Published online 18 June 2007

Nucleic Acids Research, 2007, Vol. 35, No. 13 ***4433–4440***
doi:10.1093/nar/gkm454

A new systematic computational approach to predicting target genes of transcription factors

Xinbin Dai, Ji He and Xuechun Zhao*

Plant Biology Division, the Samuel Robert Noble Foundation, Ardmore OK 73401, USA

Received December 21, 2006; Revised May 17, 2007; Accepted May 21, 2007

ABSTRACT

Identifying transcription factor target genes (TFTGs) is a vital step towards understanding regulatory mechanisms of gene expression. Methods for the *de novo* identification of TFTGs are generally based on screening for novel DNA binding sites. However, experimental screening of new binding sites is a technically challenging, laborious and time-consuming task, while computational methods still lack accuracy. We propose a novel systematic computational approach for predicting TFTGs directly on a genome scale. Utilizing gene co-expression data, we modeled the prediction problem as a 'yes' or 'no' classification task by converting biological sequences into novel reverse-complementary position-sensitive *n*-gram profiles and implemented the classifiers with support vector machines. Our approach does not necessarily predict new DNA binding sites, which other studies have shown to be difficult and inaccurate. We applied the proposed approach to predict auxin-response factor target genes from published *Arabidopsis thaliana* co-expression data and obtained satisfactory results. Using ten-fold cross validations, the area under curve value of the receiver operating characteristic reaches around 0.73.

INTRODUCTION

Understanding genetic regulatory networks is a key to elucidating various biological processes. In eukaryotes, an integrated regulatory network comprises transcription factors (TFs), target genes and their relationships. Although recent studies have also implicated small non-coding RNAs in the regulation of gene expression at the post-transcriptional level (1), TFs and their corresponding target genes are still regarded as key components of regulatory networks.

Screening new transcription factor binding sites (TFBSs) is the most common approach to identifying transcription factor target genes. Various experimental methods have been applied to search for new TFBSs, e.g. electrophoresis mobility shift assays (2), enzyme activity analysis of cellulose D (CELD) fusion protein (3) and the high-throughput Chromatin Immunoprecipitation (ChIP) chip approach (4). However, such experiments are difficult to perform on a large scale because they are costly and time-consuming. For example, there are nearly 2000 TFs in the first complete sequenced model plant species, *Arabidopsis thaliana* (5), but fewer than 20 of these have been experimentally validated so far. To circumvent the low efficiency of experimental methods, many computational approaches have been proposed to screen for new TFBSs. Moses *et al*. (6) and Wang *et al*. (7) reported TFBSs prediction algorithms based on phylogenetic data and multiple alignments of nucleotide sequences among different species. Anand *et al*. (8) proposed the prediction of TFBSs using an n-gram algorithm by analyzing the results of single base substitution experiments. Holloway *et al*. combined gene expression data with genomic sequence data to predict new DNA binding sites (9,10). Hoglund *et al*. (11) discussed the prospect of employing 3D structural information about protein–DNA complexes to improve the prediction of binding motifs. These methods have yet to address the problems of weak conservation in the upstream regions of genes and/or lack of protein structural data.

Known TFBSs are also used to identify transcription factor target genes; Position-Specific Scoring Matrices (PSSMs) are generally created to improve prediction performance (12,13). However, these approaches overlook the interdependence and variable distances among different bases (11), and the genetic contexts of TFBSs in the whole cell are ignored in the computational analysis. As a result, there is a very high frequency of false positive hits, especially when only one PSSM is applied, e.g. more than 30% of genes in the *A. thaliana* genome could be considered auxin response factor (ARF) targets because they all have the ARF-binding site 'TGTCTC' in their promoter regions. Several improvements have been made to reduce false positive predictions. Frith *et al*. (14)

*To whom correspondence should be addressed. Tel: +1 580 224 6725; Fax: +1 580 224 6692; Email: pzhao@noble.org

introduced genetic context (i.e. a set of functionally related PSSMs) into the model to fine-tune the PSSMs, thus improving the prediction performance. Suckow *et al*. (15) introduced variable gaps between two or more different motifs. Other similar attempts have also been reported, such as including a spacing rule between the TFs (16), and limiting the numbers of each contributing TF and combinations of TF positions (17). Unfortunately, these improvements must rely on known TFBSs, and this remains rather a sparse resource.

