Journal of Bioinformatics and Computational Biology
Vol. 16, No. 1 (2018) 1840003 (16 pages)
© World Scientific Publishing Europe Ltd.
DOI: 10.1142/S0219720018400036

# Bacterial promoter prediction: Selection of dynamic and static physical properties of DNA for reliable sequence classification

Artem Ryasik[*,§], Mikhail Orlov[*,¶], Evgenia Zykova[*,†,‖],
Timofei Ermak[‡,**] and Anatoly Sorokin[*,††]

[*] *Mechanism of Cell Genome Functioning Laboratory*
*Institute of Cell Biophysics, ul. Institutskaya 3*
*Pushchino 142290, Russia*

[†] *Department of Applied Research Informatization*
*State Institute of Information Technologies*
*and Telecommunications (SIIT&T Informika)*
*per. Brusov 21 st.2, Moscow, 125009, Russia*

[‡] *Laboratory of Molecular Genetics Systems*
*Institute of Cytology and Genetics*
*pr. Akademika Lavrentyeva 10*
*Novosibirsk 630090, Russia*
[§] *ryasik.aa@gmail.com*
[¶] *orlovmikhailanat@gmail.com*
[‖] *evgenia.teml@gmail.com*
[**] *timofei.ermak@live.com*
[††] *lptolik@gmail.com*

Received 1 August 2017
Revised 15 November 2017
Accepted 8 December 2017
Published 31 January 2018

Predicting promoter activity of DNA fragment is an important task for computational biology. Approaches using physical properties of DNA to predict bacterial promoters have recently gained a lot of attention. To select an adequate set of physical properties for training a classifier, various characteristics of DNA molecule should be taken into consideration. Here, we present a systematic approach that allows us to select less correlated properties for classification by means of both correlation and cophenetic coefficients as well as concordance matrices. To prove this concept, we have developed the first classifier that uses not only sequence and static physical properties of DNA fragment, but also dynamic properties of DNA open states. Therefore, the best performing models with accuracy values up to 90% for all types of sequences were obtained. Furthermore, we have demonstrated that the classifier can serve as a reliable tool enabling promoter DNA fragments to be distinguished from promoter islands despite the similarity of their nucleotide sequences.

*Keywords*: Machine learning; promoter recognition; DNA physical properties.

## 1. Introduction

Despite almost 40 years of efforts, our understanding of processes involved in bacterial transcription initiation is still far from complete. As the number of experimentally identified promoters grows, it becomes evident that the analysis of nucleotide composition is insufficient to establish location and activity of bacterial promoters. Accordingly, there is a growing interest in identifying other characteristics that are common to promoter regions and could be applied for promoter search.[1–5] The most frequently analyzed features include distribution of physical parameters around DNA molecule,[4–7] patterns in nucleotide composition as well as combinations of motives[8,9] and free energy values.[10] Although all of them are entirely encoded by DNA primary structure, their distributions contain valuable information that cannot be replaced with the analysis of nucleotide composition (text analysis).[4]

Molecular recognition of DNA by DNA-binding proteins occurs at the level of DNA physical properties and complementarity between the protein and the binding site surface. Some of these physical properties correlate well with the nucleotide composition of DNA fragment. For example, geometry of DNA grooves is known to correlate with GC-content.[11] On the contrary, several physical characteristics including the distribution of electrostatic potential demonstrate identical profiles that could be obtained from a broad range of sequences.[7] Furthermore, minor variations in the sequence are known to dramatically alter the physical property profile of the whole DNA molecule.[12]

Each of those characteristics by itself could not explain functional properties and predict promoter location. Accordingly, several DNA features (textual as well as physical) should be included in the whole-genome promoter search. Numerous studies of regulatory DNA regions in prokaryotic[1,5,6,13,14] and eukaryotic[15–17] genomes have proven the efficiency of analysis employing several characteristics distributed throughout regulatory DNA regions. Among them, the most notable physical characteristics are: electrostatic potential,[18–20] stress-induced duplex destabilization,[12] bendability[21]; and textual characteristics: $z$-curve,[22] CG-skew,[23] triplets, tetramers, pentamers and hexamers.[24]

A DNA molecule can be described by a set of different physical properties. Most of these characteristics are static including electrostatic potential,[18] free energy,[25] propensity to bend,[26] stress-induced duplex destabilization,[4] etc. All of them are widely used for regulatory sites' classification and prediction. These characteristics are usually easy to calculate and generally sufficient to perform the task.

In addition to static characteristics of a DNA molecule, several dynamic ones are described including activation energy and size of DNA open states as well as DNA sound velocity.[27,28] These characteristics are affected by local context and consider changes in time. In contrast to static physical properties, the dynamic ones are less studied and used as their application requires complex partial differential equations to be solved in order to characterize a DNA molecule.

Taking dynamic characteristics into consideration might provide valuable information that would help to investigate complex and multi-step transcription initiation process. DNA open states were demonstrated to move along the DNA chain during transcription initiation.[29] Moreover, promoter regions usually have lower values of activation energy of DNA open states (i.e. energy needed for the open state to appear) than other DNA regions.[27,30] We can therefore assume that promoter regions have less 'energy cost' to evolve a DNA open state which in turn could play a significant role in transition of the polymerase-promoter complex from closed to open form and consequently obtain successful transcription and RNA synthesis.

A new method described in Refs. 27,30,31 enabled not only the dynamic properties of DNA open states to be calculated but also their distribution throughout the whole genome to be estimated without solving nonlinear partial differential equations. In the current work, we have combined both static and dynamic physical properties of DNA in an attempt to develop more robust classification models for the whole-genome promoter search.

