doi:10.1093/nar/gkr795

Recognition of prokaryotic promoters based on a novel variable-window Z-curve method

Kai Song[1,2,*]

[1]School of Chemical Engineering and Technology and [2]Institute of Life Science and Biotechnology, Tianjin University, Tianjin, 300072, China

Received June 27, 2011; Revised September 5, 2011; Accepted September 11, 2011

ABSTRACT

Transcription is the first step in gene expression, and it is the step at which most of the regulation of expression occurs. Although sequenced prokaryotic genomes provide a wealth of information, transcriptional regulatory networks are still poorly understood using the available genomic information, largely because accurate prediction of promoters is difficult. To improve promoter recognition performance, a novel variable-window Z-curve method is developed to extract general features of prokaryotic promoters. The features are used for further classification by the partial least squares technique. To verify the prediction performance, the proposed method is applied to predict promoter fragments of two representative prokaryotic model organisms (*Escherichia coli* and *Bacillus subtilis*). Depending on the feature extraction and selection power of the proposed method, the promoter prediction accuracies are improved markedly over most existing approaches: for *E. coli*, the accuracies are 96.05% (σ^{70} promoters, coding negative samples), 90.44% (σ^{70} promoters, non-coding negative samples), 92.13% (known sigma-factor promoters, coding negative samples), 92.50% (known sigma-factor promoters, non-coding negative samples), respectively; for *B. subtilis*, the accuracies are 95.83% (known sigma-factor promoters, coding negative samples) and 99.09% (known sigma-factor promoters, non-coding negative samples). Additionally, being a linear technique, the computational simplicity of the proposed method makes it easy to run in a matter of minutes on ordinary personal computers or even laptops. More importantly, there is no need to optimize parameters, so it is very practical for predicting other species promoters without any prior knowledge or prior information of the statistical properties of the samples.

INTRODUCTION

In genetics, a promoter is a region of DNA that facilitates the transcription of particular genes. In bacteria, the promoter is recognized by RNA polymerase (RNAP) and associated sigma factors, which may be recruited to the promoter by regulatory proteins binding to specific sites in the region. Thus, control of transcription initiation accounts for much of the overall regulation of gene expression (1). The continued development of large, sophisticated databases and repositories has made vast amounts of biological data accessible to researchers. Additionally, advances in molecular biology and computational techniques are enabling the systematic investigation of the complex molecular processes underlying biological systems. Many algorithms have been developed for the detection of promoters in prokaryotic genomes. For example, Askary *et al.* and Rangannan and Bansal developed a promoter prediction algorithm based on the difference in stability between neighbouring upstream and downstream regions in the vicinity of experimentally determined transcription start sites (TSSs) (2,3). Rani and Bapi used *n*-grams ($n = 3$) as features for a neural network classifier for promoter prediction in *Escherichia coli* and achieved 67.75% prediction sensitivity and 86.10% specificity (4). Mann *et al.* used a hybrid technique combining profile hidden Markov models (HMMs) and artificial neural networks (ANNs) methods with Viterbi scoring optimizations (5). Burden *et al.* and Bland *et al.* also used ANNs to improve the promoter prediction accuracy (6,7). Lin and Li developed a hybrid approach (called IPMD) combining position correlation score function and increment of diversity with modified Mahalanobis Discriminant to predict eukaryotic and prokaryotic promoters (8). By applying the IPMD to *E. coli*

[*]To whom correspondence should be addressed. Tel: +86 138 201 86019; Email: ksong@tju.edu.cn

and *Bacillus subtilis* promoter sequences, they achieved the sensitivities and specificities of 84.9% and 91.4% for *E. coli*, as well as 80.4% and 91.3% for *B. subtilis*.

Although these attempts, which employ sophisticated machine-learning methods to identify promoters, offer increased accuracy in certain circumstances, the improvements may not justify the heavy computational requirements they impose for training classifiers. Moreover, the selection and optimization of parameters (such as the type and parameters of kernel functions, number of hidden layer nodes, etc.) need enough prior knowledge of the statistical properties of the samples, which makes it unpractical for the analysis of new genome sequences.

The regular Z-curve (or Z-curve) method originally proposed by Zhang is a powerful tool in visualizing and analysing DNA sequences (9,10). It is a 3D curve or point representation for a DNA sequence in the sense that each can be uniquely reconstructed given the other. The resulting curve has a zigzag shape, hence the name Z-curve. The 3D curve or point of a given DNA sequence is calculated from the frequencies of the four bases occurring in it to evaluate the sequence from three main components: distribution of purine/pyrimidine, distribution of amino/keto and distribution of strong H-bonds/weak H-bonds. Z-curve method has been used in many different areas of genome research, such as replication origin identification (11,12), *ab initio* gene prediction (13), isochore identification (14), genomic island identification (15) and comparative genomics (16). However, the regular Z-curve method could not able to extract the information of *w*-nucleotides sequence patterns occurring in DNA sequences, the promoter recognition accuracy based on it is far from satisfactory.