In this article, we propose a novel systematic computational approach to predicting transcription factor target genes (TFTGs) directly. Our approach does not necessarily predict new DNA binding sites, which other studies have shown to be difficult. Utilizing known binding sites and gene co-expression data, we modeled the prediction problem as a 'yes' or 'no' classification task and implemented the classifiers with support vector machines (SVMs). Here the 'feature generation, feature selection, feature integration' paradigm was followed to build the SVM classifiers. The promoter sequences of both target and non-target genes within 1000 bp from the transcription start site (TSS) were first profiled by a novel reverse-complementary position-sensitive (RCPS) *n*-gram profiling algorithm, which refers to the position of a known binding site or gene's TSS. Then, by applying measurements of the information gain (reduction in entropy), representative RCPS *n*-grams of positive and negative samples (target and non-target genes) were selected to create a vector space in which the promoter sequences were represented. Finally, these vectorial *n*-grams were fed to the SVMs to build prediction models.

We used the proposed approach to predict ARF target genes on the basis of published *A. thaliana* co-expression data (18) and obtained satisfactory results. Using 10-fold cross validation, the AUC value (area under curve) from our model reaches around 0.73, which is significantly higher than that from a random guess (0.5).

MATERIALS AND METHODS

Datasets

We extracted sequences up to 1000 bp upstream from gene TSS from *A. thaliana* genome sequences (ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/whole_chromosomes) by referring to gene locus data (TAIR6, 01/22/2004 release, ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR6_genome_release).

Auxin response factors (ARF) are capable of activating/suppressing the expression of primary auxin response genes at the transcriptional level by recognizing the specific 'TGTCTC' box binding site (19,20). Of the 7720 genes on the Affymetrix 8K AG Chip, 2787 with the 'TGTCTC' motif or its reverse complement 'GAGACA' in their upstream sequences were chosen as candidate ARF target genes. Goda *et al*. (18) reported that 637 genes are possibly affected by IAA/BL on the basis of a gene expression study using Affymetrix 8K AG-Chips. Therefore, we selected the 186 genes validated by Goda's experiments from the 2787 candidate genes as ARF-target genes and treated the remaining 2601 as ARF-non-target genes (highlighted in Table 1). Their 1000 bp upstream sequences were analyzed by the following steps. The sequences are available at our online supplementary page, http://bioinfo.noble.org/manuscript-support/TF_Supp/

Table 1. The classification of 7720 genes represented by probes on Affymetrix 8K AG-Chips

	Number of IAA/BL-affected genes (18)	Number of IAA/BL-unaffected genes	Sum
Number of genes with TGTCTC/GAGACA in their 1000 bp upstream regions	**186**	**2601**	2787
Number of genes without TGTCTC/GAGACA in their 1000 bp upstream regions	451	4482	4933
Sum	637	7083	7720

Reverse-Complementary Position-Sensitive *n*-gram algorithm

An *n*-gram is a subsequence of n letters from a given string (21). The *n*-gram profiling is a popular technique for converting natural language strings into histograms, i.e. generating statistics of all the n-grams occurring in a sequence stream. It has been applied to the recognition of splice sites in a genome (22). The current *n*-gram algorithm is generally little affected by the order in which different *n*-grams occur; in other words, the *n*-gram profile generated only includes *n*-gram frequencies. However, this is not necessarily valid for our TFTG prediction problem. According to existing models of the regulation of gene expression, the same DNA motif at different positions in the upstream region may exert different regulatory effects via a specific TF. In view of this, we extend the definition of an *n*-gram profile into a position-sensitive *n*-gram (PSNG) profile, formalized as follows.

Definition 1(PSNG): A PSNG of an *n*-length sequence $x = x_1,\ x_2 \ldots x_r \ldots x_{r+k} \ldots x_N$ relative to a *k*-length reference sequence $r = x_r \ldots x_{r+k}$ is defined as a subsequence of n continuous characters $s = x_i x_{i+1} \ldots x_{i+n-1}$ that satisfies $i > r + k$ or $i + n - 1 < r$, with corresponding relative distance

$$d = \begin{cases} i - (r+k), & \text{if } i > r+k \\ r - (i+n-1), & \text{if } r > i+n-1 \end{cases} \quad \mathbf{1}$$

denoted $\text{psng}(n) \equiv (s|d)$.