## 2. Materials and Methods

We used K-12 MG1655 of *E. coli* (GenBank identifier U00096.2). All annotated and experimentally validated 699 promoters and 3427 genes locations were taken from RegulonDB version 8.5.[32] Three sets of non-promoter DNA sequences were prepared as follows:

(i) *Non-promoter chromosome regions* (1880 sequences) are located at least 300 bps away from any known or suggested transcription start sites (TSS) and have no assigned genomic function.
(ii) *Lowscore sequences* (2000 sequences) are the least similar to promoters regions of genome by nucleotide composition according to Platprom software analysis.[8]
(iii) *Promoter islands* (2228 sequences) are genome regions that contain densely packed hypothetical TSSs and are predicted by Platprom software.[8,33] Although most of promoter islands could be recognized by RNA-polymerase, no experimentally validated productive transcriptional activity was shown for the sequences. Promoter islands were first described in Ref. 34 but their function still remains unclear.[33,35,36]

For each sequence, we have calculated profiles of the following characteristics: electrostatic potential distribution along the DNA chain ($EP$),[18] DNA open states activation energy ($E0$) and size ($d$), DNA sound velocity ($C$), as well as GC-content ($GC$).[27]

All characteristics applied were examined in the interval of 200 bps, with the exception of the electrostatic potential with 720 Å interval used. We have calculated a full profile of each characteristic for the whole chromosome for convenience sake. Later on, the interval [−150; +50] bps with respect to TSS (or pseudo-TSS for the non-promoter sequence) was taken for each sequence. Calculating distribution profiles for electrostatic potential was made in the same way, except that we used

720 Å interval [−540; 180] Å around TSS and pseudo-TSS, respectively. To treat all types of DNA sequences in the same way, in case of non-promoter sequences pseudo-TSSs were chosen as a point at 151 bps from the left boundary.

The electrostatic profile was calculated with algorithm described in Ref. 18 by R the package *reldna*.[37] The GC-content and dynamical properties of DNA such as DNA open states activation energy, their size, and DNA sound velocity were calculated using the DNA open states dynamics model.[27] For this purpose, the algorithm implemented by the authors in Matlab/Octave (available on request) was applied. All analysis was performed in the R statistical environment (version 3.2.2),[38] principal component analysis (PCA) and visualization of results were prepared with FactoMineR (`http://factominer.free.fr/`) and factoextra (`http://www.sthda.com/english/rpkgs/factoextra/`) libraries for R. Dendrograms were analyzed with dendextend library (`https://cran.r-project.org/web/packages/dendextend/`). Classification models training was performed by means of caret (`https://github.com/topepo/caret/`) library for R.

## 3. Results

DNA physical characteristics could be divided into three classes according to the peculiarities of their distribution along the DNA helix axis:

(i) *Global smooth* are affected by the sequence changes remote from the selected point. They are characterized by continuous profiles within this range. For example, electrostatic potential profile[18] is a global smooth feature.

(ii) *Global spike* are influenced by the long interval of DNA from the selected point, yet they have profiles with local sharp spikes. The stress induced duplex destabilization (SIDD)[12] profile is an example of this type of features.

(iii) *Local* are influenced by the sequence on short distances from the point of consideration and have smooth profiles. Dynamic characteristics of DNA open states are the examples of local feature.

Since more than half of annotated promoters do not have significant peaks in their SIDD profiles, global spike characteristics including SIDD have been excluded from further analysis. So that we have considered global smooth, dynamical, and textual characteristics (the latter two are local).

In most cases, we do not need a full set of characteristics as they might have inner interrelations. So for further classifiers training, we have selected variables by a two-step procedure. We initially evaluated the suitability of each chosen physical property and later on solely the least correlated properties were utilized. Stages of the research described above are shown in Fig. 1 as a brief flow chart.

### 3.1. *Preliminary data analysis*

The presence of correlated features is known to cause low classification performance. In this paper, clusterization of DNA physical and textual properties was used as a

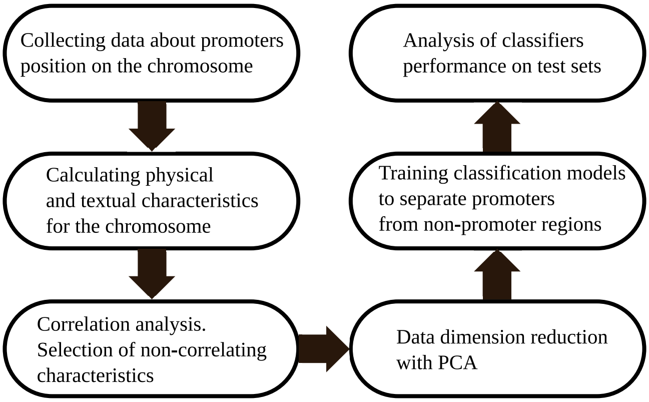
Collecting data about promoters position on the chromosome
Calculating physical and textual characteristics for the chromosome
Correlation analysis. Selection of non-correlating characteristics
Data dimension reduction with PCA
Training classification models to separate promoters from non-promoter regions
Analysis of classifiers performance on test sets

Fig. 1. Bacterial promoter prediction flowchart shows the methods applied in the research.

preliminary step prior to applying supervised machine learning algorithms. Our aim here was to assess the correlation of considered properties for several types of DNA sequences. Pearson's correlation coefficient cannot be used for comparison of characteristics of several diverse sequence types and to evaluate the respective clustering results. Clusterization yielded several sets of clusters corresponding to each physical property used. The similarity among the sets can be measured, i.e. it is possible to make an assumption about the similarity among features themselves using hierarchical clusterization results. If the sets are similar, dendrograms ought to be similar as well; the latter could be evaluated using cophenetic correlation coefficient and concordance coefficient analysis. Several dendrograms demonstrated high similarity; therefore, analyzing all of them would be excessive. On the contrary, dendrograms that differ significantly appear to be more informative, which implies low correlation of the corresponding properties. Since the most distinct characteristics would allow to describe DNA physics more efficiently, they were selected for further analysis.