Hence, a novel variable-window Z-curve (vw Z-curve) method is proposed here as a feature-extraction tool for prokaryotic promoter recognition for the first time. The features extracted by it (with window size $w = 1, 2, \ldots, 6$) are used as the input variables for further classification by a partial least squares (PLS) classifier. Promoter fragments of two prokaryotic model organisms (*E. coli* and *B. subtilis*) are used to verify the prediction performance of the proposed method. The feature extraction power of the vw Z-curve method and the iterative feature selection power of the PLS technique make the prediction performance improved markedly over most existing approaches: for *E. coli*, the accuracies are 96.05% (σ^{70} promoters, coding negative samples), 90.44% (σ^{70} promoters, non-coding negative samples), 92.13% (known sigma-factor promoters, coding negative samples), 92.50% (known sigma-factor promoters, non-coding negative samples), respectively; for *B. subtilis*, the accuracies are 95.83% (known sigma-factor promoters, coding negative samples) and 99.09% (known sigma-factor promoters, non-coding negative samples). The results are verified relying on a 10-fold cross-validation jackknife test. Moreover, the proposed method is a linear technique, thus its computational simplicity makes it possible to be run on ordinary personal computers or laptops with run times of several minutes. In particular, because there is no need to optimize parameters, this method is very practical for predicting other species promoters without any prior knowledge or prior information of the statistical properties of the samples.

MATERIALS AND METHODS

Databases

The complete genomic sequences of *E. coli* K-12 and *B. subtilis* are obtained from NCBI GenBank (17). The positions of experimentally determined TSSs of them are retrieved from RegulonDB version 7.0 (18) and DBTBS (19). Then promoter regions [TSS-60 . . . TSS+19] (the site of TSS is +1) are taken as the positive examples. The positive sample database of *E. coli* consists of two kinds of promoter fragments: 576 experimentally confirmed σ^{70} promoters and 825 experimentally confirmed promoters of several known sigma factors (576 σ^{70} promoters, 63 σ^{38} promoters, 40 σ^{38} and σ^{70} promoters, 64 σ^{24} promoters, 4 σ^{24} and σ^{70} promoters, 9 σ^{28} promoters, 44 σ^{32} promoters, 7 σ^{32} and σ^{70} promoters, 18 σ^{54} promoters). Considering the comparatively small size of experimentally confirmed *B. subtilis* promoters, all 660 promoters of known sigma factors (e.g. σ^{43}, σ^{54}, σ^{37} and so on) are used as the positive samples of *B. subtilis*.

As there is no enough experimentally confirmed negative data (i.e. the positions that are confirmed not to be TSS), the risk has to be taken to choose the negative examples randomly from the same chromosome. Approximately, for *E. coli* K-12, 81% of known TSSs are located in the intergenic non-coding regions and 19% in the coding regions (20). So two kinds of negative examples are prepared:

(i) Coding negative examples: fragments extracted from the coding regions (genes). For *E. coli*, the coding negative sample set contains 836 80-bp fragments extracted from the start of the open reading frames (ORFs) with lengths of 80–380 bp. For *B. subtilis*, the coding negative sample set contains 665 80-bp fragments extracted from the start of the ORFs with lengths of 80–335 bp.

(ii) Non-coding negative examples: fragments extracted from the non-coding regions (convergent intergenic spacers). For *E. coli*, the non-coding negative sample set contains 825 fragments with lengths of 80 bp. For *B. subtilis*, the non-coding negative sample set contains 331 fragments with lengths of 80 bp.

The data sets and the corresponding detailed descriptions are shown in Table 1.

The novel variable-window Z-curve feature extraction method

Being the first transcription step, initiation promoted by interaction of RNAP with gene promoter is a key level of control of gene expression. RNAP holoenzyme is recruited at a given promoter through the recognition of a promoter by transcriptional factors, called 'sigma (σ) factors', which are variable subunit of RNAP holoenzyme.

Typically, housekeeping σ^{70} factors of *E. coli* bind to the −35 and −10 DNA sequence elements in a promoter

Table 1. The detailed descriptions of data sets

Data set	Positive samples	Negative samples
Data set-1	576 σ^{70} promoters of *E. coli*	836 coding fragments of *E. coli*
Data set-2	576 σ^{70} promoters of *E. coli*	825 non-coding fragments of *E. coli*
Data set-3	825 known sigma-factor promoters of *E. coli*	836 coding fragments of *E. coli*
Data set-4	825 known sigma-factor promoters of *E. coli*	825 non-coding fragments of *E. coli*
Data set-5	660 known sigma-factor promoters of *B. subtilis*	665 coding fragments of *B. subtilis*
Data set-6	660 known sigma-factor promoters of *B. subtilis*	331 non-coding fragments of *B. subtilis*

with the consensus sequences TTGACA at position −35 and TATAAT at position −10, respectively (positions indicate the location of each sequence with respect to the TSS). Two other important sites are the extended −10 element with the consensus sequence 'TGN' and the AT-rich UP element (21,22). Alternatively, σ^{54} factors, which control several ancillary processes including the degradation of xylene and toluene, transport of dicarboxylic acids and so on, bind to 'GG' at −24 location and 'GC' at −12 location of promoters (23). For *B. subtilis*, DegU promoter has the 'GNCATTTA' consensus DNA-binding sequence (24), σ^{E}-independent *sig*G promoters have 'TTT' and 'AAA' motifs (25) and so on.

It is well known that different sigma factors bind to different motifs of promoters. One genome may encode many different σ-factors. In general, bacterial housekeeping sigma-factors, which regulate genes that are involved in cellular growth, σ-factors are similar to the *E. coli* σ^{70} factors (26,27). Several members of the σ^{70} factor family have been described: *E. coli* K-12 has six σ^{70} family σ-factors (28), whereas *B. subtilis* has 17 known variants of σ^{70} (19). A specific subfamily of σ-factors that directly incorporates signals from the extracellular environment in regulating transcription (ECF σ-factors) also exists (29). More details about promoter architecture and sigma factors are available in the Supplementary Data.