Definition 2 (PSNG Profile): The PSNG profile of an *n*-length sequence $x = x_1,\ x_2 \ldots x_r \ldots x_{r+k} \ldots x_N$ relative to a *k*-length reference sequence $r = x_r \ldots x_{r+k}$, denoted PSNP(*n*), is the enumeration of all possible

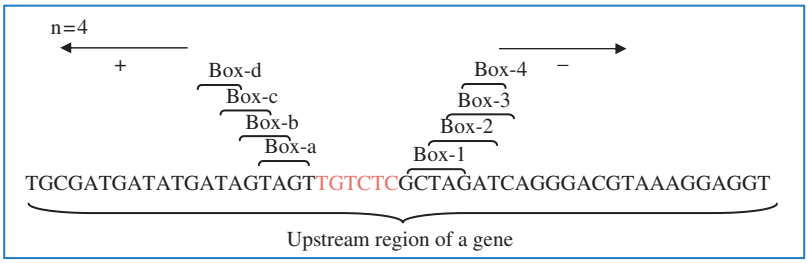

Figure 1. An example of constructing a reverse-complementary position-sensitive four-gram profile. The 'TGTCTC' highlighted in red is marked as the core motif. List of reverse-complementary position sensitive 4-grams of the given sequence: Box-a: TAGT/ACTA|+1; Box-1: GCTA/TAGC|-1; Box-b: GTAG/CTAC|+2; Box-2: CTAG/CTAG|-2; Box-c: AGTA/TACT|+3; Box-3: TAGA/TCTA|-3; Box-d: TAGT/ACTA|+4; Box-4: AGAT/ATCT|-4.

PSNG, psng(n), in the sequence. If the same numbers of n-grams are counted on the two flanks of a reference sequence, this number is represented as C, which is set to either '$r-1$' or '$N-r-k+1$' at most.

An example is given below. Given a sequence $x = \mathrm{CAACAAATGGAT}$, with the bold reference sequence AAT, its position-sensitive bi-gram profile is

$$\mathrm{PSNP}(2) = \{(\mathrm{CA}|-5), (\mathrm{AA}|-4), (\mathrm{AC}|-3), (\mathrm{CA}|-2), (\mathrm{GG}|3), (\mathrm{GA}|4), (\mathrm{AT}|5)\},$$

and its position-sensitive tri-gram profile is

$$\mathrm{PSNP}(3) = \{(\mathrm{CAA}|-5), (\mathrm{AAC}|-4), (\mathrm{ACA}|-3), (\mathrm{GGA}|3), (\mathrm{GAT}|4)\}.$$

With a PSNG profile, we are able to represent the upstream region of a gene by a limited set of subsequences with corresponding positional information. The reference sequence may simply be a shorter DNA sequence or a known DNA binding site. Here, 'TGTCTC' and its complementary sequence 'GAGACA' were the reference binding sites, since they were reported to be key motifs in the ARF-related regulatory model.

A potential binding site remains active over a range of nucleotides, not just at the point of the promoter region, because the 3D conformations of both protein and DNA are flexible. Thus, the same n-grams with different but neighboring positions may also have the same function. In view of this consideration, a parameter P, the position-sensitive factor, was introduced to control the position-sensitivity of PSNP profiling. The n-grams located in neighboring P bp regions are considered to be the same n-gram as long as they have identical sequences. The two examples above, the bi-gram and tri-gram profiles, were generated with $P = 1$.

Considering the base-pairing property of DNA double strands, a specific TF may bind to either strand of a ds-DNA molecule. We therefore extended the position-sensitive n-gram to a reverse-complementary position-sensitive n-gram, denoted rcpsng(n) $\equiv$ (s/rcs/d), e.g. (ATCG/CGAT|3). Figure 1 illustrates the process of building reverse-complementary position-sensitive four-grams from a given sequence.

Selection of representative features

The standard n-gram algorithm produces (in principle) 4^n possible n-grams when DNA sequences are profiled. When our reverse-complementary PSNG profiling algorithm is applied, the maximum possible number of n-grams would be $4^n \times C \times 2$ if we did not consider n-gram repetition. The number of n-grams is equivalent to the dimension of vector spaces in SVM algorithm. The demand of computation power will increase exponentially when the dimension of vector spaces increases. In our model building phase, our test indicated the optimal combination should include 4, 5, 6, 7, 8 and 9-grams. Thus, the total numbers of n-grams reaches $\sum_{n=4}^{9} 4^n \times C \times 2$, i.e. SVM needs to search $\sum_{n=4}^{9} 4^n \times C \times 2$ dimension of space. Therefore, it is not practical to train an SVM classifier with such a large volume of reverse-complementary PSNGs.

An effective feature selection process, which removes noise and outliers, would improve the prediction performance and reduce the computational cost. In our study, we used the measurement of information gain (IG) to select representative features from both positive and negative samples (target and non-target genes). The idea of IG is based on the evaluation of entropy in fuzzy datasets; it represents the change in entropy after a specific signal (a particular rcpsng n-gram) is observed (23).