The problem of comparing partitions has long been studied; Sokal and Rohlf[39] introduced the first efficient numerical method known as the 'cophenetic correlation' technique. The analysis of dendrograms is an approach allowing to assess the similarity among a large number of partitions. In statistics, particularly in biostatistics, cophenetic correlation coefficient serves as a measure of how precisely a dendrogram preserves the pairwise distances between the original unmodeled data points. It has been widely applied in biostatistics (e.g. taxonomic models) and also might be used for processing raw data that tend to occur in clusters. Additionally, this technique has been proposed to be used as a test for nested clusters.[40]

First, Ward clusterization[41] was performed for each characteristic profile corresponding to promoter sequences. Second, the obtained dendrograms were studied using cophenetic coefficients[42] (Table 1). As it can be seen from Table 1, most of the non-diagonal elements of the table are equal to zero. The $p$-values were in the range of $10^{-319}$–$10^{-3}$. We obtained a slight correlation between two pairs of values: DNA open state activation energy ($E0$), its size ($d$), and DNA sound velocity ($C$).

Table 1. Table of cophenetic correlations coefficients for dendrograms. The dendrograms represent the results of hierarchical clusterization of $E0, d, C, EP$ and $GC$. Low values of the coefficients prove that selected characteristics are not correlated.

| | $E0$ | $d$ | $C$ | $EP$ | $GC$ |
|---|---|---|---|---|---|
| $E0$ | 1 | 0.04 | 0.09 | 0 | 0 |
| $d$ | 0.04 | 1 | 0.01 | 0 | 0 |
| $C$ | 0.09 | 0.01 | 1 | 0.01 | 0 |
| $EP$ | 0 | 0 | 0.01 | 1 | 0 |
| $GC$ | 0 | 0 | 0 | 0 | 1 |

One possible explanation for the correlation lies in the fact that these characteristics are obtained using the same model of DNA open states dynamics.[27,30]

The dendrograms were also compared using tanglegrams,[43] i.e. diagrams where two dendrograms are plotted together side by side and points corresponding to profiles for the same sequence are joined by a straight line. Tanglegrams for $E0$ dendrogram compared to $C$ and $GC$ dendrograms correspondingly are shown as examples (Fig. 2). On the left figure, one can see that characteristics such as activation energy of open states and sound velocity demonstrate significant similarity since a portion of connecting lines is joined in bundles, and several common subtrees are also present. On the other hand, we cannot observe the same behavior on the right figure in case of $E0$ and $GC$ dendrograms. For this reason, $C$ was excluded from the final set of characteristics. The same tanglegram analysis was performed for all pairs of chosen characteristics and no correlation was obtained.

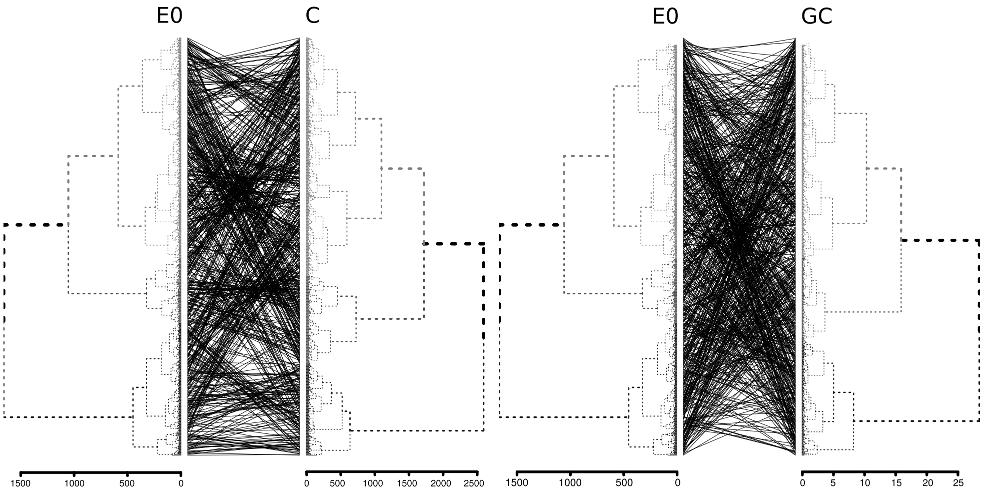

Fig. 2. Left panel: Tanglegrams for DNA open states activation energy (E0) and DNA sound velocity (C) clusterizations. As lines between two dendograms are joined into bundles, it is clear that studied clusterizations are correlated. Right panel: Tanglegrams for DNA open states activation energy (E0) and GC-content (GC) clusterizations. The figure illustrates barely correlated clusterizations.

