Sept. 2017; revised 14 Jan. 2018; accepted 16 Feb. 2018. Date of publication 15 Mar. 2018; date of current version 5 Aug. 2019.
(Corresponding author: Bingqiang Liu).
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2816032

1545-5963 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

In previous decades, several algorithms have been developed to predict promoters [15], [16]. From the computational point of view, the promoter sequences can be formulated as a classification problem, i.e., how to determine a given sequence is a promoter of a specific σ-factor or not based on its encoded features. Therefore, several classification techniques, including support vector machine (SVM), hidden Markov model, position weight matrix, artificial neural network, random forest, *etc*., have been used to predict promoters from given sequences, and different types of sequence features, including *k*-mers and z-curve, *etc*., have been extracted based on these models [10], [17], [18], [19], [20], [21].

Abeel et al. designed a core promoter prediction tool, named ProSOM [22] based on unsupervised clustering by using the self-organizing maps, by which the core promoters can be distinguished from the rest of genome. ProSOM has a comparative performance than other existing tools but results in a significantly lower false positive rate [23]. Touzain et al. developed a tool, named SIGffRid [9], to search for σ-factor binding sites in bacterial genomes. SIGffRid performs a simultaneous analysis for promoter regions of orthologous genes and identifies the over-represented sequence patterns. This method considered both statistical and biological criteria and obtained an excellent performance on the prediction of $\sigma - 70$ binding sites. Recently, Lin et al. proposed a very promising method to identify $\sigma - 54$ promoter in *E. coli* genomes and built a high-quality benchmark dataset to train and test the predictor [10]. The new method is based on analysis of pseudo *k*-tuple nucleotide composition, which has been successfully applied in many fields [19], [24], [25], [26], [27], [28], [29], [30], [31]. This idea was further optimized by the incremental feature selection procedure aiming to achieve a better performance than existing methods. A software named iPro54-PseKNC and a user-friendly web server at http://lin.uestc.edu.cn/ server was provided to facilitate its application in public domain and have been widely downloaded and visited [32], [33], [34]. Despite these efforts on computational prediction of promoters, further progress is still needed to improve the state-of-the-art performance. The $\sigma - 70$ promoter has been well studied, but other sigma factors in bacterial still lack more study on their binding specificity. In addition, the universality needs to improve, since most of the prediction methods are generated on a model organism, and their performance on other species have not been well validated.

In this paper, we, for the first time, combined motif finding and classification strategy to design a new $\sigma - 54$ promoter prediction method. Different with previous tools, the new method starts on motif finding, for conserved sequence patterns, on both training promoter sequences and generate negative data. The identified motifs can capture the subtle variation of functional DNA elements, and take full advantage of the regulatory elements related to σ-factor binding. Multiple classification methods are applied to the features generated based on motif profiles, and the feature dimension was reduced to improve efficiency and accuracy of the new method. We test the new method on the benchmark data [10] and compare with iPro54-PseKNC, which is one of the best sigma-54 promoter prediction methods based on our knowledge. The overall performance of our method is comparable with iPro54-PseKNC. In addition, we applied the new methods in other three bacterial genomes, and the results indicated that it is robust for different bacterial genomes. It worth noting that, the method can be readily applied to predict other promoters with annotated training data. The algorithm is implemented into a tool, which can be download at https://github.com/maqin2001/PromotePredictor with source code and instruction file.

2 Materials and Methods

2.1 Data Preparation

High-quality benchmark data is required to select features and train the criteria for prediction, including positive dataset, i.e., experimentally confirmed $\sigma - 54$ promoter sequences, and negative dataset, i.e., randomly generated data or randomly selected non-promoter data from a genome. Lin et al. [10] created a benchmark data set on the *E. coli* genome from the RegulonDB database (http://regulondb.ccg.unam.mx/) [35] and Barrios's previous study [4]. They took the 81bp-long sequences (from -60 to $+20\text{bp}$ of transcription start site) and filtered the sequences with ambiguous nucleotides. The redundant sequences (sharing high sequence similarity) were also removed from benchmark data set by considering the statistical representativeness. The random sequences with the same length are generated from the coding regions and intergenic regions those do not contain $\sigma - 54$ binding sites. The final data set includes 161 positive and 161 negative sequences.

Furthermore, we collected computationally predicted $\sigma - 54$ promoters from the Sigma 54 Promoter Database (http://www.sigma54.ca/) to evaluate the performance and the robustness of our method in different species. The database was built by Dr. Boddy's lab in University of Ottawa [36] and three species were included in our study, which are *Bacillus subtilis* (NC_000964), *Clostridium acetobutylicum* (NC_003030), and *Lactobacillus brevis* (NC_008497). B. *subtilis* is another well-studied model organism, and is considered the best studied Gram-positive bacterium [37]; Clostridium is one of the most important biomass degraders in bioenergy research [38], and Lacto is a well-known bacterial genome related to food science and human health [39]. The prediction was based on the positional weighted matrix (PWM) of the $\sigma - 54$ promoter consensus sequence from Barrios's experimental data [4] and the algorithm named PromScan (http://molbiol-tools.ca/promscan/) [40], which has been widely used in $\sigma - 54$ promoter analysis and obtained excellent performance [41], [42], [43], [44], [45]. Similarly, we also generate negative dataset by randomly cut DNA sequences form both coding regions and intergenic regions without known $\sigma - 54$ promoters. Finally, our dataset contains 78, 30, and $68\sigma - 54$ promoters, with the length as of 81bp, for genome NC_000964, NC_003030, NC_008497, respectively. The negative datasets have the number of sequences and the same length with positive data sets for each of the three genomes.