Let S be the set of N DNA sequences being studied and T be the k classes of sequences; in particular, $T = \{TG, \overline{TG}\}$ for the target gene/non-target gene binary cases ($j = 2$, target and non-target) in our study. The entropy (expected information) of the set S is evaluated as

$$e(S) \equiv -\sum_{i=1}^{j} P(T_i,S) \cdot \log(P(T_i,S)) = -\left(\frac{N_{TG}}{N} \cdot \log\left(\frac{N_{TG}}{N}\right) + \frac{N_{\overline{TG}}}{N} \cdot \log\left(\frac{N_{\overline{TG}}}{N}\right)\right) \quad \mathbf{2}$$

where N_{TG} is the number of upstream sequences corresponding to target genes and $N_{\overline{TG}}$ is the number of upstream sequences corresponding to non-target genes.

With a conserved region c, the sequence set S has two distinct values: Sc are the sequences containing the conserved region c, and $S\bar{c}$ are the sequences that do not contain c. The entropy with respect to c is then given by

$$\begin{aligned} ec(S) &\equiv \sum_{i=1}^{v} P(S_i,S)\cdot e(S_i) \\ &= \frac{N_c}{N}\cdot e(S_c) + \frac{N_{\bar{c}}}{N}\cdot e(S_{\bar{c}}) \\ &= -\frac{N_c}{N}\cdot\left(\frac{N_{TG\cdot c}}{N_c}\cdot\log\left(\frac{N_{TG\cdot c}}{N_c}\right)+\frac{N_{\overline{TG}\cdot c}}{N_c}\cdot\log\left(\frac{N_{\overline{TG}\cdot c}}{N_c}\right)\right) \\ &\quad -\frac{N_{\bar{c}}}{N}\cdot\left(\frac{N_{TG\cdot \bar{c}}}{N_{\bar{c}}}\cdot\log\left(\frac{N_{TG\cdot \bar{c}}}{N_{\bar{c}}}\right)+\frac{N_{\overline{TG}\cdot \bar{c}}}{N_{\bar{c}}}\cdot\log\left(\frac{N_{\overline{TG}\cdot \bar{c}}}{N_{\bar{c}}}\right)\right) \end{aligned} \quad \mathbf{3}$$

where $N_c/N_{\bar{c}}$ are the numbers of sequences upstream of genes containing/not containing the conserved region c, respectively; $N_{TG\cdot c}/N_{TG\cdot \bar{c}}$ are the numbers of sequences upstream of target genes containing/not containing c, respectively; and $N_{\overline{TG}\cdot c}/N_{\overline{TG}\cdot \bar{c}}$ are the numbers of sequences upstream of non-target genes containing/not containing c, respectively.

The difference between ec and e is then used to define the information gain by the partitioning of S according to c:

$$IG(c) = e(S) - ec(S) \quad \mathbf{4}$$

Vector representation of DNA sequences

A higher IG value indicates greater information significance, and thus suggests that the corresponding n-gram is better able to represent an important feature of the sequence. Here, we chose the top K (K = 500, 1000, 1500 and 2000) n-grams to represent the features of the DNA sequences, thereby constructing a K-dimensional vector space. The upstream DNA sequences were then converted into the K-dimensional vector space according to their n-gram profiles. Figure 2 is an example showing the conversion of a sequence into vector format in terms of these featured n-grams.

Training and testing

A large number of classification algorithms have been applied successfully to text classification and information search tasks. In some cases, heuristic learning-based supported SVM perform better than other machine learning methods, because both positive and negative samples are utilized to train the models (24,25).

In our study, target genes served as positive samples and non-target genes as negative samples. Once the sequence upstream of a gene is represented in vector format, the TFTG prediction problem can be modeled as a two-class ('yes' or 'no') classification task. We adopted a two-class, linear-kernel SVM as a TFTG classifier that finds the boundary between the given target and non-target gene samples, and then makes predictions about unknown instances.

The training process of linear SVM classifier is summarized below.

Given a set of linearly separable vectors $Q = \{X_1, X_2, \ldots, X_N\}$, where X_i denotes the k-dimensional vector representing the corresponding training sequence and N = total number of input sequences, each belonging to one of the two classes labeled $y_i \in \{-1, +1\}$, (-1: *non-TG*, $+1$: *TG*), SVM seeks a separating hyperplane, $Y = W - X + b$, that divides Q into two parts, each containing vectors that have the same class label only by estimation on an optimal separating hyperplane (OSH) that has the maximal margin in both parts. This is done by minimizing $(1/2)\|W\|^2$, subject to $y_i(\mathrm{W}-\mathrm{X}_i) \geqslant 1$. Those vectors closest to the OSH are termed support vectors.

During classification, SVM makes decisions based on the OSH. It determines on which side of the OSH an unknown instance should be located and assigns the corresponding class label to the unknown instance. Since the SVM classifier is trained with data in two classes, i.e. target and non-target genes, SVM predicts whether a gene falls into the target gene class or the non-target gene class.