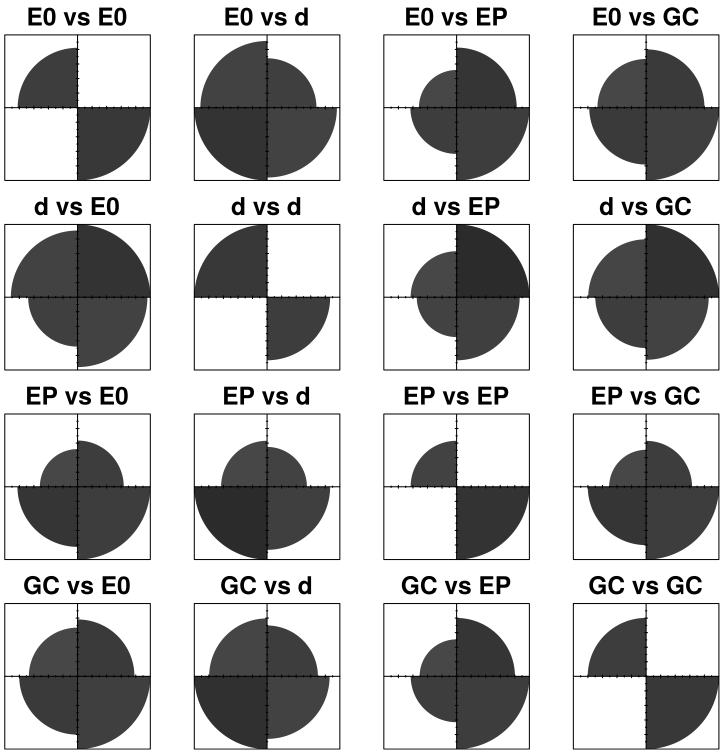

Fig. 3. Concordance matrices for DNA open states activation energy ($E0$), DNA open state size ($d$), electrostatic potential ($EP$), GC-content ($GC$). Most of the elements were placed in non-diagonal cells, which means they were put into different clusters. This fact proves that dendrograms have low concordance degree and considered characteristics are uncorrelated.

Finally, to ensure the independence of the characteristics, concordance matrices of the dendrograms were analyzed[44] (Fig. 3). A concordance matrix shows the number of elements classified as belonging or not belonging to the same cluster with respect to datasets (diagonal segments). It separates the elements of the different datasets by placing them into different clusters (non-diagonal segments). As it can be seen from circle diagrams, most of the elements in non-diagonal cells were put into different clusters, which suggests that dendrograms have low concordance degree.

These tests for correlation of characteristics allowed us to exclude sound velocity $C$ from consideration and thus reduce the amount of possible errors for classifiers, that will be described in the next section. Furthermore, each of the remaining characteristics was shown to be completely different. Although these characteristics (GC-content in our case) are completely defined by DNA sequence, they cannot be thoroughly examined solely by means of text analysis. Moreover, all characteristics barely correlate with each other, which indicates that each of the selected characteristic could be valuable for describing differences in investigated DNA sequence types and distinguishing these sequences.

### 3.2. *Classifiers training*

On the next step, the dimensionality of the dataset was reduced using principal component analysis. As the scree plot demonstrates (Fig. 4), the first 50 principal components describe more than 97% of the cumulative variance. For further machine learning, sets of 50, 100 and 150 first-principal components were taken. Therefore, the size of the new dataset was reduced by 9–26 times.

Each class of non-promoter sequences could differ from promoters in its distinct way; consequently, we have trained four different classifiers to distinguish promoters from genes, promoter islands, low-score sequences and non-promoter sequences, separately. Thus, we split the resulting dataset into four subsets for further application of machine learning: promoters–genes (G–P), promoters–promoter islands (I–P), promoters–lowscore sequences (L–P) and promoters–non-promoters sequences (N–P). We applied Naive Bayes and Random Forest machine learning algorithms to train classifiers for each of the four subsets. In each case, the training set was taken randomly and contained an equal number of sequences of two types and included 70%, 80% and 90% of the dataset, accordingly. Resampling iterations were performed 10 times and the number of cross-validation set resampling was equal to 3. The training models were validated on the testing samples that contained the rest 30%, 20% and 10% of the original set of two types of sequences respectively. The training and the testing sets of sequences did not intersect. Finally, for each possible dataset, learning algorithm, size of training set and the number of principal components, 10 classifier models were trained. Therefore, the overall number of trained models was equal to 720.

The classifiers performance was estimated by the following properties: sensitivity — number of correctly defined promoters, specificity — number of correctly defined sequences that are not promoters, and accuracy (Fig. 5). The Random Forest

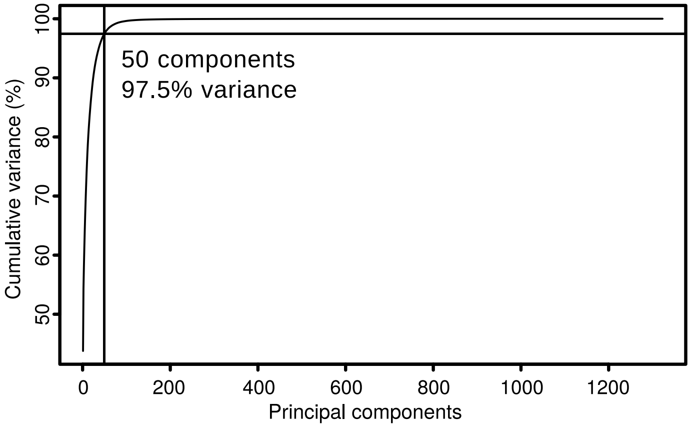

Fig. 4. Scree plot showing cumulative variance percentage as the function of principal components. First 50 principal components contain more than 97% of the cumulative variance.