2.2 Motif Finding on *E. coli* Data

De novo motif finding analysis was carried out through the DMINDA web server [46], [47], aiming to capture the sequence features of $\sigma - 54$ promoter sequences. DMINDA is an integrated regulatory DNA motif prediction and analysis platform. It contains five computational functions including *de novo* motif finding [48], [49], motif searching, motif

TABLE 1
Performance of Our Method on *E. coli* Benchmark Data Before Dimension Reducing

Classification	Acc	Precision	Recall	Sn	Sp	F_measure	F_score	MCC
Nearest Neighbors	0.7671	0.7312	0.8447	0.8447	0.6894	0.7475	0.7475	0.5407
Logistic Regression	0.8043	0.7882	0.8323	0.8323	0.7764	0.7987	0.7987	0.6096
Bagging	0.8075	0.7592	0.9006	0.9006	0.7143	0.7877	0.7877	0.6259
Gradient Boosting	0.7950	0.7714	0.8385	0.8385	0.7516	0.7857	0.7857	0.5923
SGD	0.7609	0.7308	0.8261	0.8261	0.6957	0.7442	0.7442	0.5262
LibSVM	0.8075	0.7463	0.9317	0.9317	0.6832	0.7801	0.7801	0.6348
LinearSVC	0.8012	0.7771	0.8447	0.8447	0.7578	0.7922	0.7922	0.6048
Decision Tree	0.7764	0.7633	0.8012	0.8012	0.7516	0.7707	0.7707	0.5535
Random Forest	0.7764	0.7572	0.8137	0.8137	0.7391	0.7677	0.7677	0.5543
ExtraTrees	0.8168	0.7629	0.9193	0.9193	0.7143	0.7958	0.7958	0.6473
AdaBoost	0.7826	0.7542	0.8385	0.8385	0.7267	0.7697	0.7697	0.5688
Naive Bayes	0.7857	0.7473	0.8634	0.8634	0.7081	0.7677	0.7677	0.5784

scanning, motif co-occurrence analysis [50], phylogenetic-footprinting-based motif prediction [51], and co-regulated gene modules prediction [52]. It is noteworthy that 2,125 species with complete genomes have been integrated to DMINDA to support the above five functions, covering animals, plants and bacteria (DOOR) [53]. The *de novo* motif finding tool BOBRO [48] can evaluate each nucleotide about how possible it belongs to a DNA motif by a two-step sequence alignment strategy. The followed dense subgraph identification will build the motif seed on which the motif profile with a statistical significance *p*-value is determined. The *p*-values are calculated based on the following assumption. We defined a set of DNA segments, which similar with motif seed, as motif closure of it. The number of motif patterns from the input data belongs a motif closure is assumed to follow a binomial distribution, and them approximated by a Poisson distribution, under the consideration of computational convenience. The significance p-value is calculated as the tail probability of the occurrence number of motif patterns from input data on Poisson distribution [48], [49].

The motif finding was performed on both positive and negative datasets with input motif length range between 12 and 16. We ignored the short motifs $(< 12\,\text{bps})$ with the considering that prokaryotic DNA motifs are longer than eukaryotic motifs, and a short length may bring in too many noises in genome-scale motif finding from the input sequences. We take the motifs in negative data into consideration because probably there are some conserve patterns in coding regions or intergenic regions without σ-factor binding, which tend to be excluded by sigma binding area. Each predicted motif was transferred into a position frequency matrix: $M = \{f_{i,j}\}_{4\times l}$, where $f_{i,j}$ is the frequency of nucleotide i (belongs to $\{\mathrm{A, G, C, T}\}$) at each position j (from 1 to l, and l is motif length) of aligned motif profiles.

2.3 Feature Selection

The sequence features are extracted based on the predicted motif profiles from both positive and negative input sequences. We try to evaluate how possible that there is an occurrence pattern for any predicted motifs in a given sequence. For a motif m (with length l and position frequency matrix M) and an $l - \mathrm{mer}\, s$ in any given sequence S, we calculate the similarity score of them as:

$$S(m,s) = \frac{1}{l}\sum\nolimits_{j=1}^{l}\sum\nolimits_{i=1}^{4} p_{ij}, \tag{1}$$

where $p_{ij} = M_{ij}$ if the $j\mathrm{th}$ nucleotide is i in s, otherwise $p_{ij} = 0$. And the score for S on motif m is:

$$S(m,S) = \max_{all\ l-mers\, s\, in\, S} S(m,s). \tag{2}$$

The scores on each sequence for all predicted motif compose a feature vector. Based on these vectors, the classification method will train a criterion to distinguish whether the given sequence is potential σ-factor promoter or not.

