OPEN ACCESS Freely available online

PLOS ONE

An Improved Systematic Approach to Predicting Transcription Factor Target Genes Using Support Vector Machine

Song Cui[1]*, Eunseog Youn[2], Joohyun Lee[3], Stephan J. Maas[3]

1 School of Agribusiness and Agriscience, Middle Tennessee State University, Murfreesboro, Tennessee, United States of America, **2** Department of Computer Science, Texas Tech University, Lubbock, Texas, United States of America, **3** Department of Plant and Soil Science, Texas Tech University, Lubbock, Texas, United States of America

Abstract

Biological prediction of transcription factor binding sites and their corresponding transcription factor target genes (TFTGs) makes great contribution to understanding the gene regulatory networks. However, these approaches are based on laborious and time-consuming biological experiments. Numerous computational approaches have shown great potential to circumvent laborious biological methods. However, the majority of these algorithms provide limited performances and fail to consider the structural property of the datasets. We proposed a refined systematic computational approach for predicting TFTGs. Based on previous work done on identifying auxin response factor target genes from *Arabidopsis thaliana* co-expression data, we adopted a novel reverse-complementary distance-sensitive *n*-gram profile algorithm. This algorithm converts each upstream sub-sequence into a high-dimensional vector data point and transforms the prediction task into a classification problem using support vector machine-based classifier. Our approach showed significant improvement compared to other computational methods based on the area under curve value of the receiver operating characteristic curve using 10-fold cross validation. In addition, in the light of the highly skewed structure of the dataset, we also evaluated other metrics and their associated curves, such as precision-recall curves and cost curves, which provided highly satisfactory results.

Citation: Cui S, Youn E, Lee J, Maas SJ (2014) An Improved Systematic Approach to Predicting Transcription Factor Target Genes Using Support Vector Machine. PLoS ONE 9(4): e94519. doi:10.1371/journal.pone.0094519

Editor: Hans A Kestler, University of Ulm, Germany

Received October 11, 2012; **Accepted** March 17, 2014; **Published** April 17, 2014

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: song.cui@mtsu.edu

Introduction

Unraveling the gene regulatory networks is regarded as one of the fundamental problems challenging biologists [1]. Gene expression is systematically controlled by regulatory proteins known as transcription factors (TFs) that bind to specific cognate DNA sites known as transcription factor binding sites (TFBSs). Through interacting with other *cis*-elements, these TFs can function as repressors preventing transcription by inhibiting the activity of RNA polymerization complex either by directly binding to TFBSs or indirectly modifying transcription factor target genes (TFTGs). Transcription factors can also function as activators, which promote the expression of TFTGs. In addition to post-transcriptional gene regulations, there are also post-translational gene modification regulations, including biochemical alteration and RNA interference [2,3]. However, the interplay among the corresponding TFs, TFBSs, and TFTGs remains the predominant mechanism governing the gene regulatory processes.

In order to circumvent the laborious biological experiments for screening TFBSs and their corresponding TFTGs, a number of computational algorithms have been proposed in the last decade on the basis of pre-established biological results [4–16]. Instead of directly searching for TFTGs, the majority of these algorithms focused on the nucleotide sequence information to screen potential TFBSs and ignored the structural property of DNA molecules. Local search-based algorithms, such as Gibbs sampling, have been applied on certain microorganisms with some success but lacked global optimality [4–7]. Position weight matrix-based approaches were popular but suffered greatly from high false-positive prediction rate and the independence assumption among different TFBSs [8–10]. Most recently, He *et al.* [11] refined the traditional *n*-gram profile algorithm based on the fact that a specific TF may bind on either a DNA strand or its reverse complement and produced satisfactory results. Following their work, Dai *et al.* [12] incorporated a positional signal into each potential TFBS and greatly improved prediction performances. Additionally, Meysman *et al.* [13] discussed a prediction algorithm using DNA structural information alone to predict TFBSs. *De novo* methodology-based predictions did not require any model training based on the prior knowledge of TFTGs thus showing its advantages in terms of computational cost and classification accuracy [14–16]. Unfortunately, all these approaches provided limited classification performances and failed to consider the dataset structure when interpreting the final results. Particularly, the method proposed by Dai *et al.* [12], which requires arbitrarily choosing thresholds for feature selection, could provide limited performances because the optimal threshold was not identified. Taking all these weaknesses into account, an improved new systematic computational approach for TFTG prediction was proposed in our study and produced great results.

Materials and Methods

Using the well-documented information domain of the corresponding TFs, TFBSs, and TFTGs, we constructed a binary classifier based on support vector machines (SVM). A standard feature extraction, feature selection, model construction, and dataset testing paradigm was followed. The feature extraction region was limited to within 1000-bp upstream from the transcription start point. This frame was verified to contain the most amounts of TFBSs from previous biological studies [11,12]. Once these 1000-bp sequences with identified class labels (TFTGs or non-TFTGs) were generated, they were then profiled by a new reverse-complementary distance-sensitive *n*-gram profiling (RCDSNGP) algorithm designed to better capture the patterns of potential conserved motifs and their corresponding positions relative to recognized TFBSs. For feature selection, we adopted Monte Carlo simulation based on information gain (IG) measurements to select features that have a *p*-value smaller than 0.01. Finally, each upstream sequence of either TFTGs or non-TFTGs was represented by a single data point in a multi-dimensional feature space and was later fed to SVM to build prediction models.

Feature extraction, selection, model training, and testing were performed on the basis of a 10-fold cross validation (10-fold CV). That is, the entire dataset is randomly and evenly split into 10 disjoint subsets. Each subset contains a proportion of TFTG sequences similar to that of the pre-split dataset, e.g. 19 TFTGs +260 non-TFTGs = 279. The final result is calculated from a composite of 10 trials. Within each trial, a different subset of samples is selected for testing and the other nine subsets of samples are used for training. Feature extraction and selection are performed within the nine training subsets during each trial, thus, selected features are different in each trial.

Datasets

The procedures for generating sequence datasets were described previously by Dai *et al.* [12]. In general, auxin response factors (ARFs) regulate their target genes by recognizing the primary conserved motif 'TGTCTC' or its reverse complement 'GAGACA' in the upstream region [17]. However, the presence of this conserved motif by itself may not guarantee that the corresponding sequence belongs to TFTGs. Goda *et al.* [18] used Affymetrix Genechip to investigate the gene expressions of *A. thaliana* treated with auxin and brassinosteroid and reported that only 186 out of 2787 genes containing the conserved motif ('TGTCTC' or 'GAGACA') in their 1000-bp upstream region were TFTGs.