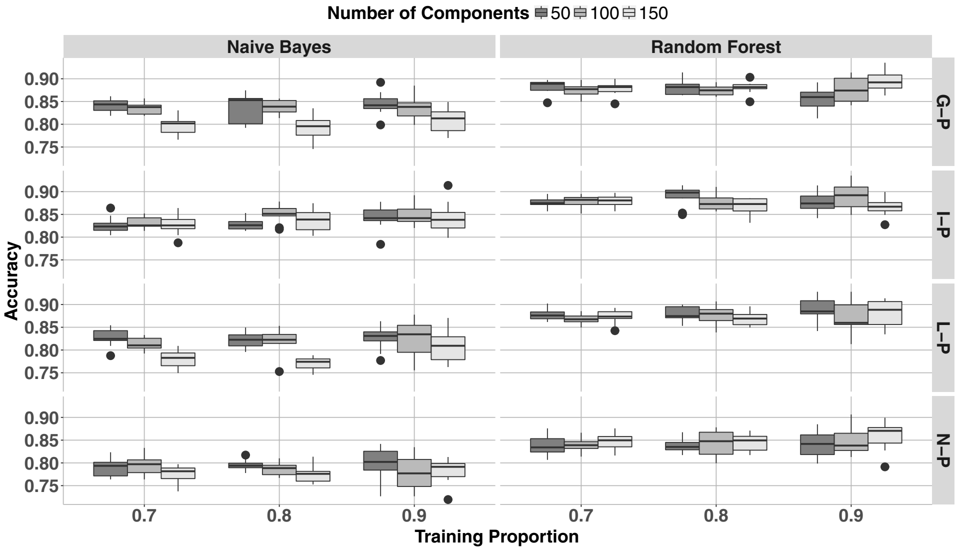

Fig. 5. Mean accuracy values for classification models of promoters versus genes (G–P), promoters versus promoter islands (I–P), promoters versus lowscore sequences (L–P), promoter versus non-promoter sequences (N–P) trained with Naive Bayes (left) and Random Forest (right). The outliers are represented by black dots.

Table 2. The best accuracy, sensitivity and specificity properties of classification models.

| Model | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| Naive bayes | | | |
| Promoters versus Genes | 89% | 93% | 86% |
| Promoters versus Promoter islands | 91% | 96% | 87% |
| Promoters versus Lowscore sequences | 88% | 97% | 78% |
| Promoters versus Non-promoters regions | 84% | 87% | 81% |
| Random forest | | | |
| Promoters versus Genes | 94% | 94% | 93% |
| Promoters versus Promoter islands | 94% | 99% | 88% |
| Promoters versus Lowscore sequences | 93% | 94% | 91% |
| Promoters versus Non-promoters regions | 91% | 89% | 93% |

classification models outperformed Naive Bayes models by accuracy, and, as the Table 2 shows, the best Random Forest accuracy values are 91–94% compared with only 84–91% accuracy in case of Naive Bayes models.

## 4. Discussion

Identifying the precise promoter location on the bacterial chromosome is a challenging task. Each case requires analysis of upstream and downstream TSS regions due to chances that there are multiple overlapping promoters, promoters partly located in genes, etc.[8] For this reason, we have decided that one classifier is not sufficient for promoter search and have trained four classifiers capable of distinguishing a promoter from its context sequence at various locations on the chromosome.

Another long-standing problem in the promoter search is selecting features significant for promoter recognition by RNA-polymerase. Several approaches have been tested to work around this issue, starting from basic similarity of consensus hexanucleotides[9] and length of spacer,[45] to more sophisticated sequence analysis[46] and finally to various physical properties of promoter DNA.[1,2,7,10,12,14,15,18,20,47–49] Most previous works considered only static properties of DNA molecules: electrostatic profile,[7,18,20,47,48] thermodynamical stability,[4,12,17] geometry of double helix,[5,6,49] presence of certain chemical groups,[49] elastic properties,[50,51] etc. We have also applied dynamic properties of the DNA open states for the analysis of promoter sequences. The importance of DNA open states could not be underestimated as the formation of a transcription bubble is the key step in promoter activation process. The recent development of fast algorithms[27] allows us to calculate the DNA open states activation energy, its velocity, size and other characteristics. Therefore, we are able to calculate distributions of all these characteristics for the whole chromosome and include the data into our analysis.

All features considered in this paper are calculated for the same DNA sequence of the *E. coli* chromosome. The initial dataset of physical properties requires thorough feature selection before applying any further algorithm. Eliminating correlated noisy features is quite important for better classification accuracy. This task is not obvious, as we work with physical properties distributions around DNA molecule. The distribution profiles generally have different length and variance that causes certain difficulties in the analysis. Here, we employed the fact that profiles vary significantly for different sequences, and each physical characteristic emphasizes various aspects of the interaction between DNA and RNA-polymerase.

We built hierarchical clustering for each of the physical properties of promoter sequences separately and assessed the similarity among dendrograms with several techniques. We defined redundant characteristics as the ones producing similar dendrograms. The similarity between the dendrograms indicates that two features influence the classification results in a similar way, so one of them has to be excluded from consideration.

Comparison of dendrograms using cophenetic correlation coefficient demonstrated minimum similarity among all selected DNA characteristics which in most cases have values less than 0.05. Small cophenetic coefficient values imply that all studied properties are barely correlated and could be considered together to describe regulatory DNA regions (e.g. promoters) and predict their location. The next correlation test was tanglegram analysis that showed similar groups of promoters to be located in the same clusters in case of dendrograms based on different characteristics – $E0$ and $C$. Accordingly, we excluded DNA sound velocity $C$ from further consideration. The final test consisted in analyzing concordance matrices. The study revealed that clusterization based on the remaining characteristics has low values of concordance. All these tests emphasized that each of the selected characteristics