2.4 Classification

Twelve classification strategies were applied in this work and prediction performance of them are thoroughly evaluated (Table 1). Then, three of the twelve are chosen to do further analysis, including bagging [54], random forest [55], and SVM [56], [57], because they tend to have higher and more robust prediction performance. Bagging, as a special case of the model averaging, can be used to improve the stability and accuracy of classification algorithms, reduce variance, and help avoid overfitting. Random forest is a classification method and operates by constructing a multitude of decision trees. SVM transferred the input vectors into a higher dimension space and made them separable by a hyperplane. The separating hyperplane will be the trained boundary for classification. In this study, we use a machine-learning package, WEKA [58], to perform bagging and random forest and use the libSVM package [59] to carry out SVM analysis. The feature vectors obtained in the last section will be the input for classification. The ten-fold cross-validation test was used to test the performance.

2.5 Dimension Reducing

In order to improve the efficiency and accuracy of the algorithm, and facilitate the subsequent analysis on important features, we perform dimension reducing on the three selected classification methods. MRMD (Maximum-Relevance-Maximum-Distance) [60] is a feature selection algorithm based on the maximum relevance maximum distance strategy, which can be download at https://github.com/ShixiangWan/MRMD. It takes Pearson correlation coefficient to calculate the correlation between features and classes. It also adopts many distance metrics to evaluate the similarity between a pair of features, and then reduce the reticence of features. The distance functions used in this study include Euclidean distance, cosine distance, Mean distance, Manhattan distance, Canberra distance, Bhattacharyya distance, Modified Cosine distance, and Spearman correlation coefficient. Specifically, for a pair of feature vectors X and Y:

$$X : (x_1, x_2, \ \ldots, \ x_n); Y : (y_1, y_2, \ \ldots, \ y_n). \tag{3}$$

The distance functions are following.

Euclidean distance:

$$ED(X,Y) = \sqrt{\sum\nolimits_{k=1}^{N} (x_k - y_k)^2} \tag{4}$$

Cosine distance:

$$CosD(X,Y) = \frac{\vec{X} * \vec{Y}}{\vec{X} * \vec{Y}} \tag{5}$$

Tanimoto coefficient:

$$TC(X,Y) = \frac{\vec{X} * \vec{Y}}{\vec{X} + \vec{Y} - \vec{X} * \vec{Y}} \tag{6}$$

Mean distance:

$$MeanD(X,Y) = \sqrt{\frac{1}{N}\sum\nolimits_{k=1}^{N} (x_k - y_k)^2} \tag{7}$$

Manhattan distance:

$$MD(X,Y) = \sum\nolimits_{k=1}^{N} |x_k - y_k|, \tag{8}$$

Canberra distance:

$$CD(X,Y) = \sum\nolimits_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{9}$$

Bhattacharyya distance:

$$BD(X,Y) = \ln\left(\sum\nolimits_{i=1}^{n} \sqrt{x_i y_i}\right) \tag{10}$$

Modified Cosine distance:

$$MCosD(X,Y) = -\ln\left[1 + \frac{\cos\left(\vec{X}, \vec{Y}\right)}{2}\right], \tag{11}$$

Spearman correlation coefficient:

$$SCC(X,Y) = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2 - 1)}. \tag{12}$$

2.6 Performance Evaluation

As described above, we carried out cross-validation on benchmark data to evaluate the performance of classification. For the classification results before and after dimension reducing, multiple metrics are calculated including accuracy (Acc), precision, recall, sensitivity (Sn), specificity (Sp), F-measure, F-score, and Matthews correlation coefficient (MCC). These metrics are defined based on true positive (TP), true negative (TN), false positive (FP) and false negative (FN) of predictions:

$$Acc = \frac{(TP \ + \ TN)}{(TP \ + \ FN \ + \ FP \ + \ TN)} \tag{13}$$

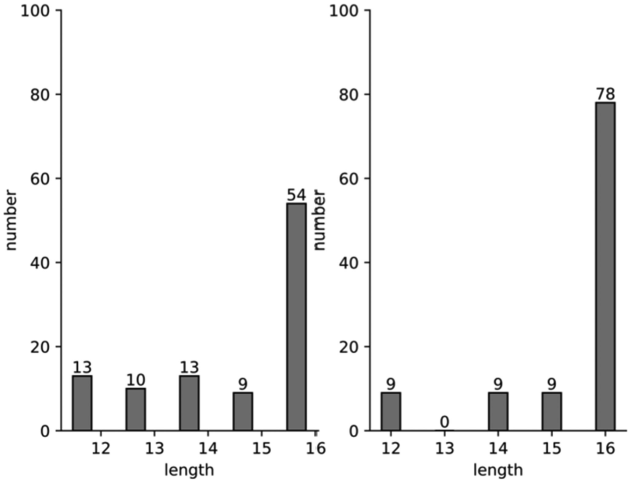


Fig. 1. The distribution of motif length from positive (A) and negative (B) benchmark dataset from *E. coli*. The x-axis is length of predicted motif, and the y-axis is the number of predicted motifs with the corresponding length.

$$precision = \frac{TP}{TP + FP} \tag{14}$$

$$recall = \frac{TP}{TP + FN} \tag{15}$$

$$Sn = \frac{TP}{(TP + FN)} \tag{16}$$

$$Sp = \frac{TN}{FP \ + \ TN} \tag{17}$$

$$F - measure = \frac{2 * recall * precision}{(recall + precision)} \tag{18}$$

$$F - score = recall * precision \tag{19}$$

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{20}$$

The performance of the new method is compared with the performance of iPro54-PseKNC, in terms Sn, Sp, Acc, and MCC. Further, to test the robustness of the method on other species, we applied the trained criteria of the new method and iPro54-PseKNC on three other genomes, NC_000964, NC_003030, and NC_008497, and compare the performance. The application of iPro54-PseKNC was performed by its webserver.