Mismatches between RNAP, σ-factors and the given binding sites can be tolerated and even allow for the modulation of promoter strength at some specific genes. Multiple occurrences of promoters in the same regulatory region of one gene can be found for different regulatory functions (30). Unless mutagenesis is performed, each site has the chance to be the place chosen by the RNAP to bind the DNA. Unlike eukaryotic promoters, tightly packed prokaryotic genes and promoters frequently overlap each other (18) obscuring promoter motifs.

Experimental procedures are efficient to identify individual promoters but not conceivable for sets of genes at the whole genome scale. This motivated the search for computational methods based on the knowledge gained about the properties of known promoters or based on an efficient representation of DNA motifs by means of combinatorial or stochastic methods. Unfortunately, the absence of relatively strong sequence patterns identifying true promoters, the diversity of the motifs, the comparatively uncertainty of the locations of the motifs and the incompletely understood mechanisms of the regulation of promoters confound exact predictions of prokaryotic promoters.

The aims of this work are not only to predict promoters with very high accuracy, but to predict promoters of different sigma factors that have different recognition motifs in one collective data set (Tables 1 and 3). So it is important to draw out these distinctions with different sigma factors whose motifs usually comprise more than 1 nt. While the regular Z-curve parameters are only derived from the frequencies of mononucleotides occurring in a DNA sequence. Consequently, the features extracted by regular Z-curve method are not enough for promoter recognition problems and the promoter prediction accuracy based on these features is far from satisfactory. Up to now, only Yang *et al.* (31) used Z-curve method in Human Pol II promoter recognition.

According to the key motifs mentioned above, it is reasonable to assume that the parameters derived from the distributions of *w*-nucleotides patterns (the window size $w \in \mathbf{N}$) are the essential features which could able to distinguish between promoter regions and non-promoter regions successfully. Hence, a novel variable-window Z-curve (vw Z-curve) method which introduces variable window technique into the regular Z-curve method is developed and used in prokaryotic promoter recognition for the first time to improve the prediction accuracy of the issue. The following paragraphs provide a detailed explanation of the methodology of the vw Z-curve method.

Let Word is a set consisting the 4 nt A, G, C and T, that is: Word = {A,G,C,T}, S_w^i (the window size $w \in \mathbf{N}$, $i = 1, \ldots, 4^w$) is a string constructed by picking w elements from the set Word with order and repetition. For example: when $w = 2$, $S_2^1 =$ 'AA', $S_2^2 =$ 'AT', $S_2^3 =$ 'AG', …, $S_2^{16} =$ 'TT'.

Let the frequency of sequence pattern '$S_{w-1}^i X$' occurring in an ORF or a fragment of DNA sequence be denoted by $p(S_{w-1}^i X)$, where $X =$ A, C, G and T. Using the Z-curve method of DNA sequences (32,33), the uniform definition of vw Z-curve variables (the window size $w \in \mathbf{N}$) could be deduced as Equation (1)

$$\begin{cases} x_{S_w^i} = [p(S_{w-1}^i\mathrm{A}) + p(S_{w-1}^i\mathrm{G})] - [p(S_{w-1}^i\mathrm{C}) + p(S_{w-1}^i\mathrm{T})] \\ y_{S_w^i} = [p(S_{w-1}^i\mathrm{A}) + p(S_{w-1}^i\mathrm{C})] - [p(S_{w-1}^i\mathrm{G}) + p(S_{w-1}^i\mathrm{T})] \\ z_{S_w^i} = [p(S_{w-1}^i\mathrm{A}) + p(S_{w-1}^i\mathrm{T})] - [p(S_{w-1}^i\mathrm{C}) + p(S_{w-1}^i\mathrm{G})] \\ \qquad w \in N \quad i = 1, 2, \ldots, 4^{w-1} \end{cases} \tag{1}$$

It can be easily seen that the mono-nucleotide, di-nucleotides and tri-nucleotides phase-independent Z-curve parameters illustrated by Gao and Zhang (32) are the special instances of the vw Z-curve method where $w = 1, 2, 3$.

The detailed descriptions of them are shown in Equations (2–4), respectively.

(1) The Z curve parameters for frequencies of phase-independent mononucleotides (window size $w = 1$, variable number $n = 3 \times 4^0 = 3$): the frequencies of bases A, C, G and T occurring in a DNA sequence are denoted by *a, c, g* and *t*, respectively. Based on the Z-curve method, *a*, *c*, *g* and *t* are mapped onto a point ***P*** in a 3D space ***V***, which are denoted by *x*, *y*, *z* (33).

$$\begin{cases} x = (a+g) - (c+t) \\ y = (a+c) - (g+t) \\ z = (a+t) - (c+g) \\ \quad x,y,z \in [-1,+1], \end{cases} \qquad \textbf{(2)}$$

(2) The Z curve parameters for frequencies of phase-independent di-nucleotides (window size $w = 2$, variable number $n = 3 \times 4^1 = 12$): let the frequency of di-nucleotides XY be denoted by $p(XY)$, where X, $Y = \mathrm{A}$, C, G and T. Using the Z-curve method of DNA sequences, the following equation could be deduced as:

$$\rangle \begin{cases} x_X = [p(X\mathrm{A})+p(X\mathrm{G})] - [p(X\mathrm{C})+p(X\mathrm{T})] \\ y_X = [p(X\mathrm{A})+p(X\mathrm{C})] - [p(X\mathrm{G})+p(X\mathrm{T})] \\ z_X = [p(X\mathrm{A})+p(X\mathrm{T})] - [p(X\mathrm{C})+p(X\mathrm{G})] \\ \qquad\qquad X = \mathrm{A,C,G,T} \end{cases} \qquad \textbf{(3)}$$

where x_X, y_X and z_X are the coordinates of a point in a 3D space.