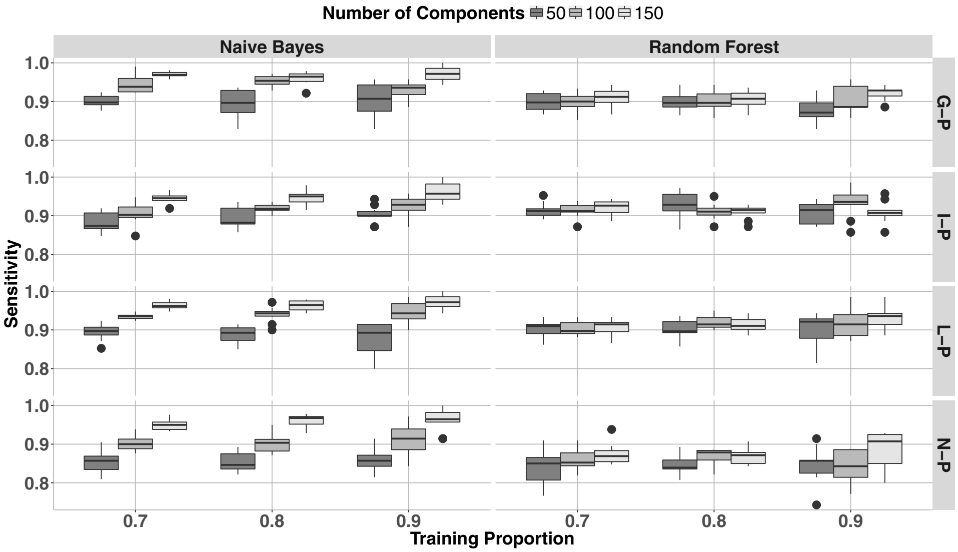

Fig. 6. Mean sensitivity values for classification models of promoters and genes (G–P), promoters and promoter islands (I–P), promoters and lowscore sequences (L–P), promoter versus non-promoter sequences (N–P) trained with Naive Bayes (left) and Random Forest (right). The outliers are represented by black dots.

contains specific information that could be applied for distinguishing between different types of DNA sequences.

On the next step, vectors containing selected profiles were combined into a matrix where each row represents a DNA sequence. After that, the dimensionality of the matrix was reduced by means of PCA. Based on the result of PCA, we formed three samples containing 50, 100, and 150 principal components. These samples were used for classifier training with Naive Bayes and Random Forest algorithms.

As Fig. 6 demonstrates for Naive Bayes classification models, the sensitivity tends to grow with the number of principal components involved in training increases. The opposite trend is observed for the specificity of the studied models. Random Forest classification models appear to be more stable in respect to variation of both parameters. Figs. 5 and 7 show that both specificity and accuracy values for Random Forest exceed the ones for Naive Bayes models. At the same time, sensitivity values (Fig. 6) values are almost equal for both algorithms. As a result, the number of correctly classified promoter sequences is higher for Random Forest models than that for Naive Bayes.

Noticeably, the high accuracy values were obtained for promoter–promoter islands classifier (Table 2). Promoter islands are identified by the Platprom as regions that contain multiple possible TSSs within 300 bps interval.[33,35,36] Here, we have proven that promoter islands and promoters could be distinguished by considering physical properties of DNA. This observation requires further interpretation, which could lead to better understanding of physical mechanisms of promoter function.

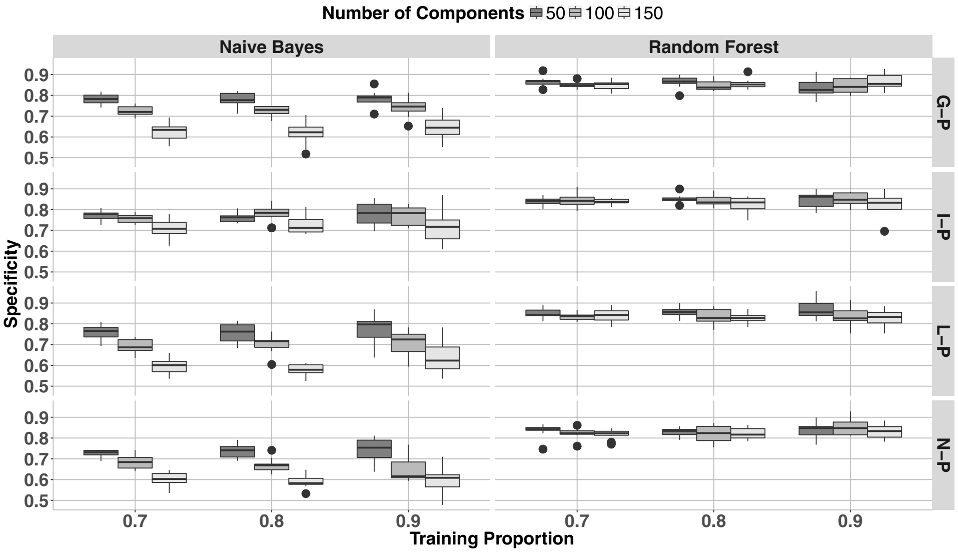

Fig. 7. Mean specificity values for classification models of promoters and genes (G–P), promoters and promoter islands (I–P), promoters and lowscore sequences (L–P), promoter versus non-promoter sequences (N–P) trained with Naive Bayes (left) and Random Forest (right). The outliers are represented by black dots.

It is also worth comparing the parameters of our classifiers with other promoter prediction software such as recently developed CNNProm based on text analysis only.[52] The authors declared 90% sensitivity and 96% specificity for the algorithm applied for *E. coli* promoters search. The parameters of our classifiers yield similar values — 89–94% sensitivity and 88–93% specificity. So, we can conclude that using DNA physical properties, static and dynamic, is a reliable and promising approach for prediction of bacterial promoters.

The source code of the algorithms used in this study is available on the link `https://github.com/FVortex/DNAClassifiers`.

## Acknowledgments

The research was supported by RFBR, research project 16-37-00303 mol_a.