3 Results

3.1 Motif Finding and Feature Generation on Benchmark Data

We performed the *de novo* motif finding on *E. coli* benchmark data by BOBRO on the DMINDA web server, with the input motif length ranges from 12 to 16. We obtained 99 motif profiles from positive sequence set and 105 from negative sequence set. Fig. 1 showcases the distribution of length of prediction motifs, in which we can see the identified motifs tend to belong especially in the negative set. We

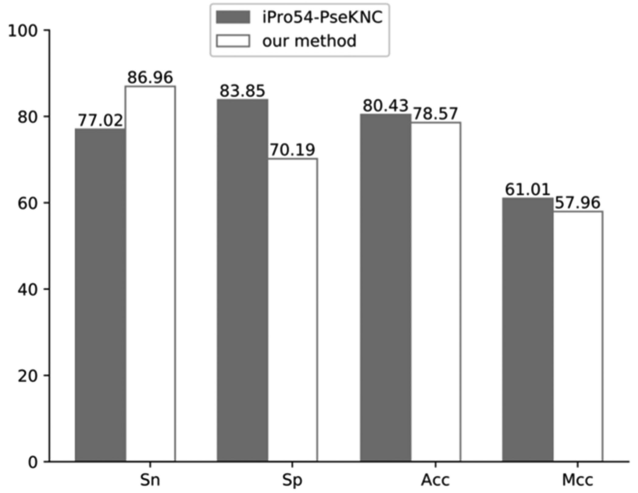


Fig. 2. The performance comparison between our method and iPro54-PseKNC. The gray bars show the metrics of iPro54-PseKNC, and the white bars indicate the average metrics of our method by adopting 12 classification methods.

calculated the position frequency matrix for these 204 motifs, by which the feature vectors of the sequences in benchmark data were calculated. For each sequence and each motif profile, we scanned the sequences using a reading frame, having the same length of motif profile. The score of the best matching between a motif and substrings of the sequence was taken as a feature value for this sequence. And the feature values by 204 motifs composed the feature value of the sequence. Finally, we obtained a $322 * 204$ feature matrix on 322 sequences in benchmark data.

3.2 Classification of Features

We applied 12 classification tools on the feature matrix obtained in last section (Table 1): Nearest Neighbors, Logistic Regression, Bagging, Gradient Boosting, SGD, SVM, LinearSVC, Decision Tree, Random Forest, Extra Trees, AdaBoost, Naive Bayes. These classification methods performed well on benchmark data, while having little difference on various metrics. The highest Sn reached 93.17 percent by SVM, and the highest Accuracy, Precision, and recall reached 81.68 percent, 78.82 percent, and 93.17 percent, respectively. We took the average value of Sn, Sp, Acc, and MCC for all classification tools and compared with iPro54-PseKNC's performance in Fig. 2. We can see that the average performance of classification based on our strategy has higher Sn but lower Sp compared to iPro54-PseKNC, and the Acc and MCC are comparable between these two methods.

3.3 Dimension Reducing

We performed dimension reducing based on the above classification. Here, we chose three popular classification methods, bagging, random forest, and SVM to do further analysis. The most important parameters in the application of libSVM are the cost $(-\mathrm{c})$ and gamma $(-\mathrm{g})$, and we used default value 1 for cost and set gamma as $1/k$, where k is the number of records for input data. The parameters for random forest were set as following: the default value of the size of attribute set considered by each split (m_KValue) was used as $\log_2(m) + 1$, where *m* is the number of the input samples; the sample size of the training learner (batchSizes) was set to 100; and the number of iterations was set to 100.

TABLE 2
Performance after Dimension Reducing

Distance functions	Classification methods	Best dimensions	accuracy
Euclidean distance	bagging	192	85.87%
	rf	62	90.53%
	svm	44	80.43%
Cosine distance	bagging	30	85.71%
	rf	94	90.37%
	svm	40	80.75%
Tanimoto coefficient	bagging	27	86.18%
	rf	63	91.30%
	svm	43	80.59%
Mean distance	bagging	128	85.56%
	rf	91	91.61%
	svm	49	81.06%
Manhattan	bagging	110	85.56%
	rf	89	90.86%
	svm	38	80.12%
Canberra	bagging	50	85.09%
	rf	175	90.21%
	svm	38	81.05%
Bhattacharyya	bagging	189	86.42%
	rf	195	88.08%
	svm	195	79.14%
Modified Cosine	bagging	55	86.18%
	rf	80	90.68%
	svm	68	81.53%
Spearman	bagging	50	85.09%
	rf	175	90.21%
	svm	38	81.05%

rf: randome forest

The feature selection algorithm MRMD was performed with nine distance functions, Euclidean distance, cosine distance, Mean distance, Manhattan distance, Canberra distance, Bhattacharyya distance, Modified Cosine distance, and Spearman correlation coefficient.