By referring to the accession IDs verified by Goda *et al.* [18] and location information (transcription start point and chromosome ID) obtained from TAIR6 Arabidopsis Information Resource (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR6_genome_release), the 1000-bp gene upstream sequences of *A. thaliana* with the conserved motif (186 TFTGs +2601 non-TFTGs = 2787) were extracted from the genome sequences (ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/whole_chromosomes). The entire dataset can be downloaded from the Samuel Roberts Noble Foundation online supplementary data source (available via http://bioinfo.noble.org/manuscript-support/TF_Supp/dataset/).

Feature extraction

Each upstream sequence has to be converted into a series of numerical values corresponding to its coordinates in a high-order feature space for SVM training and prediction purposes. An *n*-gram profiling algorithm was used previously to represent a sequence stream by a set of *n* continuous characters and their corresponding frequencies [19]. This approach is analogous to the *k*-mer approach used in other gene sequence studies [20].

Because of the double helix structure and base-pairing property of DNA, TFs may bind on either strand of a DNA molecule. Thus, a conserved motif and its reverse complement should be treated equally for each TF. In the light of this, He *et al.* [11] proposed the reverse-complementary n-gram profile (RCNP) algorithm, formalized as follows.

Definition 1 (RCNP): Given an *m*-length sequence $s = s_1, s_2 \ldots s_m$, the RCNP of s is a set of K 2-tuples, denoted as $\mathrm{RCNP}(s) = \{(\{f_1, r_1\}, c_1), (\{f_2, r_2\}, c_2) \ldots (\{f_K, r_K\}, c_K)\}$, f_k being a distinct *n*-gram, r_k being the reverse complement of f_k, and c_k being the sum of frequency counts of f_k and r_k in s. Additionally, $\{f_k, r_k\}$ $(k = 1, 2 \ldots K)$ includes all possible combinations of an *n*-gram and its reverse complement in s.

The essence of RCNP was that an occurrence of either an *n*-gram or its reverse complement will be counted equally as one increment of that feature ($\{f_i, r_i\}$). In addition, by limiting the feature extraction region within a finite window evenly neighboring the center motif (such as 100-bp window size with 50 bp on each flank), He *et al.* [11] considered the presence of other possible synergic TFBSs within the window beside the center motif. This approach was based on the assumption that the closer an *n*-gram is to the primary TFBSs, the stronger its influence on regulating TF binding processes. An optimal area under curve (AUC) value of 0.8949 was obtained on a similar dataset using this RCNP algorithm [11].

Immediately after He *et al.*'s work, Dai *et al.* [12] expanded the RCNP algorithm into a reverse-complementary position-sensitive *n*-gram profile (RCPSNP) algorithm by incorporating a positional information parameter and a position-sensitive parameter into the RCNP. The position sensitive parameter was introduced to mainly account for the possibility that two identical *n*-grams extracted from a certain window flanking the center motif on the same DNA strand may have similar impacts on regulating TF binding processes regardless of their positional differences. This feature generation scheme yielded an AUC value of 0.73 for the receiver operating characteristic (ROC) curve [12].

In this study, we propose an improved feature generation algorithm. Studies have shown the existence of a composite structure containing constitutive elements adjacent to the 'TGTCTC' binding site for ARFs [17,21]. As a result, the auxin inducibility was likely affected incrementally by multiple elements. In addition, their contribution differences should be related with the distance from each element to the primary TFBS. None of the previous studies investigated the impact differences between the upstream-region elements and the downstream-region elements around the primary binding sites. Therefore, it was not logical to incorporate each *n*-gram with a signed integer representing its direction and distance relative to the primary TFBS as proposed by Dai *et al.* [12]. Considering all factors described above, we introduced the reverse-complementary distance-sensitive *n*-gram profile (RCDSNGP), formalized as follows.

Definition 2 (RCDSNGP): Given an *m*-length sequence $s = s_1, s_2 \ldots s_i \ldots s_{i+j} \ldots s_m$, the RCDSNGP of s with respect to a *j*-length reference subsequence $x = s_i \ldots s_{i+j-1}$ is a set of K 2-tuples, denoted as $\mathrm{RCDSNGP}(s) = \{(\{f_1, r_1, d_1\}, c_1), (\{f_2, r_2, d_2\}, c_2) \ldots (\{f_K, r_K, d_K\}, c_K)\}$, f_k being a distinct *n*-gram, r_k being the reverse complement of f_k, d_k being the relative distance parameter, and c_k being the sum of frequency counts of f_k and r_k with the same d_k relative to x in s. Additionally, $\{f_k, r_k, d_k\}$ $(k = 1, 2 \ldots K)$ include all possible combinations of an *n*-gram, its reverse complement, and its distance to x in s.

If we denote either f_k or r_k as an n-gram, $g = s_t, s_{t+1} \ldots s_{t+n-1}$ ($n-1 < t+n-1 < i$ or $m-n+2 > t > i+j-1$), then its relative distance to x is calculated as follows.

$$\begin{aligned} d_k &= t-(i+j-1), if\,(m-n+2) > t > (i+j-1) \\ d_k &= i-(t+n-1), if\,(n-1) < (t+n-1) < j \end{aligned} \quad (1)$$

Each set ($\{f_k, r_k, d_k\}$) within a 2-tuple of an RCDSNGP is a reverse-complementary distance-sensitive n-gram (RCDSNG), synonymously a feature in our study.

By adopting RCDSNGP, an occurrence of either an n-gram or its reverse complement with the same distance to the central TFBS will be counted equally as one increment of that feature ($\{f_k, r_k, d_k\}$).

Table 1 demonstrates an example of an RCDSNGP of a given sequence with respect to a designated reference subsequence.

Feature selection

We used n-grams of $n = 4$–9 for profiling each upstream sequence using RCDSNGP algorithm because $n = 4$–9 were verified (data not shown) to give optimal performances and can be handled efficiently in a moderate computing environment. Given the maximum distance d_{max}, a total number of $d_{max} \times 4^n$ features at most can be generated for each n, which is less than half of the number produced by Dai *et al.* [12].

Given a d_{max}, 10 trials were conducted to generate the final result (10-fold CV). Within each trial, a separate IG-based feature ranking was used on the nine training subsets [22]. The IG measure is based on information theory [23], which calculates the entropy differences before and after observing a specific feature. The entropy of the set S, which contained e.g. $N = 2509$ (168 TFTGs +2341 non-TFTGs $= 2509$) upstream sequences and two distinct class labels $T = \{TG, \overline{TG}\}$ for the TFTGs and non-TFTGs (number of classes $y = 2$), can be calculated as

$$\begin{aligned} Entropy(S) &= -\sum_{x=1}^{y} P(Tx,S) \times \log(P(Tx,S)) \\ &= -\frac{N_{TG}}{N} \times \log\left(\frac{N_{TG}}{N}\right) - \frac{N_{\overline{TG}}}{N} \times \log\left(\frac{N_{\overline{TG}}}{N}\right) \end{aligned} \quad (2)$$

where NTG and $N\overline{TG}$ are the numbers of sequences in S belonging to TFTGs and non-TFTGs, respectively.

