A Systematic Computational Approach for Transcription Factor Target Gene Prediction

Ji He, Xinbin Dai and Xuechun Zhao‡
Bioinformatics Group, Plant Biology Division, The Samuel Roberts Noble Foundation
2510 Sam Noble Parkway, Ardmore, OK 73401, USA
{jhe, xdai, pzhao‡}@noble.org

‡ Author of correspondence.

***Abstract*— Computational prediction of transcription factor's binding sites and regulatory target genes has great value to the biological studies of cellular process. Existing practices either look into first-hand gene expression data which could be costly for large scale analysis, or apply statistical or heuristic learning methods to discover potential binding sites which have limited accuracy due to the complexity of the data. Based on well-studied information retrieval theories, this paper proposes a novel systematic approach for transcription factor target gene prediction. The key of the approach is to model the prediction problem as a classification task by representing the features of the sequential data into vector data points in a higher-order domain. The proposed approach has produced satisfactory results in our controlled experiment on Auxin Response Factor (ARF) target gene prediction in *Arabidopsis*.**

I. Introduction

A transcription factor (TF) is a protein that regulates gene's transcription by binding the DNA at a specific promoter, or enhancer region, or site. The interactions between TFs and the cognate upstream or promoter sequences primarily determine the establishment of spatial and temporal gene expression patterns. Briefly saying, there are three known classes of TFs, namely *general TFs* which interact with the core promoter region surrounding the transcription start sites of all class II genes, *upstream TFs* which bind to the upstream of transcription initiation site, and *inducible TFs* which work like upstream TFs but require activation or inhibition. Among them the latter two classes are of our research interests, as each upstream TF or inducible TF regulates a specific selection of genes, commonly referred to as target genes (TGs), under certain biological conditions. The study of the trio, namely TF, it's binding site (BS, or TFBS in some other literature), and the corresponding TGs, has great value to the understanding of various cellular processes.

As of today, all TFs reported in the literature were discovered and verified through biological experiments. On each known TF, even the identification of corresponding BSs/TGs is an experiment intensive and time consuming work. A good computational prediction algorithm could greatly improve the efficiency of the biological experiments by identifying a small collection of candidate BSs and/or TGs out of a large DNA sequence database. Recent research work on this topic is bringing a wider attention in the community. Software programs and tools for BS/TG prediction are emerging.

This paper proposes a novel, systematic approach for the prediction of primary TGs. The proposed paradigm is based on information retrieval theories. It takes use of existing knowledge on confirmed TFs and TGs, extracts the DNA sequence features into data points in a higher-order domain, and converts the TG prediction problem into a typical classification task which could be further resolved with various well-studied methods. Our experimental study on the prediction of the Auxin Response Factor (ARF) target genes in *Arabidopsis* has shown satisfactory results.

The rest of the paper is organized as following: Section II gives a brief review of existing BS/TG prediction algorithms. Section III presents our novel approach in details. Section IV reports the application of the approach to ARF TG prediction. Section V draws conclusions and proposes future work.

II. A Brief Typology Review of Existing Binding Site/Target Gene Prediction Algorithms

In general, there exist three categories of practices for computational prediction of BSs/TGs, summarized as gene expression based prediction, statistical search based prediction, and heuristic learning based prediction in our review.

1) Gene Expression Based Prediction: Methods in this category apply various data mining technologies on gene expressions, with a presumption that the TGs regulated by a TF usually have a close association to the corresponding TF in terms of the expression levels. Algorithms being applied for this purpose include various clustering based algorithms [1], [5] and Bayesian Networks [9]. Application of these practices into a new problem domain (such as a new species) requires the availability of first-hand gene expression data which could be costly and time consuming in the biological experiments.

2) Statistical Search Based Prediction: Methods in this category attempt to locate candidate BSs of known TFs through discovery of conserved motif patterns from a set of DNA sequences, then further make prediction on potential TGs according to the occurrence of candidate BSs. Among them, Expectation Maximization (EM) [2], Variable Memory Markov Model (VMM) [3] and Artificial Neural Networks (ANN) [22] are used to discover candidate BSs according to a pre-defined or dynamically generated statistical model.

1-4244-0623-4/06/$20.00 ©2006 IEEE

There are also reported work that studies the protein structure of interested regions [23]. It should be noted that since the prediction is usually based on unsupervised search which may not integrate the heuristic of known experimental results, the correctness of such prediction is not necessarily high, due to the complexity of the biological process.