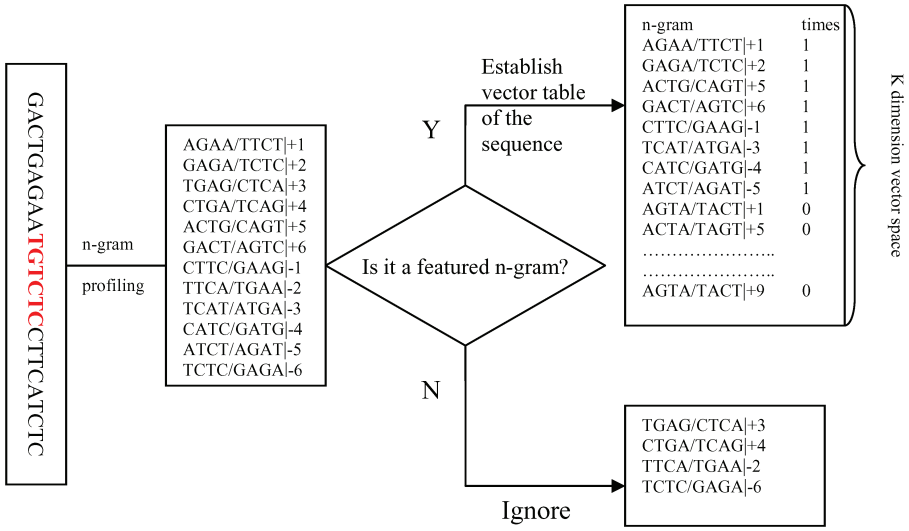

Figure 2. Vector representation of DNA sequence using the featured reverse-complementary position sensitive n-grams.

The SVM-light package (25) (http://svmlight.joachims.org/) was used to construct the SVM classifiers. We also compared the linear kernel with other kernels, such as polynomial and sigmoid kernels of the SVM. In our trial-and-error tests, the linear-kernel SVM yielded satisfactory prediction accuracy with relatively low computational cost.

Performance evaluation

We applied 10-fold cross-validation to estimate the model's performance. The entire dataset was randomly divided into 10 groups. Each time we chose a different group of sequences as the test group and the remaining nine as training groups; representative features were simultaneously extracted to create the corresponding vector space from the nine training groups. This procedure was repeated ten times to test all ten groups of sequences. The accuracy under a specific threshold value was evaluated on the basis of the following criteria:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad \mathbf{5}$$

where TP is the number of true positives (target genes with true positive predictions), TN the number of true negatives (non-target genes with true negative predictions), FP the number of false positives (non-target genes predicted as target genes) and FN the number of false negatives (target genes predicted as non-target genes). Because there were 14 times as many non-targets as target gene samples in the datasets, the high ratio of non-target genes may have generated a high accuracy value, even though the model performance was weak. Therefore, we introduced sensitivity and specificity to evaluate the models, which were computed as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \mathbf{6}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \mathbf{7}$$

The sensitivity is the ratio of target genes correctly predicted in the target gene dataset. The specificity is the ratio of non-target genes correctly predicted in the non-target gene dataset. Obviously, these two measurements generally vary when different SVM thresholds are applied. That is to say, by adjusting the threshold, we can obtain higher sensitivity and lower specificity, or vice versa.

We evaluated model performance by the area under the ROC curve (the receiver operating characteristic curve), the AUC value, which is independent of changes of threshold. The ROC curve indicates the change of sensitivity (true positive rate) versus specificity (true negative rate) under different thresholds.

We also tested model performance under different settings, e.g. different position sensitive factor P, length of n-gram n, number of n-grams C, number of representative n-grams K and SVM kernels.

RESULTS AND ANALYSIS

Position-sensitive n-gram algorithm

In our experiments, reverse-complementary PSNG profiles were generated with n set at 4–9. The numbers of n-grams (C) counted on each flank of the central motif 'TGTCTC/GAGACA' was set at 50, 75, 100, 125, 150 or 175. For $C = 100$, $n = 4$, the total number of reverse-complementary PSNG (with $P = 1$) was 27 038 in the upstream regions of both ARF-related and ARF-unrelated genes.

Selection of representative features

We applied IG algorithms to select representative features from the reverse-complementary PSNG generated, as detailed in the Materials and Methods section. Table 2 lists the top 10 four-grams with highest IG values and their frequencies in ARF-related and ARF-unrelated genes.