(3) The Z curve parameters for frequencies of phase-independent tri-nucleotides (window size $w = 3$, variable number $n = 3 \times 4^2 = 48$): using similar notations, it could be deduced as:

$$\begin{cases} x_{XY} = [p(XY\mathrm{A})+p(XY\mathrm{G})] - [p(XY\mathrm{C})+p(XY\mathrm{T})] \\ y_{XY} = [p(XY\mathrm{A})+p(XY\mathrm{C})] - [p(XY\mathrm{G})+p(XY\mathrm{T})] \\ z_{XY} = [p(XY\mathrm{A})+p(XY\mathrm{T})] - [p(XY\mathrm{C})+p(XY\mathrm{G})] \\ \qquad X = \mathrm{A,C,G,T}, \quad Y = \mathrm{A,C,G,T}, \end{cases} \qquad \textbf{(4)}$$

By the same way, the vw Z-curve parameters for frequencies of *w*-nucleotides could be deduced easily. By a selective combination of *n* variables or parameters derived from the vw Z-curve method, a DNA sequence can be represented by a point or a vector in an *n*-dimensional space ***V***.

Unlike the variables extracted by Position Weight Matrix (PWM) based algorithms (30), vw Z-curve parameters are derived from the distributions of *w*-nucleotides patterns occurring in the same sequence fragment not from their frequencies occurring in different sequence fragments. Thus, the vw Z-curve parameters are not influenced by the uncertainty of motif positions relative to the TSS. Due to the introduce of the window size *w*, the distributions of *w*-nucleotides patterns according to different sigma factors could be taken into account synchronously. Consequently, this novel vw Z-curve method is especially suitable for solving motif-finding or pattern recognition (PR) problems of DNA sequence researching.

Considering both the length of those widely known motifs and the computational requirement, the window sizes of the proposed vw Z-curve method used for promoter recognition problems are set $w = 1, 2, \ldots, 6$. The detailed descriptions of them are shown in Supplementary Table S1.

For researchers' convenience, the MATLAB codes of the vw Z-curve method are given in the Supplementary Data.

Partial least squares classifier

Supervised pattern analysis could be taken as the regression problems in which the dependent variables are defined as $l\in\{-1,+1\}$ in two-class problems or as $l\in\{1, 2, \ldots, N\}$ in multi-class problems, here N is the number of classes. Hence regression algorithms could be used as classifiers in supervised PR.

PLS algorithm is a key technique for modelling linear relationships between a set of output variables (known class-labels) and a set of input variables (predictors). PLS algorithm creates orthogonal latent variables (LVs), which are linear combinations of the original variables. The basic point of the procedure is that the weights used to determine these linear combinations of the original variables are proportional to the maximum covariance among input and output variables (34). Hence, by the projection of the PLS algorithm, the *n*-dimensional ***X***-space is compressed into the *v*-dimensional LV-space ($v<<n$ in common cases) to remove the noise and the multi-colinearity of the original data. This leads to a biased but lower variance estimate of the regression coefficients compared to the least squares method (34). PLS has been proven to be very useful in situations where the number of observed variables (*n*) is significantly greater than the number of observations (*m*) and high multi-colinearity among the variables exists (35,36). This is especially true in the case of the current study. Thus, PLS is expected to be a useful supervised PR method with potential applications in the discovery of key vw Z-curve features.

For more detailed mathematical descriptions of the PLS classifier, please refer to the Supplementary Data.

The performance of promoter prediction

To evaluate the performance of promoter prediction, the following measurements are used here.

$$\text{Sensitivity}: S_n = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad \textbf{(5)}$$

$$\text{Specificity}: S_p = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad \textbf{(6)}$$

$$\text{Accuracy}: a = \frac{S_n+S_p}{2} \qquad \textbf{(7)}$$

where TP, TN, FP and FN are fractions of positive correct, negative correct, false positive and false negative predictions, respectively.

The sensitivity *Sn* is the proportion of promoter sequences that have been correctly predicted as promoters. The specificity *Sp* is the proportion of negative samples that have been correctly predicted as negative samples. The accuracy *a* is defined as the average of *Sn* and *Sp*. Thus the goal in this study is to maximize the prediction accuracy *a* of testing set as well as make good trade-off between *Sn* and *Sp*.

To overcome the randomicity of samples and to evaluate the prediction performance forcefully, 10-fold leave-one-out method is performed as the cross-validation jack-knife test, in which the data set is divided into 10 parts and tested on the 10 different one-tenths while trained on the remaining nine-tenths.

RESULTS AND DISCUSSION

Optimum combination of the vw Z-curve features for promoter prediction using the PLS based iterative feature selection method

Apart from feature extraction, feature selection (also known as variable selection) is one of the most useful techniques for improving the performance of PR. By removing the most irrelevant and redundant features, feature selection helps to do the following:

- Alleviate the effect of the curse of dimensionality.
- Enhance generalizability.
- Speed up the learning process.
- Improve model interpretability.

To select an optimal set of features, it is necessary to quantitatively evaluate the contribution of each feature of the vw Z-curve method. As mentioned above, two-class supervised pattern analysis can be handled as a univariate regression problem in which the dependent variables are defined as $l \in \{-1,+1\}$. For univariate regression problems, the absolute value of the regression coefficient of each variable is a reasonable measurement of its contribution.