After observing a specific feature f, we can partition the original set S into two distinct subsets: Sf, a set of upstream sequences containing f; $S\bar{f}$, a set of upstream sequences without f. Thus, $S = \{Sf, S\bar{f}\}$ and the number of classes is $y = 2$. The entropy of S with respect to f is evaluated as

$$\begin{aligned} Entropy(S|f) &= \sum_{x=1}^{y} P(Sx,S) \times Entropy(Sx) \\ &= -\frac{Nf}{N} \times \left(\frac{N_{TGf}}{N_f} \times \log\left(\frac{N_{TGf}}{N_f}\right) + \frac{N_{\overline{TG}f}}{N_f} \right. \\ &\quad \left. \times \log\left(\frac{N_{\overline{TG}f}}{N_f}\right)\right) - \frac{N_{\bar{f}}}{N} \times \left(\frac{N_{TG\bar{f}}}{N_{\bar{f}}} \right. \\ &\quad \left. \times \log\left(\frac{N_{TG\bar{f}}}{N_{\bar{f}}}\right) + \frac{N_{\overline{TG}\bar{f}}}{N_{\bar{f}}} \times \log\left(\frac{N_{\overline{TG}\bar{f}}}{N_{\bar{f}}}\right)\right) \end{aligned} \quad (3)$$

where numbers of upstream sequences with at least one occurrence of feature f and no occurrence of f are denoted by N_f and $N_{\bar{f}}$, respectively; numbers of upstream sequences belonging to TFTGs with at least one occurrence of feature f and no occurrence of f are denoted by $NTGf$ and $N_{TG\bar{f}}$, respectively; and, $N_{\overline{TG}f}$ and $N_{\overline{TG}\bar{f}}$ represent the corresponding numbers of non-TFTGs. Finally, the IG obtained by dividing S according to f is calculated using:

$$IG(f) = Entropy(S) - Entropy(S|f) \quad (4)$$

and a higher IG value implies greater importance of a given feature for representing a specific sequence.

Instead of ranking features based on IGs and subjectively choosing the cutoff value, we further evaluated each feature using Monte Carlo simulations [24,25]. This approach provides information on whether we can statistically differentiate samples from two classes on a basis of a given feature. For each feature, we shuffled class labels (TFTGs or non-TFTGs) 10 000 times without changing either the feature count in each sequence or the total number of sequences in each class. A new IG was calculated for each shuffling, thus, 10 000 IGs were obtained for each feature.

Table 1. An example of a reverse-complementary distance-sensitive n-gram profile (RCDSNGP) representation with $n = 4$, 5, and 6 for a given sequence (AAGCTT**GAGACA**CAGCT) with the reference subsequence marked in bold*.

Length of n-gram (n)	Reverse-complementary distance-sensitive n-gram (RCDSNG), or feature	Frequency count
$n = 4$	{AAGC, GCTT, 1}	1
	{AGCT, AGCT, 2}	2
	{AAGC, GCTT, 3}	1
	{CAGC, GCTG, 1}	1
$n = 5$	{AAGCT, AGCTT, 1}	1
	{AAGCT, AGCTT, 2}	1
	{AGCTG, CAGCT, 1}	1
$n = 6$	{AAGCTT, AAGCTT, 1}	1

*Given an m-length sequence $s = s_1, s_2 \ldots s_i \ldots s_{i+j} \ldots s_m$, the RCDSNGP of s with respect to an j-length reference subsequence $x = s_i \ldots s_{i+j-1}$ is a set of K 2-tuples, denoted as RCDSNGP(s) RCDSNGP(s) = {($\{f_1, r_1, d_1\}, c_1$),($\{f_2, r_2, d_2\}, c_2$)…($\{f_K, r_K, d_K\}, c_K$)}, f_k being a distinct n-gram, r_k being the reverse complement of f_k, d_k being the relative distance parameter, and c_k being the sum of frequency counts of f_k and r_k with the same d_k relative to x in s. Each set in a 2-tuple ($\{f_k, r_k, d_k\}$) is a reverse-complementary distance-sensitive n-gram (RCDSNG), or a feature in our study. This RCDSNGP representation was adopted for all training sequences. For testing processes, each sequence was converted to RCDSNGP first, and then represented according to the selected RCDSNGs generated from the training datasets, including those with zero count.

doi:10.1371/journal.pone.0094519.t001

The *p*-value for a specific feature was calculated according to

$$Pvalue = \frac{Nge}{N} \quad (5)$$

where *Nge* represents the number of shuffling that gave new IG values greater than or equal to the original one and N is the total number of shuffling (N= 10 000). The smaller the *p*-value is, the stronger the contribution the feature would provide to differentiate a sample between two classes. In our study, features were considered important at $p<0.01$ level.

Data representation

Within each trial, we selected F features whose *p*-values were smaller than 0.01 from the total of 493 781 all possible features (number of RCDSNGs obtained from all sequences). Each 1000-bp upstream sequence was represented by an RCDSNGP consisting of these F features and their corresponding occurrences. The whole set of N sequences (e.g. N= 168 TFTGs +2341 non-TFTGs = 2509) were then represented by an $N\times(F+1)$ matrix (1 extra column for class labels).

Training and testing

Support vector machine-based approaches have been widely used in various problem domains, such as bioinformatics [26–28]. They often outperformed many other classification algorithms [29]. We adopted SVM in our study using the LIBSVM [30], which is based on sequential minimal optimization. The general concept of SVM is given below.