We calculated the Acc for all combinations of distance function and classification tools, which are shown in Table 2. The results indicated that the classification accuracy was improved a lot and the dimension was significantly reduced, e.g., the bagging combined with Bhattacharyya distance improved the accuracy from 80.75 percent to 86.42 percent, and reduced dimension from 204 to 189. The improvement of accuracy made by SVM and modified cosine is from 80.75 percent to 81.53 percent and reduced to 68 in dimension. For random forest under the mean distance function, the accuracy reached 91.61 percent, which is comparable with the accuracy 93.79 percent obtained by iPro54-PseKNC after feature optimization.

3.4 Performance on Other Species

Since the experimentally confirmed promoter data is not enough for the test in other bacterial genomes, we applied the new method and iPro54-PseKNC on the computationally predicted data from three different genomes, NC_000964, NC_003030, and NC_008497, as described in the above

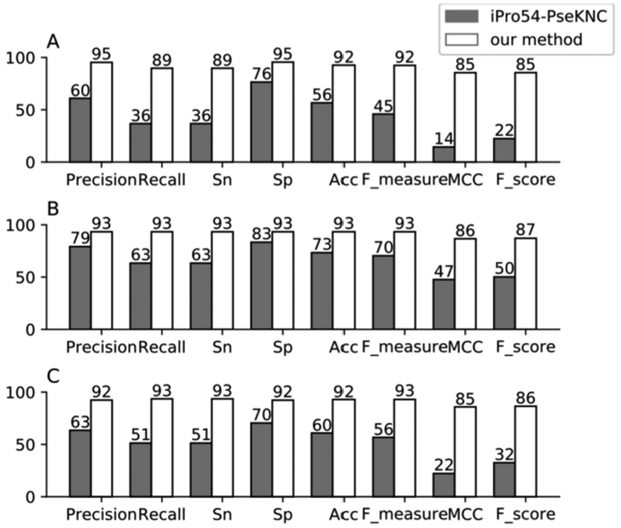


Fig. 3. The performance of our method and iPro54-PseKNC on the genomes NC_000964 (A), NC_003030 (B), and NC_008497 (C). The solid gray show the metrics of iPro54-PseKNC, and the white bars show the metrics of our method.

Method section. It worth noting that, since the positive set is also computational predicted $\sigma - 54$ promoters, the bias in evaluation on these data is inevitable. The good side is that these promoters are selected very carefully with strict criteria, which means the positive set (with 78, 30, and 68 predicted) may only be a small part of all $\sigma - 54$ promoters in these genome, but have relatively high confidence. In contrast, the negative datasets were generated randomly in a genome-scale by avoiding predicted promoters, thus have a little chance to randomly contain many unpredicted $\sigma - 54$ promoters. Therefore, the conformity between multiple methods can provide valuable information on their prediction performance. Here, we take the positive and negative data from three genomes as benchmark to evaluate the performance of our method and iPro54-PseKNC on several prediction metrics including Precision, Recall, Sn, Sp, Acc, MCC, F-measure, and F-score (Fig. 3). In the figure, we can see that the conformity between benchmark data with our method are much higher. Specifically, the Acc of our methods are 64 percent, 27 percent, and 53 percent higher than the ones of iPro54-PseKNC for the three genomes, respectively. The performance indicates that our methods, despite trained in *E. coli* dataset, may have robust performance on other bacterial genomes, even they belong to different phylum. Indeed, the performance of iPro54-PseKNC on this test data cannot represent its actual prediction capacity, since it may get the specific features neglected by test data and our method.

4 Conclusion and Discussion

In this study, we developed a new method to predict the $\sigma - 54$ promoters in bacterial genomes by organically integrating the motif finding and machine learning classification strategies. The method was trained on the *E. coli* promoters, while the application of it on additional biological data showed high robustness on a large range of bacterial genomes. It indicates that the new method proposed in this study may has better potential on capturing the sequence features of σ-factor promoters and avoiding the overfitting on a single-source dataset. The strategy of this method can also be used to train sequence features for other σ-factors. Certainly, taking computational predicted promoters as benchmark data may have bias on performance evaluation, and the high conformity between our method and the benchmark data may be partly caused by the similar intrinsic mechanism on feature capturing. Further analysis is needed in the future work to examine the prediction ability of our new method, especially on variant species of bacteria. In addition, we plan to apply this method to all the fully sequenced bacterial genomes ($> 6,500$ as of this manuscript is submitted in the NCBI database) and systematically predict all the σ-factor promoters along with their commonality and differences among various genomes. All the derived information will be integrated into the most well-developed and maintained operon database, DOOR2.0 [53], to benefit more researchers in the field of functional microbiology.