The total number of all vw Z-curve variables of the issue is $n = \sum_{w=1}^{6} 3 \times 4^{w-1} = 4095$ ($w = 1, \ldots, 6$). The absolute values of the coefficients of all 4095 variables in the PLS promoter recognition model of data set-1 and data set-5 are shown in Supplementary Figure S3. The figure shows that only a few variables stand out above the others with high absolute coefficient values. Obviously, important information gets buried in a sea of trivialities, a phenomenon known as 'information saturation'. Moreover, the methodology of the vw Z-curve method indicates there are strong multi-collinear relationships among all these features. Hence, the feature selection method relying on the PLS algorithm is used to improve the recognition performance.

Considering the large number of variables compared to the number of samples, iterative feature selection is used as a way to improve the recognition performance. The detailed procedure is as follows (take the iterative feature selection of data set-1 for an example):

(1) Selecting *m* positive and *m* negative samples (here, $m = 576$, the number of positive samples).
(2) Using *n* vw Z-curve variables for training the promoter recognition model (for first feature selection iteration $n = 4095$).
(3) Sorting variables in descending order according to their absolute coefficient values.
(4) Selecting the top *p* variables (e.g. $p = 600$) with the highest absolute coefficient values and use of a cross-validation procedure to assess the prediction performance of these selected variables.
(5) Optimizing *p* to maximize the prediction average accuracy and to ensure a good trade-off between sensitivity (*Sn*) and specificity (*Sp*) of the recognition model.
(6) Repeating steps 2–6 (setting $n = p$) until the prediction average accuracy converges.

It is notable that, for different sample sets, the iteration number of the feature selection procedure and the optimal combination of vw Z-curve variables may be different.

The prediction results shown in Tables 2–4 demonstrate that the promoter recognition performance strongly lowered by information saturation and multi-collinearity is remarkably improved by the iterative feature selection method.

From the results shown in Table 2, it can be seen that after the first iteration of feature selection, the number of useful vw Z-curve variables is markedly reduced from 4095 variables to 600 variables. By eliminating the interference of irrelevant variables, the prediction accuracy of data set-1 is improved by 15.27%, and the accuracy of

Table 2. Prediction results of the σ^{70} promoters of *E. coli* using different combination of vw Z-curve features**

Number*	Data set-1				Data set-2				
	4095	600	350	330	4095	600	500	245	220
Results (%)									
$\bar{S}_n$	80.00	92.63	95.79	**96.32**	81.40	88.42	87.02	91.40	**92.11**
$\bar{S}_p$	75.96	93.86	95.61	**95.79**	75.44	84.74	85.09	86.14	**88.77**
$\bar{a}$	77.98	93.25	95.70	**96.05**	78.42	86.58	86.05	88.77	**90.44**

*Number: number of selected vw Z-curve variables.
**The average accuracies of the vw Z-curve methods with 330 parameters for Data set-1and 220 parameters for Dataset-2, which were the best ones among the algorithms evaluated here, were shown in boldface.

data set-2 is improved by 8.16%. These 600 variables are used to build recognition models to again re-evaluate their importance clearly and fairly. To further improve the prediction accuracy, features are selected according to their re-evaluated importance. The prediction accuracy is improved until no more useless variables could be eliminated. After three iterations of feature selection, the final prediction accuracy of data set-1 reaches as high as 96.05%, much better than the accuracy obtained with any previously developed method. The final accuracy of data set-2 is only 90.44%, but it is much better than the accuracy achieved by most other methods. Furthermore, the trade-off between the *Sn* and *Sp* is also improved by the feature selection procedure: the difference between *Sn* and *Sp* for data set-1 is reduced from 4.04% of 4095 variables to 0.53% of 330 variables; for data set-2, it is reduced from 5.96% to 3.34%.

The prediction results of all experimentally confirmed promoters of known sigma-factors of *E. coli* are shown in Table 3. It is obvious that the prediction accuracies are both improved markedly using the iterative feature selection method. The highest average accuracies of data set-3 and data set-4 are 92.13% and 92.50%, respectively.

To verify the prediction performance of the proposed method further, it is used to predict promoter sequences of *B. subtilis*, a typical gram-positive model organism. The samples are contained in data set-5 and data set-6 and the details of them are also shown in Table 1. The prediction results of them are shown in Table 4. Surprisingly, the average accuracies of data set-5 and data set-6 are as high as 95.83% and 99.09% respectively, which are much higher than the accuracies obtained by any other existing method.

Comparison with other existing methods

Evaluation of the performance of the proposed method requires comparisons with other available methods. Because different algorithms use different negative sample sets and different fragment sizes for promoter samples, it is only

Table 5. The best prediction results of *E. coli* promoters obtained by different methods (fragments length is 80 bp)**

Methods	Results (%)		
	Sensitivity TP/(TP+FN)	Specificity TN/(TN+FP)	Precision TP/(TP+FP)
Negative samples: Coding segments			
IPMD (8)	84.9	91.4	–
Sequence Alignment Kernel+SVM (20)	82	–	84
The proposed method	**96.32**	**95.79**	**95.81**
Negative samples: Intergenic segments			
3-gram* (4)	67.75	86.10	–
IPMD (8)	81	92.7	–
Sequence Alignment Kernel+SVM (20)	81	–	81
The proposed method	**92.11**	**88.77**	**89.13**

*The negative sample set contained 709 sequence fragments from the coding region and 709 sequence segments from intergenic portions. Training data set size for *E. coli* was 1669. The paper did not give more details about the training and testing set.