For model training, given a set of vector-label pairs $(Xi,yi), i=1,2...N$, where N is equal to the number of upstream sequences (e.g. N= 2509); $Xi \in R^n$, where n is the dimension of Xi, equal to the number of selected features; $yi \in \{1,-1\}$ where 1 corresponds to TFTGs and -1 corresponds to non-TFTGs, the support vector machine computes the solution to the optimization problem formalized below:

$$\begin{aligned} &\min_{W,b,\varepsilon} \frac{1}{2} W^T W + C \sum_{i=1}^{N} \varepsilon_i \\ &\text{subject to } y_i(W^T \Theta(X_i) + b \geq 1 - \varepsilon_i \text{ and } \varepsilon_i \geq 0. \end{aligned} \quad (6)$$

This is equivalent to finding the maximum-margin hyperplane that separates TFTG samples (labeled as 1) and non-TFTG samples (labeled as -1) with minimized measures of errors. Predictions are made based on the geometric location of an unknown sample when fed into the model. A label is assigned to a sample according to which side of the hyperplane it resides. In order to obtain better classification accuracies, data are often projected into a high-dimension feature space with a kernel function. We evaluated other possible kernels in addition to the linear kernel suggested by Dai *et al.* [12]. Particularly, we performed a grid (factorial) search [28] for an optimal combination of a penalty factor C of SVM and a kernel width σ for the Gaussian radial basis function (RBF) kernel.

Performance measure

First, we measured the traditional accuracy defined as

$$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

where all parameters are defined in Table 2. Accuracy provides a direct and simple way of evaluating performances, however, it is highly sensitive to data distribution [29]. In our study, if a classifier predicts every TFTG sequence as a non-TFTG sequence, we can still obtain an accuracy of 0.9333 due to the fact that more than 93 percent of the sequences belong to non-TFTGs. Thus, other evaluation metrics are warranted, including precision, recall, and F_1 [29] defined as follows.

$$PRECISION = \frac{TP}{TP+FP} \quad (8)$$

$$RECALL = \frac{TP}{TP+FN} \quad (9)$$

$$F_1 = \frac{2\times TP}{TP+FN+TP+FP} \quad (10)$$

Furthermore, we adopted the ROC curve technique, which demonstrated satisfactory performances across different classifiers and datasets from previous studies [12]. Some recent studies argued that the ROC approach tends to provide an over-optimistic evaluation on highly imbalanced datasets [30]. Thus, we also examined precision-recall (PR) curves, which appeared more informative and valid than ROC curves on skewed datasets [31]. The AUC value was calculated on both the ROC curve (ROC-AUC) and the PR curve (PR-AUC) to summarize the classification results. To better visualize the misclassification costs and statistical significances, cost curves [32] were also evaluated.

Dai *et al.* [12] reported that a window size of 200 bp (100 bp on each flank around the central TFBS, 'TGTCTC' or 'GAGACA') provided the best ROC-AUC value. In our study, a new feature generation scheme was adopted. Therefore, we evaluated different maximum distances (d_{max} = half of the window size) on each flank of the central TFBSs, including d_{max} = 25, 50, 75, 100, 125, 150, 175, and 200 bps.

To summarize, we evaluated eight different d_{max} settings within each trial on the basis of 10-fold CV. Within each d_{max} setting during each trial, an IG-based *p*-value selection procedure was adopted to select the important ($p<0.01$) features based on the training dataset (nine folds out of 10). Immediately following that, we located the optimal combination of C for SVM and σ for the RBF kernel using the grid (factorial) search. Thus, the optimal combination of features, C, and σ depended on the training dataset within each trial. The final results, including accuracy, precision, recall, and F_1 as well as the corresponding AUC values

Table 2. Confusion matrix for performance evaluation with positive class label (+1) denoting transcription factor target gene (TFTG) and negative class label (-1) denoting non-TFTG.

		True class label	
		+1	−1
Predicted class label	+1	TP^{++}	FP^{-+}
	−1	FN^{--}	TN^{+-}

$^{++}$, $^{-+}$, $^{--}$, $^{+-}$ denote true positive, false positive, false negative, and true negative, respectively.

doi:10.1371/journal.pone.0094519.t002

Table 3. Number of unique features (union of selected features with p-value <0.01 based on 10-fold cross validation) and classification performances [evaluated as accuracy, precision, recall, F_1, area under the curve (AUC) value of receiver operating characteristic (ROC) curve, and AUC value of precision-recall (PR) curve] affected by the maximum distance on each flank from the central binding site (d_{max}) based on 10-fold cross-validation on transcription factor target gene prediction using reverse-complementary distance-sensitive n-gram profile algorithm with $n = 4$–9 and support vector machine with Gaussian radial basis function kernel.

d_{max}	Unique Feature	Accuracy	Precision	Recall	F_1	ROC-AUC	PR-AUC
25	893	0.9476	0.6923	0.3871	0.4966	0.7202	0.4180
50	1870	0.9541	0.7544	0.4624	0.5733	0.7562	0.5286
75	2732	0.9559	0.7739	0.4785	0.5914	0.7739	0.5554
100	3502	0.9566	0.7519	0.5215	0.6159	0.7720	0.5808
125	4255	0.9580	0.7899	0.5054	0.6164	0.7640	0.5646
150	4949	0.9602	0.8319	0.5054	0.6288	0.7626	0.5690
175	5622	0.9587	0.8034	0.5054	0.6205	0.7673	0.5639
200	6136	0.9580	0.7805	0.5161	0.6214	0.7664	0.5879

doi:10.1371/journal.pone.0094519.t003

were generated from the combined classification results containing all 2787 sequences according to different d_{max} settings.

Results

Extraction of features from sequences

Using a maximum distance $d_{max} = 150$ and $n = 4$–9 for building RCDSNGP, 2 455 252 unique features were generated from 1000-bp upstream of 2787 sequences (186 TFTGs +2601 non-TFTGs). We eliminated singleton features that contain only 1 occurrence across all sequences to reduce the feature size down to 735 624. The same technique was applied to all other settings of d_{max} as well.

Selection of representative features

Information gain calculated for each feature was used in Monte Carlo simulation to generate a p-value for that feature. We selected those features that had p-values smaller than 0.01. For different maximum distances, $d_{max} = 25$, 50, 75, 100, 125, 150, 175, and 200, we selected 893, 1870, 2732, 3502, 4255, 4949, 5622, and 6136 unique features, respectively on the basis of 10-fold CV (Table 3). Table 4 lists the top 20 features ranked by their p-values with corresponding IGs, when $d_{max} = 150$.

Classifier performances

We implemented our own 10-fold CV SVM with Python programming language on the basis of LIBSVM. Using the $p<0.01$ threshold, models were constructed based on different combinations of maximum distances and kernels (polynomial, RBF, sigmoid, and linear kernels), with n-grams of $n = 4$–9. The RBF kernel provided the best performances regardless of measurement metrics used. In our study, the best accuracy (0.9602), precision (0.8319), and $F1$ (0.6288) values were obtained with $d_{max} = 150$. The best recall (0.5215), ROC-AUC (0.7739), and PR-AUC (0.5879) values were obtained with $d_{max} = 100$, 75, and 200, respectively (Table 3).