## References

1. Abeel T, Saeys Y, Rouzé P, Van de Peer Y, Prosom: Core promoter prediction based on unsupervised clustering of DNA physical profiles, *Bioinform* **24**(13):i24–i31, 2008.
2. Djordjevic M, Integrating sequence analysis with biophysical modelling for accurate transcription start site prediction, *J Integ Bioinform* **11**(2):240, 2014.
3. Yeramian E, Genes and the physics of the DNA double-helix, *Gene* **255**(2):139–150, 2000.
4. Wang H, Benham CJ, Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress, *BMC Bioinform* **7**(1):248, 2006.

5. Duran E, Djebali S, Gonzalez S, Flores O, Mercader JM, Guigo R, Torrents D, Soler-Lopez M, Orozco M, Unravelling the hidden DNA structural/physical code provides novel insights on promoter location, *Nucleic Acids Res* **41**(15):7220–7230, 2013.
6. Meysman P, Dang TH, Laukens K, De Smet R, Wu Y, Marchal K, Engelen K, Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli, *Nucl Acids Res* **39**(2):e6, 2011.
7. Sorokin AA, Osypov AA, Dzhelyadin TR, Beskaravainy PM, Kamzolova SG, Electrostatic properties of promoter recognized by E. coli RNA polymerase Esigma70, *J Bioinform Comput Biol* **4**(2):455–467, 2006.
8. Shavkunov KS, Masulis IS, Tutukina MN, Deev AA, Ozoline ON, Gains and unexpected lessons from genome-scale promoter mapping, *Nucleic Acids Res* **37**(15):4919–4931, 2009.
9. Jacques PE, Rodrigue S, Gaudreau L, Goulet J, Brzezinski R, Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs, *BMC Bioinform* **7**:423, 2006.
10. Kanhere A, Bansal M, A novel method for prokaryotic promoter prediction based on DNA stability, *BMC Bioinform* **6**:1, 2005.
11. Parker SCJ, Tullius TD, DNA shape, genetic codes, and evolution, *Curr Opin Struct Biol* **21**(3):342–347, 2011.
12. Benham CJ, Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions, *J Mol Biol* **255**(3):425–434, 1996.
13. Mallios RR, Ojcius DM, Ardell DH, An iterative strategy combining biophysical criteria and duration hidden markov models for structural predictions of chlamydia trachomatis $\sigma$ 66 promoters, *BMC Bioinform* **10**(1):271, 2009.
14. Conilione P, Wang D, A comparative study on feature selection for E. coli promoter recognition, *Int J Inf Technol* **11**:54–66, 2005.
15. Goñi JR, Pérez A, Torrents D, Orozco M, Determining promoter location based on DNA structure first-principles calculations, *Genome Biol* **8**(12):R263, 2007.
16. Gan Y, Guan J, Zhou S, A pattern-based nearest neighbor search approach for promoter prediction using DNA structural profiles, *Bioinform* **25**(16):2006–2012, 2009.
17. Morey C, Mookherjee S, Rajasekaran G, Bansal M, DNA free energy-based promoter prediction and comparative analysis of arabidopsis and rice genomes, *Plant Physiol* **156**(3):1300–1315, 2011.
18. Polozov RV, Dzhelyadin TR, Sorokin AA, Ivanova NN, Sivozhelezov VS, Kamzolova SG, Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences, *J Biomol Struct Dyn* **16**(6):1135–1143, 1999.
19. Kamzolova SG, Sivozhelezov VS, Sorokin AA, Dzhelyadin TR, Ivanova NN, Polozov RV, RNA polymerasepromoter recognition. Specific features of electrostatic potential of early t4 phage DNA promoters, *J Biomol Struct Dyn* **18**(3):325–334, 2000.
20. Temlyakova EA, Dzhelyadin TR, Kamzolova SG, Sorokin AA, 70 electrostatic properties of bacterial DNA and promoter predictions, *J Biomol Struct Dyn* **31**(suppl1):44–45, 2013.
21. Ozoline ON, Deev AA, Trifonov EN, DNA bendability; a novel feature in E. coli promoter recognition, *J Biomol Struct Dyn* **16**(4):825–831, 1999.
22. Song K, Recognition of prokaryotic promoters based on a novel variable-window z-curve method, *Nucleic Acids Res* **40**(3):963–971, 2011.
23. Tatarinova T, Brover V, Troukhan M, Alexandrov N, Skew in cg content near the transcription start site in arabidopsis thaliana, *Bioinform* **19**(suppl_1):i313–i314, 2003.
24. Sobha Rani T, Bapi RS, Analysis of n-gram based promoter recognition methods and application to whole genome promoter prediction, *In Silico Biol* **9**(1, 2):S1–S16, 2009.
25. Rangannan V, Bansal M, High-quality annotation of promoter regions for 913 bacterial genomes, *Bioinform* **26**(24):3043–3050, 2010.