Acknowledgments

This work was supported by a National Science Foundation/EPSCoR Award (No. IIA-1355423), the State of South Dakota Research Innovation Center, the Scholarly Excellence Funds of South Dakota State University, and the Sanford Health – South Dakota State University Collaborative Research Seed Grant Program. This work was also supported by the National Nature Science Foundation of China (NSFC) [61772313, 61303084, and 61432010 to B. Liu]. B. Liu's work was also supported by the Young Scholars Program of Shandong University (YSpSDU, 2015WLJH19). J. Wu's work was supported by The China Scholarship Council. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

References

[1] Z. Wang, X. Sun, and Y. Zhao, et al., "Evolution of gene regulation during transcription and translation," *Genome Biol. Evol.*, vol. 7, no. 4, pp. 1155–1167, Apr. 14, 2015.

[2] V. Svetlov and I. Artsimovitch, "Purification of bacterial RNA polymerase: Tools and protocols," *Methods Mol. Biol.*, vol. 1276, pp. 13–29, 2015.

[3] J. D. Helmann and M. J. Chamberlin, "Structure and function of bacterial sigma factors," *Annu. Rev. Biochem.*, vol. 57, pp. 839–872, 1988.

[4] H. Barrios, B. Valderrama, and E. Morett, "Compilation and analysis of sigma(54)-dependent promoter sequences," *Nucleic Acids Res.*, vol. 27, no. 22, pp. 4305–4313, Nov. 15, 1999.

[5] U. K. Sharma and D. Chatterji, "Transcriptional switching in Escherichia coli during stress and starvation by modulation of sigma activity," *FEMS Microbiol. Rev.*, vol. 34, no. 5, pp. 646–657, Sep., 2010.

[6] A. M. Huerta and J. Collado-Vides, "Sigma70 promoters in *Escherichia coli*: Specific transcription in dense regions of overlapping promoter-like signals," *J. Mol. Biol.*, vol. 333, no. 2, pp. 261–278, Oct. 17, 2003.

[7] Q. Z. Li and H. Lin, "The recognition and prediction of sigma70 promoters in *Escherichia coli* K-12," *J. Theor. Biol.*, vol. 242, no. 1, pp. 135–141, Sep. 07, 2006.

[8] T. P. Hunt and B. Magasanik, "Transcription of glnA by purified *Escherichia coli* components: Core RNA polymerase and the products of glnF, glnG, and glnL," *Proc. Nat. Academy Sci. United States America*, vol. 82, no. 24, pp. 8453–8457, Dec., 1985.

[9] F. Touzain, S. Schbath, and I. Debled-Rennesson, et al., "SIGffRid: A tool to search for sigma factor binding sites in bacterial genomes using comparative approach and biologically driven statistics," *BMC Bioinf.*, vol. 9, pp. 73, Jan. 31, 2008.

[10] H. Lin, E. Z. Deng, and H. Ding, et al., “iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition,” *Nucleic Acids Res.*, vol. 42, no. 21, pp. 12961–12972, Dec. 01, 2014.

[11] C. A. Gross, C. Chan, and A. Dombroski, et al., “The functional and regulatory roles of sigma factors in transcription,” *Cold Spring Harb Symp. Quant. Biol.*, vol. 63, pp. 141–155, 1998.

[12] T. Zhou, N. Shen, and L. Yang, et al., “Quantitative modeling of transcription factor binding specificities using DNA shape,” *Proc. Nat. Academy Sci. United States America*, vol. 112, no. 15, pp. 4654–4659, Apr. 14, 2015.

[13] L. Yang, T. Zhou, and I. Dror, et al., “TFBSshape: A motif database for DNA shape features of transcription factor binding sites,” *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D148–D155, Jan., 2014.

[14] S. G. Carvalho, R. Guerra-Sa, and C. M. L. H. de, “The impact of sequence length and number of sequences on promoter prediction performance,” *BMC Bioinf.*, vol. 16 no. Suppl 19, 2015, Art. no. S5.

[15] A. Kanhere and M. Bansal, “A novel method for prokaryotic promoter prediction based on DNA stability,” *BMC Bioinf.*, vol. 6, Jan. 05, 2005, Art. no. 1.

[16] Y. Gan, J. Guan, and S. Zhou, “A comparison study on feature selection of DNA structural properties for promoter prediction,” *BMC Bioinf.*, vol. 13, Jan. 07, 2012, Art. no. 4.

[17] V. B. Bajic, M. R. Brent, and R. H. Brown, et al., “Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment,” *Genome Biol.*, vol. 7, no. Suppl 1, pp. S3 1–13, 2006.

[18] Y. C. Zuo and Q. Z. Li, “The hidden physical codes for modulating the prokaryotic transcription initiation,” *Physica a-Statistical Mech. Appl.*, vol. 389, no. 19, pp. 4217–4223, Oct. 1, 2010.

[19] H. Lin and Q. Z. Li, “Eukaryotic and prokaryotic promoter prediction using hybrid approach,” *Theory Biosci.*, vol. 130, no. 2, pp. 91–100, Jun., 2011.

[20] Q. Wu, J. Wang, and H. Yan, “An improved position weight matrix method based on an entropy measure for the recognition of prokaryotic promoters,” *Int. J. Data Min. Bioinf.*, vol. 5, no. 1, pp. 22–37, 2011.

[21] K. Song, “Recognition of prokaryotic promoters based on a novel variable-window Z-curve method,” *Nucleic Acids Res.*, vol. 40, no. 3, pp. 963–971, Feb., 2012.