Figure 1a shows the response of accuracy, precision, recall, and F_1 values versus d_{max}. Accuracy fails to provide adequate information on evaluating the minority samples (TFTGs). Combining different evaluation metrics tends to provide comprehensive assessment of classification on imbalanced datasets. When $d_{max} = 25$, our model suffered from low recall rate because of its poor accuracy when classifying the positive samples (TFTGs). Starting from $d_{max} = 50$, a boost in performances was observed in recall and $F1$ values as a result of increased accuracy in predicting positive samples. At $d_{max} = 150$, the model reaches the highest accuracy value of 0.9602. The best precision value is 0.8319 when $d_{max} = 150$ and declines a little as d_{max} increases. Starting from $d_{max} = 100$, recall values are above 0.5 and peak at $d_{max} = 100$ with a value of 0.5215. Likewise, the F_1 scores are above 0.6 when d_{max} is bigger or equal to 100 and reach the greatest value of 0.6288 when $d_{max} = 150$. Figure 1b shows an AUC-versus-d_{max} curve based on both the ROC curve and the PR curve. The ROC-AUC value arrives at 0.7739 when $d_{max} = 75$. A slight decrease is detected when $d_{max}>100$. Furthermore, by varying d_{max} the PR-AUC value responds more obviously than the ROC curve. Beginning at $d_{max} = 100$, our model produces above-0.56 PR-AUC values, which gradually decrease from $d_{max} = 100$ to $d_{max} = 125$ and peaks at 0.5879 when $d_{max} = 200$. Overall, a 150-d_{max} setting is likely to give a superior performance with limited complexity of computation compared to other maximum distance settings. Although it fails to produce the optimal AUC values for either ROC or PR curve, it provides the best accuracy, precision, and F_1 values. Considering the fact that most d_{max} value settings perform well in detecting the correct negative samples (non-TFTGs), the model that is most capable of identifying the correct positive samples (TFTGs; high precision value) yields the best results. A maximum distance $d_{max} = 200$ provides great PR-AUC values. However, it adds great computational cost for feature generation and selection processes (possibly $50 \times \sum_{i=4}^{9} 4^i = 17\ 472\ 000$ more features) and some of its performance metrics are even worse than $d_{max} = 150$. Therefore, we conclude that features within a maximum distance $d_{max} = 150$ around central TFBSs contain sufficient information for making accurate prediction on TFTGs.

In order to show the advantages of our proposed RCDSNGP algorithm compared to other algorithms [12], we also examined the ROC, PR, and cost curve as well as precision, recall, and F_1 value generated by different algorithms [12] based on the same 10-fold CV split. Additionally, comparisons were made among classifiers with different kernels. Figure 2a indicates that RCDSNGP-based model with RBF kernel outperformed all other models because it generates a curve that is closer to the perfect classification point (0,1) in the ROC curve compared with all others. Interestingly, the polynomial kernel produced bad result

Table 4. The top 20 smallest *p*-value reverse-complementary distance-sensitive *n*-grams (RCDSNGs; $n = 4–9$) selected as representative features with their information gain (IG) values and *p*-values in the 1000-bp upstream of 186 transcription factor target genes (TFTGs) and 2601 non-TFTGs, when the maximum distance (half of the window size) on each flank of the central transcription factor binding site $d_{max} = 150$.

Ranking	RCDSNG (feature)	IG value	*p*-value
1	{ACACGT, ACGTGT, 4}	0.001885	0
2	{CGAGAA, TTCTCG, 82}	0.001884	0
3	{AATATAA, TTATATT, 52}	0.001884	0
4	{ACTTCC, GGAAGT, 30}	0.001880	0
5	{ACACC, GGTGT, 44}	0.001880	0
6	{GTAC, GTAC, 39}	0.001701	0
7	{CAAACA, TGTTTG, 149}	0.001707	0
8	{AAAAATA, TATTTTT, 44}	0.001707	0
9	{AGTAT,ATACT, 51}	0.001713	0
10	{ATGATTA, TAATCAT, 130}	0.001656	0
11	{ACTTC, GAAGT, 30}	0.001516	0
12	{CTAAC, GTTAG, 91}	0.001476	0
13	{ACAAATA, TATTTGT, 71}	0.001463	0
14	{ATACG, CGTAT, 49}	0.001463	0
15	{AAAACC, GGTTTT, 75}	0.001463	0
16	{AAAGACA, TGTCTTT, 117}	0.001463	0
17	{TAAAACA, TGTTTTA, 85}	0.001463	0
18	{AGTATA, TATACT, 124}	0.001458	0
19	{AATGTG, CACATT, 43}	0.001412	0
20	{ATACCC, GGGTAT, 16}	0.001412	0

doi:10.1371/journal.pone.0094519.t004

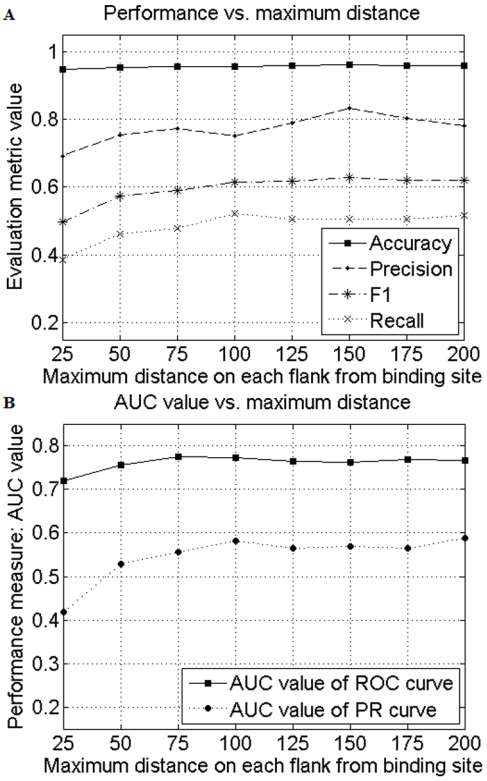

Figure 1. Performances of transcription factor target gene prediction affected by the maximum distance on each flank from the binding site (d_{max}) based on 10-fold cross-validation using reverse-complementary distance-sensitive *n*-gram profile algorithm with $n = 4–9$ and support vector machine with Gaussian radial basis function kernel. (A) Performance evaluation metric (accuracy, precision, recall, and F_1) values versus d_{max} on each flank from the central binding site. (B) Area under the curve (AUC) value of receiver operating characteristic (ROC) curve and precision-recall (PR) curve versus d_{max} on each flank from the central binding site.
doi:10.1371/journal.pone.0094519.g001

because it predicted each sample with a fixed score of -1. Classifiers that dominate in ROC space should dominate in PR space as well [33]. This is vividly presented by Figure 2b. Furthermore, our study highlighted the drawbacks of over-dependency on the ROC-AUC value when evaluating the classification performances. Although the difference in ROC-AUC values was relatively small (0.7626 versus 0.5055) between RCDSNGP and RCPSNP algorithms, the difference in PR-AUC values was remarkable (0.569 versus 0.0773).