**The best average accuracies among the algorithms evaluated here were shown in boldface.

Table 3. Prediction results of all known sigma-factor promoters of *E. coli* using different combination of vw Z-curve features**

Number*	Data set-3					Data set-4			
	4095	650	350	280	230	4095	1100	610	360
Results (%)									
$\bar{S}_n$	79.63	87.07	91.59	91.95	**92.44**	82.56	89.63	91.34	**92.20**
$\bar{S}_p$	75.98	88.17	90.49	91.46	**91.83**	84.51	90.12	91.10	**92.80**
$\bar{a}$	77.80	87.62	91.04	91.71	**92.13**	83.54	89.88	91.22	**92.50**

*Number: number of selected vw Z-curve variables

**The average accuracies of the vw Z-curve methods with 230 parameters for Data set-3 and 360 parameters for Dataset-4, which were the best ones among the algorithms evaluated here, were shown in boldface.

Table 4. Prediction results of all known sigma-factor promoters of *B. subtilis* using different combination of vw Z-curve features**

Number*	Data set-5				Data set-6		
	4095	872	405	340	4095	740	490
Results (%)							
$\bar{S}_n$	80.91	92.73	95.30	**95.76**	66.97	94.55	**98.79**
$\bar{S}_p$	81.82	91.97	94.24	**95.91**	73.03	95.76	**99.39**
$\bar{a}$	81.36	92.35	94.77	**95.83**	70.00	95.15	**99.09**

*Number: number of selected vw Z-curve variables.

**The average accuracies of the vw Z-curve methods with 340 parameters for Data set-5 and 490 parameters for Dataset-6, which were the best ones among the algorithms evaluated here, were shown in boldface.

possible to give a rough comparison between the proposed method and other methods.

Comparing the prediction performance of E. coli promoter. Most existing methods tested their prediction performance using σ^{70} promoter fragments of *E. coli* K-12 with 80 bp (TSS-60 ... TSS+19). They used two kinds of negative samples: coding segments and intergenic segments. The best prediction results of different methods are shown in Table 5 in detail. The most commonly used measurements of these methods are introduced to evaluate the performance of them. For both coding and non-coding negative samples, the performance of the proposed method is much better than that of other methods.

When taking intergenic segments as negative samples, the specificity obtained by IPMD is 92.7%, which is higher than the specificity obtained by the proposed method. But the average accuracy (the mean of sensitivity and specificity) obtained by IPMD is 86.85%, while the accuracy obtained by the proposed method is 90.44%. It is obviously, compared with IPMD, the average accuracy is improved by 3.59% by the proposed method. Additionally, the difference between specificity and sensitivity of IPMD and the proposed method is 11.7% and 3.34%, respectively. Consequently, the trade-off between *Sn* and *Sp* obtained by the proposed method is much better than that obtained by IPMD.

As mentioned above, approximately, for *E. coli* K-12, 81% of known TSSs are located in the intergenic non-coding regions and 19% in the coding regions (20). Partly due to these facts, the difference of patterns distribution between coding sequences and promoter sequences is much more statistically significant than that between intergenic sequences and promoter sequences. Consequently, from the results shown in Table 5, it could be seen that, for all listed methods, the recognition performance between promoter and coding sequences is better than that of promoter and non-coding sequences.

The promoter region is less stable and hence more prone to melting as compared to other genomic regions. Thus, there are some methods based on the differences in the stability of DNA sequences in the promoter and non-promoter region. Askary *et al.* presented a modified ANN (named N4) fed by nearest neighbours and based on DNA duplex stability (2). The promoter prediction sensitivity [TP/(TP+FN)] and precision [TP/(TP+FP)] of N4 for predicting promoters in *E. coli* are both 94%. To this author's knowledge, this represents the best result achieved in the existing literature.

Comparisons of the method presented here with that of Askary *et al.* (2) are made by using the same measurements and similar database construction methods. The positive sample set consists of 576 experimentally confirmed σ^{70} promoters fragments with 414 bp ([−207 ... TSS ...], the site of TSS is +1). The negative sample set consists of the first 414-bp fragment of the 530 ORFs with length of 414–585 bp. The best recognition results obtained by these two different methods are shown in Table 6. It is also obvious that the prediction accuracy of the proposed method is much better than the accuracy obtained by N4 method.

Table 6. The best recognition results of *E. coli* promoters obtained by different methods (fragments length is 414 bp)*

Methods	Sensitivity (%) TP/(TP+FN)	Precision (%) TP/(TP+FP)
The proposed method	**97.10**	**97.31**
N4 Neural Networks (2)	94	94

*The best average accuracies among the algorithms evaluated here were shown in boldface.

Comparing the prediction performance of B. subtilis promoter. *Bacillus subtilis*, a representative Gram positive bacterium, is often used to demonstrate the performance of the prokaryotic promoter prediction methods. Lin and Li applied the IPMD method to predict *B. subtilis* promoters (8). To this author's knowledge, this represents the best result achieved in the existing literature.

Comparisons of the proposed method with that of Lin and Li (8) are made by using the same measurements and similar database construction methods. The best recognition results obtained by these two different methods are shown in Table 7. In the case of coding negative samples, the prediction average accuracy is improved from 85.85% to 95.83% by the proposed method, as well as the difference between *Sn* and *Sp* is decreased from 10.9% to 0.15%. In the case of non-coding negative samples, the prediction average accuracy is improved by 15.54% by the proposed method, as well as the difference between *Sn* and *Sp* is decreased by 21.3%. The results strongly indicate that both the prediction accuracy and the trade-off between *Sn* and *Sp* are improved remarkably.