Classification performance

Our 10-fold cross-validation gave the AUC values of the SVM models, which were constructed from various combinations of n, C, P and K. In our experiments, the best model achieved an AUC value of 0.73, with $n = 4$–9, $C = 100$, $P = 100$ and $K = 1000$.

Figure 3 shows the raised ROC curve of the models generated. Since the ROC curve represents the relationship between sensitivity and specificity under different thresholds, weak models or completely random guesses would give a straight line at a −45 degree angle (i.e. AUC = 0.5), whereas our raised ROC curve (AUC = 0.73) indicates that our model has significant predictive power; ARF target and non-target genes can be discriminated. The detailed optimal models and their ranking are available at our online supplementary page, http://bioinfo.noble.org/manuscript-support/TF_Supp/.

Table 2. Top 10 significant four-grams screened by information gain value*

No.	Position-sensitive four-grams	Number of occurrences in 186 ARF-related genes	Number of occurrences in 2601 ARF-unrelated genes	Information gain value
1	AAAT/ATTT|-018	13	37	0.00146
2	AAAG/CTTT|-060	11	29	0.00132
3	AAGT/ACTT|-088	9	19	0.00128
4	TAGA/TCTA|-048	8	15	0.00124
5	CCCA/TGGG|-048	6	8	0.00114
6	CTAC/GTAG|+065	5	5	0.00109
7	ACTA/TAGT|-074	8	18	0.00109
8	ACAT/ATGT|-068	8	19	0.00104
9	ATTC/GAAT|-063	9	26	0.00099
10	TACA/TGTA|+024	7	15	0.00098

*The reverse-complementary position-sensitive 4-gram profile was generated from 1000 bp upstream regions of the 186 ARF-target genes and the 2601 ARF-non-target genes with position-sensitive factor $P = 1$.

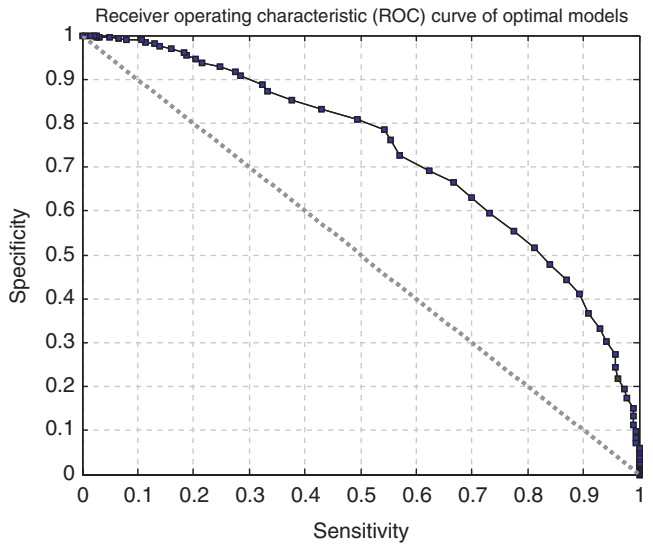

Figure 3. Receiver operating characteristic (ROC) curve of optimal models.

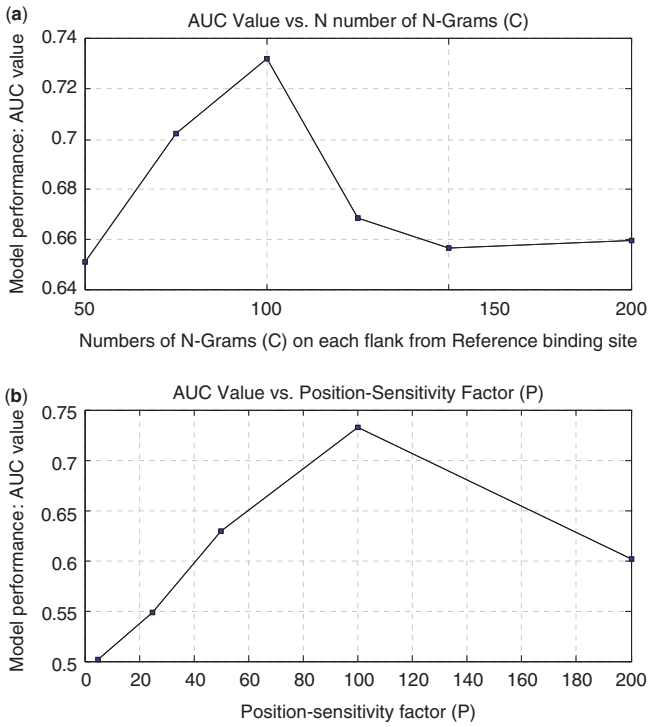

Figure 4. (**a**) AUC value versus *N* number of *n*-Grams (C) (**b**) AUC value versus position-sensitivity factor (P).