Finally, we evaluated the cost curve of each model. The cost curve, which emphasizes the expected misclassification cost for performance measure, is proposed to better visualize the misclassification cost and compare performances on the basis of statistical significance [32]. Each data point in the ROC curve gets mapped to a distinct straight line (cost line) in the cost curve by connecting point (0, FP) and point (1, FN). Multiple points in the ROC space would generate several cost lines and form a lower envelope, which is shown in Figure 2c. The *x*-axis, denoted as probability cost, includes all possible percentage values of positive samples (0 to 1). It represents the proportion of positive samples when the classifier is deployed. Thus, at $x = 0$ (no positive samples), the only possible misclassification errors are FPs. Likewise, at $x = 1$ (no negative samples), the only possible misclassification errors are FNs. The straight line connecting these two points represents the trend of misclassification cost as percentage of positive samples varies. The lower envelope generated by a non-discrete classifier such as SVM or Artificial Neural Network is the counterpart for the upper convex hull of the ROC curve. At each probability cost value, the closer the curve is to the *x*-axis, the better the classifier performs (a lower expected cost). As presented in Figure 2c, the RCDSNGP-based model with RBF kernel has the lowest cost from 0 to 68 percent of positive samples. Additionally, it is also verified to be different ($p<0.05$) from RCPSNP-based model within the 4.8 to 65 percent range of positive samples using the method proposed by Drummond and Holte [32]. In our study, the dataset contains 6.7 percent of positive samples, which is within the 4.8 to 59 percent range, thus our model outperformed RCPSNP-based model in terms of expected cost.

To summarize, using new feature generation and selection strategies to predict TFTGs of ARFs in *A. thaliana* based on published datasets [12], we drastically increased classification performances. Our best result was obtained when $d_{max} = 150$ using the RBF kernel based on an average of 2395 features. We adopted the ROC measure for efficacy evaluation and obtained an ROC-AUC value of 0.7626 (SE = 0.0021), accuracy value of 0.9602 (SE = 0.0023), precision value of 0.8319 (SE = 0.0331), recall value of 0.5054 (SE = 0.0231), and F_1 score of 0.6288 (SE = 0.0211; Table 3), which were higher than the best result reported by Dai *et al.* (12; ROC-AUC = 0.73, accuracy = 0.69, precision = 0.3684, recall = 0.1129, and $F_1 = 0.1728$) based on the same dataset but with a different 10-fold CV split.

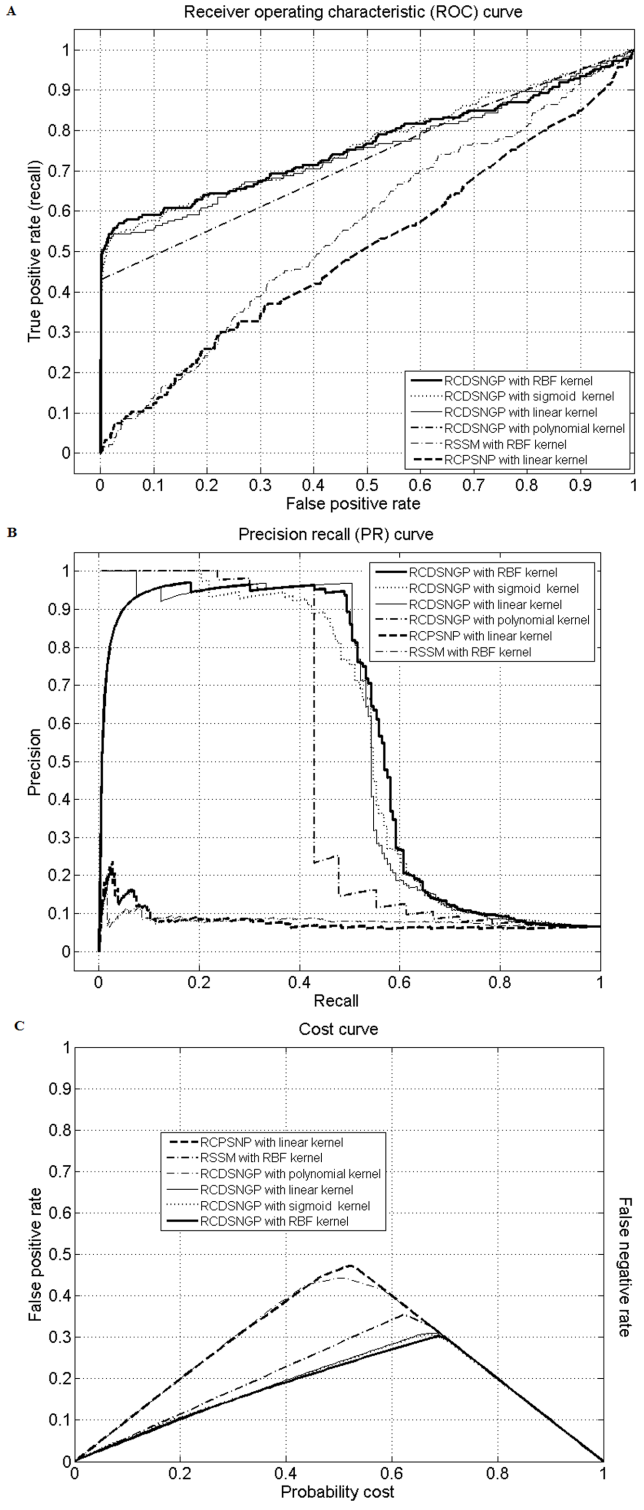

Figure 2. Classification performances using the optimal maximum distance on each flank from the binding site (d_{max} = 150). (A) Receiver operating characteristic (ROC) curve, (B) precision-recall (PR) curve, and (C) cost curve of the 10-fold cross-validation on transcription factor target gene prediction using reverse-complementary distance-sensitive *n*-gram profile (RCDSNGP) algorithm with $d_{max} = 150$ and $n = 4–9$ based on different support vector machine (SVM) kernels, reverse-complementary position-sensitive *n*-gram profile (RCPSNP) algorithm using linear-kernel SVM, and Position-Specific Scoring Matrices (PSSM)-based approach.
doi:10.1371/journal.pone.0094519.g002