CONCLUSIONS

With the explosive development of the research on synthetic biology and genetic regulatory networks, understanding the gene regulation process has been one of the

Table 7. The best recognition results of *B. subtilis* promoters obtained by different methods (fragments length is 80 bp)*

Methods	Results (%)			
	Sensitivity (*Sn*) TP/(TP+FN)	Specificity (*Sp*) TN/(TN+FP)	Average accuracy	Difference between *Sn* and *Sp*
Negative samples: coding segments				
IPMD (8)	80.4	91.3	85.85	10.9
The proposed method	**95.76**	**95.91**	**95.83**	**0.15**
Negative samples: intergenic segments				
IPMD (8)	72.6	94.5	83.55	21.9
The proposed method	**98.79**	**99.39**	**99.09**	**0.6**

*The best average accuracies among the algorithms evaluated here were shown in boldface.

main challenges for biologists. In this context, important regulatory mechanisms involve the high precise prediction of promoter regions, which promote the initialization of gene expression processes. In this paper, a novel vw Z-curve method is developed as a feature extraction tool for prokaryotic promoter recognition for the first time. The proposed method is used in promoter prediction in *E. coli* and *B. subtilis*. Together with the iterative feature selection and classification power of the PLS algorithm, recognition accuracy and the trade-off between sensitivity and specificity are improved markedly. The simplicity of this method allows it to be particularly practical for performing research without any prior knowledge or prior information and to be run on ordinary personal computers or laptops with run times of several minutes.

Although this method is developed for prokaryotic promoter recognition, and it has only been tested on samples of *E. coli* and *B. subtilis* promoter fragments, it can easily be used for the development of eukaryotic promoter prediction methods or for the development of new motif-finding methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online: Supplementary Tables S1, Supplementary Figures S1–3, Supplementary Methods, Supplementary MATLAB codes and Supplementary References [37–56].

ACKNOWLEDGEMENTS

The author is much grateful to Prof. Chun-Ting Zhang, the Department of Physics, Tianjin University, China, for instructing the methodology of Z-curve method.

FUNDING

Funding for open access charge: National Natural Science Foundation of China (31000592); Doctoral Fund of Ministry of Education of China (200800561005).

Conflict of interest statement. None declared.

REFERENCES

1. Alberts,B. (1998) *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*. Garland Pub, New York.
2. Askary,A., Masoudi-Nejad,A., Sharafi,R., Mizbani,A., Parizi,S.N. and Purmasjedi,M. (2009) N4: A precise and highly sensitive promoter predictor using neural network fed by nearest neighbors. *Genes Genet. Syst*., **84**, 425–430.
3. Bansal,M. and Rangannan,V. (2007) Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *J. Biosci*., **32**, 851–862.
4. Rani,T.S. and Bapi,R.S. (2009) Analysis of n-gram based promoter recognition methods and application to whole genome promoter prediction. *In Silico Biol*., **9**, S1–S16.
5. Mann,S., Li,J. and Chen,Y.P. (2007) A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts. *Nucleic Acids Res*., **35**, e12.
6. Burden,S., Lin,Y.X. and Zhang,R. (2005) Improving promoter prediction Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. *Bioinformatics*, **21**, 601–607.
7. Bland,C., Newsome,A.S. and Markovets,A.A. (2010) Promoter prediction in *E. coli* based on SIDD profiles and Artificial Neural Networks. *BMC Bioinformatics*, **11**, S17.
8. Lin,H. and Li,Q.Z. (2011) Eukaryotic and prokaryotic promoter prediction using hybrid approach. *Theor. Biosci*., **130**, 91–100.
9. Zhang,C.T. (1997) A symmetrical theory of DNA sequences and its applications. *J. Theor. Biol*., **187**, 297–306.
10. Zhang,C.T., Zhang,R. and Ou,H.Y. (2003) The Z curve database: a graphic representation of genome sequences. *Bioinformatics*, **19**, 593–599.
11. Gao,F. and Zhang,C.T. (2008) Origins of replication in Cyanothece 51142. *Proc. Natl Acad. Sci. USA*, **105**, E125; author reply E126–E127.
12. Zhang,R. and Zhang,C.T. (2005) Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea*, **1**, 335–346.
13. Guo,F.B., Ou,H.Y. and Zhang,C.T. (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res*., **31**, 1780–1789.
14. Zhang,C.T. and Zhang,R. (2004) Isochore structures in the mouse genome. *Genomics*, **83**, 384–394.
15. Zhang,R. and Zhang,C.T. (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics*, **20**, 612–622.
16. Zhang,R. and Zhang,C.T. (2003) Identification of genomic islands in the genome of *Bacillus cereus* by comparative analysis with *Bacillus anthracis*. *Physiol. Genomics*, **16**, 19–23.
17. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res*., **38**, D46–D51.
18. Gama-Castro,S., Salgado,H., Peralta-Gil,M., Santos-Zavaleta,A., Muniz-Rascado,L., Solano-Lira,H., Jimenez-Jacinto,V., Weiss,V., Garcia-Sotelo,J.S., Lopez-Fuentes,A. *et al*. (2011) RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*., **39**, D98–D105.
19. Sierro,N., Makita,Y., de Hoon,M. and Nakai,K. (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res*., **36**, D93–D96.
20. Gordon,L., Chervonenkis,A.Y., Gammerman,A.J., Shahmuradov,I.A. and Solovyev,V.V. (2003) Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, **19**, 1964–1971.
21. Hook-Barnard,I.G. and Hinton,D.M. (2007) Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. *Gene Regul. Syst. Biol*., **1**, 275–293.
22. Shultzaberger,R.K., Chen,Z., Lewis,K.A. and Schneider,T.D. (2007) Anatomy of *Escherichia coli* sigma70 promoters. *Nucleic Acids Res*., **35**, 771–788.
23. Barrios,H., Valderrama,B. and Morett,E. (1999) Compilation and analysis of sigma(54)-dependent promoter sequences. *Nucleic Acids Res*., **27**, 4305–4313.
24. Tsukahara,K. and Ogura,M. (2008) Promoter selectivity of the *Bacillus subtilis* response regulator DegU, a positive regulator of the fla/che operon and sacB. *BMC Microbiol*., **8**, 8.
25. Evans,L., Feucht,A. and Errington,J. (2004) Genetic analysis of the *Bacillus subtilis* sigG promoter, which controls the sporulation-specific transcription factor sigma G. *Microbiology*, **150**, 2277–2287.
26. Gruber,T.M. and Gross,C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Ann. Rev. Microbiol*., **57**, 441–466.
27. Paget,M.S. and Helmann,J.D. (2003) The sigma70 family of sigma factors. *Genome Biol*., **4**, 203.
28. Perez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res*., **28**, 1838–1847.
29. Helmann,J.D. (2002) The extracytoplasmic function (ECF) sigma factors. *Adv. Microbial Physiol*., **46**, 47–110.
30. van Hijum,S.A., Medema,M.H. and Kuipers,O.P. (2009) Mechanisms and evolution of control logic in prokaryotic