26. Brukner I, Sanchez R, Suck D, Pongor S, Sequence-dependent bending propensity of DNA as revealed by dnase i: Parameters for trinucleotides, *EMBO J* **14**(8):1812, 1995.
27. Grinevich AA, Ryasik AA, Yakushevich LV, Trajectories of DNA bubbles, *Chaos Solitons Fractals* **75**:62–75, 2015.
28. Dauxois T, Peyrard M, Bishop AR, Dynamics and thermodynamics of a nonlinear model for DNA denaturation, *Phys Rev E* **47**(1):684, 1993.
29. von Hippel PH, From "simple" DNA-protein interactions to the macromolecular machines of gene expression, *Ann Rev Biophys Biomol Struct* **36**:79–105, 2007.
30. Yakushevich LV, Ryasik AA, Dynamics of kinks in inhomogeneous polynucleotide chains, *Biophys* **58**(4):439, 2013.
31. Yakushevich LV, Grinevich AA, Ryasik AA, Simulation of a kink movement in homogeneous and heterogeneous DNA sequences taking into account the dissipation, *Russian J Numer Anal Math Model* **29**(3):197–204, 2014.
32. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, *et al.*, Regulondb v8. 0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more, *Nucleic Acids Res* **41**(D1):D203–D213, 2012.
33. Shavkunov KS, Tutukina MN, Masulis IS, Ozoline ON, Promoter islands: The novel elements in bacterial genomes, *J Biomol Struct Dyn* **28**(6):1128–1129, 2011.
34. Tutukina MN, Shavkunov KS, Masulis IS, Ozoline ON, Intragenic promotor-like sites in the genome of escherichia coli discovery and functional implication, *J Bioinf Comput Biol* **5**(02b):549–560, 2007.
35. Panyukov VV, Ozoline ON, Promoters of escherichia coli versus promoter islands: Function and structure comparison, *PLoS One* **8**(5):e62601, 2013.
36. Purtov YA, Glazunova OA, Antipov SS, Pokusaeva VO, Fesenko EE, Preobrazhenskaya EV, Shavkunov KS, Tutukina MN, Lukyanov VI, Ozoline ON, Promoter islands as a platform for interaction with nucleoid proteins and transcription factors, *J Bioinform Comput Biol* **12**(2):1441006, 2014.
37. Reldna package. `https://github.com/promodel/reldna`. Accessed: 2017-07-19.
38. R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017.
39. Sokal RR, Rohlf FJ, The comparison of dendrograms by objective methods, *Taxon* **11**(2):33–40, 1962.
40. Rohlf FJ, Fisher DR, Tests for hierarchical structure in random data sets, *Syst Biol* **17**(4):407–412, 1968.
41. Ward JH Jr. Hierarchical grouping to optimize an objective function, *J Am Stat Assoc* **58**(301):236–244, 1963.
42. Farris JS, On the cophenetic correlation coefficient, *Syst Zoology* **18**(3):279–285, 1969.
43. Scornavacca C, Zickmann F, Huson DH, Tanglegrams for rooted phylogenetic trees and networks, *Bioinform* **27**(13):i248–i256, 2011.
44. Lawrence I, Lin K, A concordance correlation coefficient to evaluate reproducibility, *Biometrics* **45**(1):255–268, 1989.
45. Mulligan ME, Brosius J, McClure WR, Characterization in vitro of the effect of spacer length on the activity of Escherichia coli RNA polymerase at the TAC promoter, *J Biol Chem* **260**(6):3529–3538, 1985.
46. Kiselev SS, Ozoline ON, Structure-specific modules as indicators of promoter DNA in bacterial genomes, *Mat Biol Bioinform* **6**(1):39–52, 2011.
47. Kamzolova SG, Sorokin AA, Dzhelyadin TR, Beskaravainy PM, Osypov AA, Electrostatic potentials of E. coli genome DNA, *J Biomol Struct Dyn* **23**(3):341–345, 2005.

48. Kamzolova SG, Osipov AA, Beskaravaĭnyĭ PM, Dzheliadin TR, Sorokin AA, Regulation of promoter activity through electrostatic interactions with RNA-polymerase, *Biofizika* **52**(2):228–236, 2007.
49. Bauer AL, Hlavacek WS, Unkefer PJ, Mu F, Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites, *PLoS Computat Biol* **6**(11):e1001007, 2010.
50. Lankas F, Sponer J, Hobza P, Langowski J, Sequence-dependent elastic properties of DNA, *J Mol Biol* **299**(3):695–709, 2000.
51. Pedone F, Mazzei F, Santoni D, Sequence-dependent DNA torsional rigidity: A tetranucleotide code, *Biophys Chem* **112**(1):77–88, 2004.
52. Umarov RKh, Solovyev VV, Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks, *PloS One* **12**(2):e0171410, 2017.

**Artem Ryasik** graduated in theoretical physics from Siberian Federal University (Krasnoyarsk, Russia), in 2011. Then, he received PhD in Biophysics from Institute of Cell Biophysics RAS (Pushchino, Russia) in 2016. Currently, he is working as a research fellow in Institute of Cell Biophysics in the field of machine learning, big data and graph analysis applied to biology.

**Mikhail Orlov** received the BS and MSc from the Lomonosov Moscow State University. Since 2015, he is a PhD student and junior research fellow with the Institute of Cell Biophysics RAS. Mr. Orlov's main areas of expertize are biophysics, bioinformatics and machine learning. His research interests include RNA promoter recognition and DNA physics.

**Evgenia Zykova** received her specialist's degree in Biochemical Physics from Siberian Federal University (Krasnoyarsk, Russia) in 2011, and she received her PhD degree in Biophysics from Institute of Cell Biophysics RAS (Pushchino, Russia) in 2016. She is a research fellow in State Institute of Information Technologies and Telecommunications (SIIT&T "Informika") and Institute of Cell Biophysics RAS. Her research interests include machine learning and data mining techniques in application to omics data, genome and metabolic regulatory networks.

**Timofei Ermak**, MSc, is a PhD student in the Department of Systems Biology, Institute of Cytology and Genetics (Novosibirsk, Russia). He was graduated from Siberian State University of Telecommunications and Information Sciences in 2013. His research interests include metabolic networks reconstruction and analysis, structural bioinformatics, high performance computing.

**Anatoly Sorokin** received his MSc in Physics from Moscow Institute of Physics and Technology in 1995 and his PhD in Biophysics from Institute of Theoretical and Experimental Biophysics RAS in 2001. From 2005 till 2011, he was a research fellow in School of Informatics of The University of Edinburgh. Since 2013, he is the Head of Mechanisms of Cell Genome Functioning Laboratory in the Institute of Cell Biophysics RAS. His research interests include large-scale modeling in systems biology, whole genome modeling, and knowledge representation. He is a member of iSCB.