To understand the model performance better, we calculated accuracy, sensitivity and specificity (threshold = −1.0), which were equal to 0.69, 0.61 and 0.70, respectively.

Experimental parameters and model performance

To test how the number of n-grams, *C*, affects the performance of our models, we generated a set of SVM models with $C = 50$, 75, 100, 125, 150 and 175. Figure 4a shows an AUC versus *C* curve. The best AUC reaches 0.73 when $C = 100$, suggesting that *n*-grams in the region of 100 bp from the core motif (TGTCTC/GAGACA) may be important in the recognition and binding of TF.

We also evaluated the effect of position-sensitive factors (*P*) on model performance by varying *P*. Figure 4b shows an AUC versus *P*-plot. The optimal *P*-value was around 100; the model's AUC value declined at higher or lower *P*-values.

The optimal model was used to predict ARF target genes in *A. thaliana*. In summary, of the total 26 751 non-redundant TAIR6 genes, 12 559 were first retrieved by searching for the 'TGTCTC/GAGACA' binding motif in their promoter regions and then were sent to our SVM for prediction and ranking. We examined the top 1000 genes in the SVM output and found 172 known ARF target genes listed. We also manually curated the remaining 828 genes and found that 574 have been reported as possible auxin-related genes (the gene list is available at online supplementary page, http://bioinfo.noble.org/manuscript-support/TF_Supp/). This cross-validation result suggests that the model has potential for predicting new target genes of a transcription factor or its family.

DISCUSSION

Prediction of the target genes of transcription factors often suffers from high false positive rates (14). In our study, 2787 of the 7720 genes observed would be identified as possible ARF target genes if we only utilized the known 'TGTCTC/GAGACA' binding site for prediction. However, Affymetrix AG-chip co-expression analysis suggested that only 186 were true ARF target genes (18,19). That is to say, the false positive rate would reach 93% if we only employed the known TFBS 'TGTCTC/GAGACA' to screen target genes of ARF. Combinations of known TFBSs are commonly used to improve the prediction performance of TFTG (14–17). We compared our approach with a typical algorithm of this kind, cluster-buster (14). In our comparison, 20 transcription factor families were found from the known auxin-related genes, and then 13 PSSMs were created from these families and inputted into the cluster-buster program as matrix files. The cluster-buster program outputted a score for each promoter region in the dataset. Our analysis shows that the AUC value of the cluster-buster algorithm can only reach 0.51, which is obviously lower than our approach (data are available at our online supplementary page). The result of this comparison indicates that, owing to the information from co-expression analysis, our approach performs very well even though the associated binding motifs are unknown. The comparison also shows that the cluster-buster algorithm may be useful for identifying TFTGs when most of the associated binding motifs are known. Since gene co-expression data from microarray experiments are rapidly accumulating in public repositories, our approach holds promise for TFTG prediction.

Although *n*-gram algorithms have been applied to the analysis of biological sequences (26), relative distance

information is generally not considered in the n-gram. In this article, by defining novel reverse-complementary PSNG, we have introduced positional information into the standard *n*-gram algorithm for the first time. Here, a conserved binding site, i.e. 'TGTCTC', served as a reference point for relative positional information, further narrowing the scope of searches for potential interacting regions in promoters. The inclusion of positional information reflects the mechanism of interaction between transcription factor proteins and DNA, in which multiple transcription factors usually interact to recognize their corresponding binding sites. For example, ARF TFs possibly interact with members of the bZIP family, which are able to recognize the 'CCTCG' motif near 'TGTCTC' (19,20). The 3D structures of TF complexes indicate that their corresponding DNA binding sites are sensitive to relative distance (27). It is reasonable to include positional information when we profile *n*-grams. On the other hand, since protein/DNA 3D structures are flexible to some degree, motifs with slightly shifted positions may still have the same binding function. Therefore, positional information that is too loose or too stringent may affect our model's performance. Here, we introduced the *P*-factor to represent the sensitivity to position differentiation. Our results (Figure 4b) suggest that the best prediction results could be achieved when P was set to a small region ($P = 100$).

In the present work, we have employed IG to choose representative *n*-grams between positive and negative training samples. A high IG value represents a strong signal to noise ratio, which indicates that the corresponding n-gram is more valuable for training SVMs. IG determines the relative difference in occurrence of *n*-grams between two groups of sequences. Moreover, compared with another popular measurement, the χ^2- test (data not shown), the IG test is more capable of simultaneously filtering out *n*-grams with too low a frequency of occurrence. We considered low frequency *n*-grams as random noise even if their occurrence was significantly different between the two groups of sequences.