transcriptional regulation. *MMBR*, **73**, 481–509, Table of Contents.

31. Yang,J.Y., Zhou,Y., Yu,Z.G., Anh,V. and Zhou,L.Q. (2008) Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. *BMC Bioinformatics*, **9**, 113.
32. Gao,F. and Zhang,C.-T. (2004) Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics*, **20**, 673–681.
33. Zhang,C.T. and Zhang,R. (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.*, **19**, 6313–6317.
34. Rosipal,R. and Trejo,L.J. (2002) Kernel partial least squares regression in reproducing Kernel Hilbert space. *J. Mach. Learn. Res.*, **2**, 97–123.
35. Wold,S., Sjöström,M. and Eriksson,L. (2001) PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab.*, **58**, 109–130.
36. Kvalheim,O.M. (1992) The latent variable. *Chemometr. Intell. Lab.*, **14**, 1–3.
37. Samal,A. and Jain,S. (2008) The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst. Biol.*, **2**, 21.
38. Gruber,T.M. and Gross,C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Ann. Rev. Microbiol.*, **57**, 441–466.
39. Wosten,M.M. (1998) Eubacterial sigma-factors. *FEMS Microbiol. Rev.*, **22**, 127–150.
40. Paget,M.S. and Helmann,J.D. (2003) The sigma70 family of sigma factors. *Genome Biol.*, **4**, 203.
41. Perez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 1838–1847.
42. Sierro,N., Makita,Y., de Hoon,M. and Nakai,K. (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, D93–D96.
43. Hook-Barnard,I.G. and Hinton,D.M. (2007) Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. *Gene Regul. Syst. Biol.*, **1**, 275–293.
44. Shultzaberger,R.K., Chen,Z., Lewis,K.A. and Schneider,T.D. (2007) Anatomy of *Escherichia coli* sigma70 promoters. *Nucleic Acids Res.*, **35**, 771–788.
45. Estrem,S.T., Ross,W., Gaal,T., Chen,Z.W., Niu,W., Ebright,R.H. and Gourse,R.L. (1999) Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev.*, **13**, 2134–2147.
46. McCracken,A., Turner,M.S., Giffard,P., Hafner,L.M. and Timms,P. (2000) Analysis of promoter sequences from *Lactobacillus* and *Lactococcus* and their activity in several *Lactobacillus* species. *Arch. Microbiol.*, **173**, 383–389.
47. Barrios,H., Valderrama,B. and Morett,E. (1999) Compilation and analysis of sigma(54)-dependent promoter sequences. *Nucleic Acids Res.*, **27**, 4305–4313.
48. Helmann,J.D. (2002) The extracytoplasmic function (ECF) sigma factors. *Adv. Microbial Physiol.*, **46**, 47–110.
49. Burnham,A.J., MacGregor,J.F. and Viveros,R. (1999) Latent variable multivariate regression modeling. *Chemometr. Intell. Lab.*, **48**, 167–180.
50. Kvalheim,O.M. (1992) The latent variable. *Chemometr. Intell. Lab.*, **14**, 1–3.
51. Rosipal,R. and Trejo,L.J. (2002) Kernel partial least squares regression in Reproducing Kernel Hilbert Space. *J. Mach. Learn. Res.*, **2**, 97–123.
52. Geladi,P. and Kowalski,B.R. (1986) Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, **185**, 1–17.
53. Höskuldsson,A. (1988) PLS regression methods. *J. Chemometr.*, **2**, 211–228.
54. Wold,H. (1974) Causal flows with latent variables: partings of the ways in the light of NIPALS modelling. *Eur. Economic Rev.*, **5**, 67–86.
55. Lindgren,F., Geladi,P. and Wold,S. (1993) The kernel algorithm for PLS. *J. Chemometr.*, **7**, 45–59.
56. Rännar,S., Lindgren,F., Geladi,P. and Wold,S. (1994) A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *J. Chemometr.*, **8**, 111–125.