[22] T. Abeel, Y. Saeys, and P. Rouze, et al., “ProSOM: Core promoter prediction based on unsupervised clustering of DNA physical profiles,” *Bioinf.*, vol. 24, no. 13, pp. i24–31, Jul. 01, 2008.

[23] T. Abeel, Y. Saeys, and E. Bonnet, et al., “Generic eukaryotic core promoter prediction using structural features of DNA,” *Genome Res.*, vol. 18, no. 2, pp. 310–23, Feb., 2008.

[24] K. C. Chou, “Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes,” *Bioinf.*, vol. 21, no. 1, pp. 10–19, Jan. 1, 2005.

[25] Y. K. Chen and K. B. Li, “Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou’s pseudo amino acid composition,” *J. Theor. Biol.*, vol. 318, pp. 1–12, Feb. 07, 2013.

[26] Z. Hajisharifi, M. Piryaiee, and M. Mohammad Beigi, et al., “Predicting anticancer peptides with Chou’s pseudo amino acid composition and investigating their mutagenicity via Ames test,” *J. Theor. Biol.*, vol. 341, pp. 34–40, Jan. 21, 2014.

[27] L. Nanni, S. Brahnam, and A. Lumini, “Prediction of protein structure classes by incorporating different protein descriptors into general Chou’s pseudo amino acid composition,” *J. Theor. Biol.*, vol. 360, pp. 109–16, Nov. 07, 2014.

[28] P. Du, S. Gu, and Y. Jiao, “PseAAC-General: fast building various modes of general form of Chou’s pseudo-amino acid composition for large-scale protein datasets,” *Int. J. Mol. Sci.*, vol. 15, no. 3, pp. 3495–3506, Feb. 26, 2014.

[29] P. Du, X. Wang, and C. Xu, et al., “PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou’s pseudo-amino acid compositions,” *Anal. Biochem.*, vol. 425, no. 2, pp. 117–119, Jun. 15, 2012.

[30] D. S. Cao, Q. S. Xu, and Y. Z. Liang, “Propy: A tool to generate various modes of Chou’s PseAAC,” *Bioinf.*, vol. 29, no. 7, pp. 960–962, Apr. 01, 2013.

[31] S. X. Lin and J. Lapointe, “Theoretical and experimental biology in one,” *J. Biomed. Sci. Eng.*, vol. 6, no. 4, pp. 435–442, 2013.

[32] B. Liu, J. Chen, and X. Wang, “Protein remote homology detection by combining Chou’s distance-pair pseudo amino acid composition and principal component analysis,” *Mol. Genetic Genomics*, vol. 290, no. 5, pp. 1919–1931, Oct., 2015.

[33] Q. Zou, J. Li, and Q. Hong, et al., “Prediction of MicroRNA-disease associations based on social network analysis methods,” *Biomed. Res. Int.*, vol. 2015, 2015, Art. no. 810514.

[34] Q. Zou, J. Guo, and Y. Ju, et al., “Improving tRNAscan-SE annotation results via ensemble classifiers,” *Mol. Inform.*, vol. 34, no. 11-12, pp. 761–770, Nov., 2015.

[35] H. Salgado, M. Peralta-Gil, and S. Gama-Castro, et al., “RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D203–D213, Jan, 2013.

[36] D. C. Stevens, K. R. Conway, and N. Pearce, et al., “Alternative sigma factor over-expression enables heterologous expression of a type II polyketide biosynthetic pathway in *Escherichia coli*,” *PLoS One*, vol. 8, no. 5, 2013, Art. no. e64858.

[37] M. M. Nakano and P. Zuber, “Anaerobic growth of a “strict aerobe” (Bacillus subtilis),” *Annu. Rev. Microbiology* vol. 52, pp. 165–190, 1998.

[38] C. J. Sund, S. Liu, and K. L. Germane, et al., “Phosphoketolase flux in clostridium acetobutylicum during growth on L-arabinose,” *Microbiology*, vol. 161, no. Pt 2, pp. 430–440, Feb., 2015.

[39] C. C. Kern, R. F. Vogel, and J. Behr, “Differentiation of lactobacillus brevis strains using matrix-assisted-laser-desorption-ionization-time-of-flight mass spectrometry with respect to their beer spoilage potential,” *Food Microbiol.*, vol. 40, pp. 18–24, Jun., 2014.

[40] D. J. Studholme, M. Buck, and B. T. Nixon, “Identification of potential sigma(N)-dependent promoters in bacterial genomes,” *Microbiology-UK*, vol. 146, pp. 3021–3023, Dec., 2000.

[41] L. Rajeev, E. G. Luning, and P. S. Dehal, et al., “Systematic mapping of two component response regulators to gene targets in a model sulfate reducing bacterium,” *Genome Biol.*, vol. 12, no. 10, Oct. 12, 2011, Art. no. R99.

[42] K. M. Giglio, N. Caberoy, and G. Suen, et al., “A cascade of coregulating enhancer binding proteins initiates and propagates a multicellular developmental program,” *Proc. Nat. Academy Sci. United States America*, vol. 108, no. 32, pp. E431–E439, Aug. 09, 2011.