Additionally, Dai *et al.* [12] reported the performance of a Position-Specific Scoring Matrices (PSSM)-based approach using the cluster-buster algorithm [34], which only yielded an ROC-AUC value of 0.51. We implemented a traditional approach based on the position frequency matrix method using a similar feature encoding algorithm explained in Youn *et al.* [26]. Each sequence was parsed according to a $d_{max} = 150$ setting and the sequence conservation was evaluated using a four (four nucleotides) by 300 ($2 \times d_{max}$) position frequency matrix (150-bp flanking each side of the primary conserved motif). At each residue (1 out of 300), we considered 20-bp window size (left:10, right: 10) to construct the frequency count for each nucleotide. The standard SVM-based training and testing was performed based on generated PSSMs. Likewise, the algorithm only provided an ROC-AUC value of 0.5569 and a PR-AUC value of 0.0801 using the same 10-fold CV split as our RCDSNGP algorithm (Table 5). The performance curves based on PSSM are also presented in Fig. 2.

Furthermore, we also implemented the RCPSNP algorithm proposed by Dai et al [12] using the optimal settings (e.g. $n = 4–9$, linear-kernel SVM) and applied it to the same 10-fold CV split as our RCDSNGP algorithm, which yielded an ROC-AUC value of 0.5055, PR-AUC value of 0.0773, accuracy value of 0.9300, precision value of 0.2000, recall value of 0.0161, and F_1 score of 0.0299 (Table 5). Our classifier generated points much closer to the perfect classification point (0,1) in the ROC curve than those generated by RCPSNP algorithm (Fig. 2a; 12). Most importantly, considering that traditional metrics for measuring classification performances tended to provide deceiving or inadequate information of imbalanced datasets, we also evaluated other metrics and their corresponding curves such as PR and cost curves, which showed greatly improved results as well.

The detailed model files, 10-fold CV datasets represented as matrices, and classification results are available at the supplementary online data source.

Discussion

Understanding the mechanism of gene regulatory network is a challenging task. As of today, there is still much uncertainty in identifying the corresponding TFBSs and TFTGs. More TF and TF-dependent target gene regulation studies are required to evaluate the biological function and mechanism of more gene regulation players. The activity and affinity of TF would be the ultimate balanced result of the various check points of biological regulation. The binding efficiency of TF to its corresponding TFBS is regulated by various factors, including TF synthesis, ligand binding to the TFs, and DNA binding mechanism through post-translational modifications such as phosphorylation and glycosylation of the TFs. In addition, the DNA binding process, dimerization, and interactions with cofactors for the functional complex formation are important parameters controlling the TF activity [35,36]. As more information of the interplay among TF, its corresponding TFBS, and TFTG accumulates, it could be possible to understand the precise TFTG expression affected by different TFs. A number of computational approaches that rely on well-known TFBSs have been proposed, but a majority of these algorithms suffered from high FP rate [12,37]. Therefore, much effort was put on reducing FP rates and increasing prediction accuracies [12], whereas the importance of the dataset structure was ignored. In our study, only 186 out of 2787 genes that all contain the binding site ('TGTCTC' or 'GAGACA') in their 1000-bp upstream region were TFTGs. If our model correctly predicted all negative samples (non-TFTGs) and miss-predicted all positive samples (TFTGs), it still yielded an accuracy value of 0.9333

Table 5. Classification performances [evaluated as accuracy, precision, recall, F_1, area under the curve (AUC) value of receiver operating characteristic (ROC) curve, and AUC value of precision-recall (PR) curve] using different feature encoding algorithms with optimal parameter settings for SVM and $d_{max} = 150$, including reverse-complementary distance-sensitive *n*-gram profile (RCDSNGP), reverse-complementary position-sensitive *n*-gram profile (RCPSNP), and a Position-Specific Scoring Matrices (PSSM)-based algorithms.

Feature Encoding Algorithm	Accuracy	Precision	Recall	F_1	ROC-AUC	PR-AUC
RCDSNGP	0.9602	0.8319	0.5054	0.6288	0.7626	0.5690
RCPSNP	0.9300	0.2000	0.0161	0.0299	0.5055	0.0773
PSSM	0.9332	NAN	0	0	0.5569	0.0801

doi:10.1371/journal.pone.0094519.t005

(2601/2787) but with precision value undefined and zero values for both recall and F_1 score. Therefore, minimizing the FP rate or maximizing the accuracy contributes little to improving overall performances when analyzing a highly skewed dataset. It is important to analyze different evaluation metrics to better assess the classification performances.

We deployed a novel feature extraction method (RCDSNGP) that incorporated a relative distance parameter into each feature to count for the positional information of each motif relative to the central TFBS. For feature selection, we adopted a Monte Carlo simulation-based statistical approach rather than arbitrarily choosing thresholds. We compared our results with the RCPSNP-based approach [12] on the same dataset. Our best model achieved an accuracy of 0.9602 and an ROC-AUC value of 0.7627 when $d_{max} = 150$ compared with 0.69 and 0.73 reported by Dai *et al.* [12], respectively. Dai *et al.* introduced three parameters for constructing RCPSNP, including a number of *n*-grams *C* (analogous to our maximum distance d_{max}), a top *F* representative features based on IG, and a position sensitive factor *P* (the identical *n*-grams located within a *P*-bp region neighboring the central binding site are counted equally). The best result was obtained when $n = 4$–9, $C = 100$, $P = 100$, and $F = 1000$. Their detailed results containing prediction scores can be found in their supplementary web data [12]. Moreover, little positional information is considered when *C* equals *P*. In other words, RCPSNP behaves almost the same as RCNP [11] when *C* and *P* hold the same value. The significant performance increase based on our RCDSNGP algorithm indicated that ARFs function by recognizing multiple consensus motifs that might be co-occurring TFBSs or subsequences functioning coordinately. More importantly, the relative distance from each motif to the binding site should always play an important role in the gene regulation process. The structural complexities of protein and DNA may result in a type of mutual recognition that relies more on the distance from the conserved motif to the TFBS, regardless of where the motif lies (downstream or upstream of the central TFBS). The PSSM-based approaches may be useful for TFTG identification when more associated TFBSs are known.