3) Heuristic Learning Based Prediction: Methods in this category take advantage of known TF, BS and TG information in a well-studied problem domain to build a computational model that encodes the association of the trio, then apply the model to make prediction on a new problem domain, provided that the regulatory associations in the known domain is well conserved in the new domain. The Hidden Markov Models (HMM) [12] are typically used for this purpose. A HMM model is normally based on the multiple alignment of the upstream DNA sequence of the known TGs. Essentially, this is a task of heuristic learning from positive samples only, as the information of the genes that are known as not to be the TGs (*non-TG* in short in the rest of the paper) of a TF is not utilized by the HMM model. It is understandable that with a positive-sample-only learning paradigm, the system could produce either a high false-positive rate or a high false-negative rate, as there is no sound way of fine-tuning the learning parameters.

III. The Proposed Systematic Approach

This section proposes a novel approach for TG prediction. The proposed approach generally falls in the heuristic learning category as reviewed above. Unlike a HMM model that learns the knowledge of known TGs only, the proposed approach encodes the information of the upstream regions of both TGs and non-TGs. The approach can be understood as a hybrid of three typical information retrieval processes. First, a series of feature selection measures are used to highlight the conserved motif patterns (not necessarily BSs) that are statistically significant to separate the features of TGs and non-TGs. Then the upstream region of each gene is represented as a data point in a higher-order feature space with the selected motif patterns. Lastly a classifier is built to learn the differences between the features of the upstream regions of TGs and non-TGs. With such a classification model we are then able to make prediction on unknown genes, based on its upstream region. Details of the approach is given below.

A. Input Data and Output Results

The proposed approach requires below data to be available:

1) A known TF, or a set of known TFs that are closely related, in terms of biological function.
2) The sequences of the upstream region of the TF's TGs.
3) The sequences of the upstream region of some genes that are confirmed not to be the TF's TGs (non-TGs).
4) The sequence of the upstream region of the gene to be predicted.

The proposed approach makes prediction on whether a gene could be regulated by the known TF(s), i.e. whether it is a TG. It does not make an affirmative prediction on which motif is a candidate BS. However the selected features containing statistically significant motif patterns could be a good reference for BS identification in downstream analysis. The downstream analysis anyway is not the focus of this paper.

B. Reverse-Complementary N-Gram Profile (RCNP) for Representation of DNA Sequences

As first proposed by Shannon [18], an *n-gram* is a selection of n continuous characters from a given character stream. An *n-gram profile* refers to the statistics on the frequencies of all occurring n-grams in the stream [7]. N-gram profile is a popular technique in statistical natural language processing for converting sequential data into histogram data (in other words, data points in a vector space) [14], [4].

In many information retrieval applications, as long as two n-grams are literally different, their frequencies are counted independently. This however is not necessarily the best practice on DNA sequences. Considering a specific TF may bind on either a DNA strand or its complemental strand, a motif sequence and its reverse-complementary motif sequence are biologically identical in this sense, though they appear literally different in the DNA sequence. With this understanding, we extend the n-gram profile to a *reverse-complementary n-gram profile* (RCNP). The modification made on the RCNP is that, any occurring motif that is the reverse-complementary sequence of an existing motif will be counted towards the frequency increment of the existing motif.

With RCNP, we are able to represent the feature of the DNA sequence $\mathbf{x}$ using a set of M motifs with the corresponding frequencies, denoted as $\text{RCNP}(\mathbf{x}) = \{(\mathbf{s}_1 : f_1), (\mathbf{s}_2 : f_2), \cdots, (\mathbf{s}_M : f_M)\}$, $\mathbf{s}_i$ being a distinct n-length motif, and f_i being the frequency of the motif (and its equivalence reverse-complimentary motif) in the sequence.

C. Conserved Motif Search

Conserved motifs intuitively present the commons of various sequences, which also suggest potentially similar functions. Thus various BS/TG prediction algorithms start with conserved motif search. Typically, conserved motifs are discovered through multiple sequence alignment, which could be computationally complex and time consuming, especially when the input data set is large and the conserved motifs are relatively sparsely distributed.

With a RCNP representation of sequence features, it is possible to discover potential conserved motifs through a simple frequency count. Specifically, we adopted an *sequence-frequency filter* in our study. The sequence-frequency of a n-gram refers to the number of sequences that contains the specific motif (or its reverse-complemental motif). Understandably, a n-gram with a high sequence-frequency tends to be a conserved motif, and a n-gram with a low sequence-frequency tends to be a noise/outliner in the data set.

With the count of sequence-frequency of all n-grams across the whole sequence set, we may either select top N n-grams with the highest sequence-frequency, or those n-grams with sequence-frequencies exceeding a threshold. Thus we are able

to generate a collection of candidate conserved motifs, denoted as $\mathbf{C} = \{\mathbf{c}_i\}$, each $\mathbf{c}_i$ being a distinct motif.