Comparative genomics and phylogenetic studies suggest that gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes (28). Our model should be applicable to other species, not just *A. thaliana*. However, in view of the variation in regions between genes, this claim needs to be further verified by biological experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful for the comments and suggestions raised by their coworkers, Drs Haiquan Li, Michael Udvardi, Rujing Chen and the anonymous reviewers of this article. Financial support for this project and funding to pay the Open Access publication charge was provided by Samuel Roberts Noble Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Chang,K., Elledge,S.J. and Hannon,G.J. (2006) Lessons from nature: microRNA-based shRNA libraries. *Nat. Methods*, **3**, 707–714.
2. Choi,H.-I., Hong,J.-H., Ha,J.-O., Kang,J.-Y. and Kim,S.Y. (2000) ABFs, a family of ABA-responsive element binding factors. *J. Biol. Chem.*, **275**, 1723–1730.
3. Xue,G.P. (2005) A CELD-fusion method for rapid determination of the DNA-binding sequence specificity of novel plant DNA-binding proteins. *Plant J.*, **41**, 638–649.
4. Moses,A.M., Pollard,D.A., Nix,D.A., Iyer,V.N., Li,X.-Y., Biggin,M.D. and Eisen,M.B. (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput. Biol.*, **2**, e130.
5. Guo,A., He,K., Liu,D., Bai,S., Gu,X., Wei,L. and Luo,J. (2005) DATF: a database of *Arabidopsis* transcription factors. *Bioinformatics*, **21**, 2568–2569.
6. Moses,A., Chiang,D., Pollard,D., Iyer,V. and Eisen,M. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, **5**, R98.
7. Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
8. Anand,A., Fogel,G., Tang,E.K. and Suganthan,P.N. (2006) Feature selection approach for quantitative prediction of transcriptional activities. *2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2006)*Toronto, Ontario, Canada, pp. 57–62.
9. Holloway,D.T., Kon,M. and DeLisi,C. (2005) Integrating genomic data to predict transcription factor binding. *Genome Inform. Ser. Workshop Genome Inform.*, **16**, 83–94.
10. Hampson,S., Kibler,D. and Baldi,P. (2002) Distribution patterns of over-represented k-mers in non-coding yeast DNA. *Bioinformatics*, **18**, 513–528.
11. Hoglund,A. and Kohlbacher,O. (2004) From sequence to structure and back again: approaches for predicting protein-DNA binding. *Proteome Sci.*, **2**, 3.
12. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
13. Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.
14. Frith,M.C., Li,M.C. and Weng,Z. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.
15. Suckow,M., Kisters-Woike,B. and Hollenberg,C.P. (1999) A novel feature of DNA recognition: a mutant Gcn4p bzip peptide with dual DNA binding specificities dependent of half-site spacing. *J. Mol. Biol.*, **286**, 983–987.
16. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.
17. Kielbasa,S.M., Korbel,J.O., Beule,D., Schuchhardt,J. and Herzel,H. (2001) Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics*, **17**, 1019–1026.
18. Goda,H., Sawa,S., Asami,T., Fujioka,S., Shimada,Y. and Yoshida,S. (2004) Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in *Arabidopsis*. *Plant Physiol.*, **134**, 1555–1573.
19. Liu,Z.B., Ulmasov,T., Shi,X., Hagen,G. and Guilfoyle,T.J. (1994) Soybean GH3 promoter contains multiple auxin-inducible elements. *Plant Cell*, **6**, 645–657.
20. Ulmasov,T., Liu,Z.B., Hagen,G. and Guilfoyle,T.J. (1995) Composite structure of auxin response elements. *Plant Cell*, **7**, 1611–1623.
21. Shannon,C. (1997) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

22. Sonnenburg,S., Ratsch,G. and Scholkopf,B. (2005) Large scale genomic sequence SVM classifiers. *Proceedings of the 22nd international conference on Machine learning*, ACM Press, Bonn, Germany. pp. 848–855.
23. Yang,Y. and Pedersen,J.P. (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. Yang,Y. and Liu,X. (1999) A Re-Examination of Text Categorization Methods. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, Berkeley, California, USA, pp. 42–49.
25. Joachims,T. (1999) Making large-Scale SVM Learning Practical. In Scholkopf, B., Burges, C., and Smola, A. (eds), *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, USA, pp. 41–56.
26. Tomovic,A., Janicic,P. and Keselj,V. (2006) n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Comput. Meth. Prog. Bio.*, **81**, 137–153.
27. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, 10.1186/gb-2000-1-1-reviews001.
28. Snel,B., van Noort,V. and Huynen,M.A. (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic Acids Res.*, **32**, 4725–4731.