Identifying patterns of other potential TFBSs and the binding property of ARFs by enumerating all possible *n*-grams is a computationally expensive work. The complexity becomes even greater when a distance parameter is included. Therefore, better feature selection methods become necessary. We employed a statistical systematic approach. Based on a given feature, two class samples are different from each other if, and only if, the probability of obtaining a bigger IG value than original is below a certain level (*p*-value). This probability value is obtained using Monte Carlo simulations [38]. We verified that our feature selection algorithm is robust for a range of *p*-values (between 0.005 and 0.01). However, when the *p*-value becomes bigger, feature number increases drastically, which greatly increases computational cost. Moreover, we also evaluated a number of important features that have a *p*-value smaller than the 0.01 threshold versus different d_{max} values (data not shown). The slope of the curve reached the maximum value between $d_{max} = 25$ and $d_{max} = 50$ and began to dwindle when $d_{max} > 50$, suggesting that flank regions closer to the core motif contain more important features for predicting TFTGs.

Precision-recall curve is used in information retrieval as an alternative to ROC curve when analyzing imbalanced datasets [33]. Optimal prediction models tend to generate curves close to the upper-left corner in the ROC curve and upper-right corner in the PR curve. Likewise, the cost curve is introduced to measure the performances by varying class probabilities to generate confidence intervals [32]. Regardless of which curve was used, RCDSNGP-based approaches using RBF kernel demonstrated significant advantages over the RCPSNP-based approach [12]. The polynomial kernel somehow yielded much poorer performances than others. Cost curves verified the similar effects by showing that superior models always generate a lower envelope curve than inferior ones. In other words, superior models always have significantly lower misclassification cost within a certain percentage range of positive samples.

Taken altogether, the RCDSNGP algorithm combined with statistical feature selection methods provides an efficient and highly accurate way to predict TFTGs on the basis of well-studied TFBSs. We believe that this improved methodology can be employed when analyzing other species besides *A. thaliana*. It might also provide new insights into the understanding of gene regulatory networks.

Acknowledgments

We thank the associated editor and two anonymous reviewers for their valuable comments and constructive suggestions on the earlier draft of the manuscript.

Supplementary Data

Online supplementary data are available at http://redwood.cs.ttu.edu/~euyoun/TFTG.html.

Author Contributions

Conceived and designed the experiments: SC EY JL SM. Performed the experiments: SC EY. Analyzed the data: SC EY JL. Contributed reagents/materials/analysis tools: SC EY. Wrote the paper: SC EY JL SM.

References

1. Sinha S, Tompa M (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res 30: 5549–5560.
2. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, et al. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 411: 494–498.
3. Ruvkun G (2001) Molecular biology. Glimpses of a tiny RNA world. Science 294: 797–799.
4. Lawrence CE, Altschul S, Boguski M, Liu J, Neuwald A, et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262: 208–214.
5. Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296: 1205–1214.
6. Robison K, McGuire AM, Church GM (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. J Mol Biol 248: 241–254.
7. McCue LA, Thompson W, Carmack CS, Ryan M, Liu J, et al. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. Nucleic Acids Res 39: 774–782.
8. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16–23.
9. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108–D110.
10. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576–3579.
11. He J, Dai X, Zhao X (2006) A systematic computational approach for transcription factor target gene prediction. 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2006) Toronto, Ontario, Canada, pp. 385–391.
12. Dai X, He J, Zhao X (2007) A new systematic computational approach to predicting target genes of transcription factors. Nucleic Acids Res 35: 4433–4440.
13. Meysman P, Dang TH, Laukens K, Smet RD, Wu Y, et al. (2010) Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. Nucleic Acids Res 39: e6.
14. Boeva V, Surdez D, Guillon N, Tirode F, Fejes AP, et al. (2010) De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. Nucleic Acids Res 38: e126.
15. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137–144.
16. Friberg M, von Rohr P, Gonnet G (2005) Scoring functions for transcription factor binding site prediction. Bmc Bioinform 6: 84.
17. Ulmasov T, Liu ZB, Hagen G, Guilfoyle TJ (1995) Composite structure of auxin response elements. Plant Cell 7: 1611–1623.
18. Goda H, Sawa S, Asami T, Fujioka S, Shimada Y, et al. (2004) Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in Arabidopsis. Plant Physiol 134: 1555–1573.
19. Shannon C (1997) A mathematical theory of communication. Bell Syst Tech J 27: 379–423.
20. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261–5267.
21. Liu ZB, Ulmasov T, Shi X, Hagen G, Guilfoyle TJ (1994) Soybean GH3 promoter contains multiple auxin-inducible elements. Plant Cell 6: 645–657.
22. Youn E, Jeong MK (2009) Class dependent feature scaling method using naive Bayes classifier for text datamining. Pattern Recognit Lett 30: 477–485.
23. Yang Y, Pedersen JP (1997) A Comparative Study on Feature Selection in Text Categorization. Proceedings of the Fourteenth International Conference on Machine Learning Morgan Kaufmann Publishers Inc., Nashville, TN, USA, pp. 412–420.
24. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5: e1000352.
25. Pitta DW, Pinchak WE, Dowd SE, Osterstock J, Gontcharova V, et al. (2010) Rumen bacterial diversity dynamics associated with changing from bermudagrass hay to grazed winter wheat diets. Microb Ecol, 59: 511–522.
26. Youn E, Peters B, Radivojac P, Mooney SD (2007) Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 16: 216–226.
27. Tzahor S, Aharonovich DM, Kirkup B, Yogev T, Frank IB, et al. (2009) A supervised learning approach for taxonomic classification of core-photosystem-II genes and transcripts in the marine environment. BMC Genomics 10: 229.
28. Patil K, Haider P, Pope P, Turnbaugh P, Morrison M, et al. (2011) Taxonomic metagenome sequence assignment with structured output models. Nat Methods 8: 191–192.
29. Joachims T (1999) Making large-Scale SVM Learning Practical. In: Scholkopf B, Burges C, Smola A, editors. Advances in Kernel Methods - Support Vector Learning. Cambridge: MIT press. pp. 41–56.
30. Chang CC, Lin CJ (2011) LIBSVM: a library for Support Vector Machines. ACM Trans Intell Syst Technol 2: 1–27.
31. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans. Knowledge Data Eng 21: 1263–1284
32. Drummond C, Holte RC (2006) Cost curve: an improved method for visualizing classifier performance. Mach Learn 65: 95–130.
33. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. Proceedings of the twenty-third International Conference on Machine Learning, Pittsburgh, PA, USA, pp. 233–240.
34. Siggers T, Duyzend MH, Reddy J, Khan S, Bulyk ML (2011) Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Mol Syst Biol 7: 555.
35. Stower H (2012) Gene regulation: Resolving transcription factor binding. Nat Rev Genet 13: 71.
36. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21: 2933–2942.
37. Frith MC, Li MC, Weng Z (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Res 31: 3666–3668.
38. Draminski M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, et al. (2008) Monte carlo feature selection for supervised classification. Bioinformatics 24: 110–117.