[43] T. Ueki and D. R. Lovley, “Novel regulatory cascades controlling expression of nitrogen-fixation genes in Geobacter sulfurreducens,” *Nucleic Acids Res.*, vol. 38, no. 21, pp. 7485–7499, Nov., 2010.

[44] K. Zhao, M. Liu, and R. R. Burgess, “Promoter and regulon analysis of nitrogen assimilation factor, sigma54, reveal alternative strategy for *E. coli* MG1655 flagellar biosynthesis,” *Nucleic Acids Res.*, vol. 38, no. 4, pp. 1273–1283, Mar., 2010.

[45] T. Ueki and D. R. Lovley, “Genome-wide gene regulation of biosynthesis and energy generation by a novel transcriptional repressor in Geobacter species,” *Nucleic Acids Res.*, vol. 38, no. 3, pp. 810–821, Jan., 2010.

[46] Q. Ma, H. Zhang, and X. Mao, et al., “DMINDA: An integrated web server for DNA motif identification and analyses,” *Nucleic Acids Res.*, vol. 42, no. Web Server issue, pp. W12–W19, Jul., 2014.

[47] J. Yang, X. Chen, and A. McDermaid, et al., “DMINDA 2.0: Integrated and systematic views of regulatory DNA motif identification and analyses,” *Bioinf.*, vol. 33, no. 16, pp. 2586–2588, Aug. 15, 2017.

[48] G. Li, B. Liu, and Q. Ma, et al., “A new framework for identifying cis-regulatory motifs in prokaryotes,” *Nucleic Acids Res.*, vol. 39, no. 7, Apr, 2011, Art. no. e42.

[49] G. Li, B. Liu, and Y. Xu, “Accurate recognition of cis-regulatory motifs with the correct lengths in prokaryotic genomes,” *Nucleic Acids Res.*, vol. 38, no. 2, Jan., 2010, Art. no. e12.

[50] Q. Ma, B. Liu, and C. Zhou, et al., “An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale,” *Bioinf.*, vol. 29, no. 18, pp. 2261–2268, Sep. 15, 2013.

[51] B. Liu, H. Zhang, and C. Zhou, et al., “An integrative and applicable phylogenetic footprinting framework for cis-regulatory motifs identification in prokaryotic genomes,” *BMC Genomics*, vol. 17, Aug 09, 2016, Art. no. 578.

[52] B. Liu, C. Zhou, and G. Li, et al., “Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses,” *Sci. Rep.*, vol. 6, Mar. 15, 2016, Art. no. 23030.

[53] X. Mao, Q. Ma, C. Zhou, et al., "DOOR 2.0: Presenting operons and their functions through dynamic and integrated views," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. 654–659, 2014.

[54] A. Shinde, A. Sahu, and D. Apley, et al., "Preimages for variation patterns from kernel PCA and bagging," *Iie Trans.*, vol. 46, no. 5, pp. 429–456, May 1, 2014.

[55] L. Tolosi and T. Lengauer, "Classification with correlated features: Unreliability of feature ranking and solutions," *Bioinf.*, vol. 27, no. 14, pp. 1986–1994, Jul. 15, 2011.

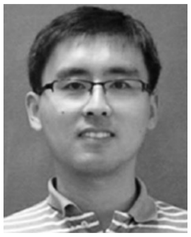
[56] S. Q. Wang, J. Yang, and K. C. Chou, "Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition," *J. Theoretical Biol.*, vol. 242, no. 4, pp. 941–946, Oct. 21, 2006.

[57] Y. D. Cai, G. P. Zhou, and K. C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical J.*, vol. 84, no. 5, pp. 3257–3263, May, 2003.

[58] E. Math and S. Davis, *Statistical Genomics: Methods and Protocols*, New York, NY, USA: Humana Press, 2016.

[59] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011.

[60] Q. Zou, J. C. Zeng, and L. J. Cao, et al., "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, Jan. 15, 2016.

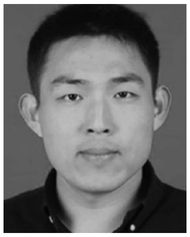


Bingqiang Liu received the doctoral degree in mathematics from Shandong University, in 2010. He is currently an associate professor with the School of Mathematics, Shandong University. His main research interests include bioinformatics algorithm design, combinatorial optimization, and graph theory.

Ling Han is currently working toward the master's degree in mathematics applied at Bioinformatics. She received the bachelor degree in mathematics from Qufu Normal University, in 2014. She works on the development of mathematics and computer methods for computational problems in gene regulation.

Xiangrong Liu received the PhD degree from the Department of Control Science and Engineering, Huazhong University of Science and Technology, in 2007. He is currently a professor with the Department of Computer Science, Xiamen University. His research interests include bioinformatics and bio-inspire computing.

Jichang Wu received the first doctoral degree in mathematics from Northwestern Polytechnical University, in 2003, and the second doctoral degree from the University of Twente, in 2009. He is currently an associate professor with the School of Mathematics, Shandong University. His main research interest includes graph theory.



Qin Ma received the doctoral degree in operational research from Shandong University, in 2010. He is currently an associate professor with the Department of Mathematics & Statistics and the Department of Agronomy, Hoticulture & Plant Sciences, South Dakota State University. His main research interests include bioinformatics and computational systems biology.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.