D. Selection of Conserved Motif Features and Vector Representation of DNA Sequences

Feature selection have been an essential pre-processing technology in various pattern recognition applications. While it is possible to adopt all candidate conserved motifs as the bases of the feature space, an effective feature selection process removes noises and outliers, drastically improves the prediction accuracy and reduces the computational cost. In our study, we applied the *information gain* (IG) measure [26] for feature selection. The IG measure is based on the evaluation of a fuzzy data set's entropy, details giving as below.

Let $\mathbf{S}$ be the set of N sequences, $\mathbf{T}$ be the k "classes" of the sequences, particularly $\mathbf{T} = \{TG, \overline{TG}\}$ for the target gene/non-target gene binary cases in our study. The entropy (expected information) of the fuzzy set $\mathbf{S}$ is evaluated as

$$\begin{aligned} e(\mathbf{S}) &\equiv -\sum_{i=1}^{k} P(T_i, S) log(P(T_i, S)) \\ &= -(\tfrac{N_{TG}}{N} log(\tfrac{N_{TG}}{N}) + \tfrac{N_{\overline{TG}}}{N} log(\tfrac{N_{\overline{TG}}}{N})), \end{aligned} \tag{1}$$

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

With a motif $\mathbf{c}$, the sequence set $\mathbf{S}$ has correspondingly two distinct subsets, namely $\mathbf{S}_c$ being sequences containing the motif, and $\mathbf{S}_{\overline{c}}$ being sequences not containing the motif. The entropy with respect to $\mathbf{c}$ is then given by

$$\begin{aligned} ec(\mathbf{S}) &\equiv \sum_{i=1}^{v} P(S_i, S) e(S_i) \\ &= \tfrac{N_c}{N} e(\mathbf{S}_c) + \tfrac{N_{\overline{c}}}{N} e(\mathbf{S}_{\overline{c}}) \\ &= -\tfrac{N_c}{N}(\tfrac{N_{TG \cdot c}}{N_c} log(\tfrac{N_{TG \cdot c}}{N_c}) + \tfrac{N_{\overline{TG} \cdot c}}{N_c} log(\tfrac{N_{\overline{TG} \cdot c}}{N_c})) \\ &\quad -\tfrac{N_{\overline{c}}}{N}(\tfrac{N_{TG \cdot \overline{c}}}{N_{\overline{c}}} log(\tfrac{N_{TG \cdot \overline{c}}}{N_{\overline{c}}}) + \tfrac{N_{\overline{TG} \cdot \overline{c}}}{N_{\overline{c}}} log(\tfrac{N_{\overline{TG} \cdot \overline{c}}}{N_{\overline{c}}})), \end{aligned} \tag{2}$$

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

The difference between ec and e, termed information gain (IG), is then used to evaluate the information "gained" by the partitioning of $\mathbf{S}$ according to $\mathbf{c}$:

$$IG(\mathbf{c}) = e(\mathbf{S}) - ec(\mathbf{S}). \tag{3}$$

IG measures the significance of the observed information ec in contrast to the background e. Compared with some other widely used feature selection methods such as frequency based selection and CHI statistics selection [26], we consider IG more appropriate in our study, as it measures the distributions of both the positive samples (TGs) and negative samples (non-TGs).

A higher IG value indicates a higher information significance, and thus suggests a higher capability of the corresponding motif $\mathbf{c}$ in representing the feature of the sequence. With a simple ranking and threshold cut-off we are able to obtain a set of K features $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_K\}$, being the top K conserved motifs that has the highest IG metrics.

With reference to the selected feature set, we follow a standard practice to represent the feature of the DNA sequence into vector format, based on the occurrence frequency of the feature motif in the sequence. Specifically, with the feature set $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_K\}$, given the sequence $\mathbf{x}$ and its reverse-complementary n-gram profile RCNP($\mathbf{x}$) $= \{(\mathbf{s}_1 : f_1), (\mathbf{s}_2 : f_2), \cdots, (\mathbf{s}_M : f_M)\}$, its vector representation is $\mathbf{v} = (v_1, v_2, \ldots, v_i, \ldots, v_K)$, of which, each v_i is calculated according to

$$v_i = \begin{cases} f_j & \text{if } \exists j \in [1, M] \text{ that } \mathbf{s}_j = \mathbf{f}_i \text{ or } \mathbf{s}_j \text{ is the} \\ & \text{reverse-complemental sequence of } \mathbf{f}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

E. Building Classification Model

Once the features of the gene upstream regions are represented in vector format, without difficulty the TG prediction problem can be modeled into a two-class classification task. In order words, we are to build a classifier that learns the boundary (differences) between the given TG and non-TG samples and make prediction on unknown instances accordingly.

A large number of classification algorithms have been successfully applied to the information retrieval domain. These include, but not limited to, Naive Bayes (NB) [11], Bayesian Networks (BN) [16], [20], k Nearest Neighbor (kNN) [6], [8] and Linear List Square Fit (LLSF) [24] which are based on statistical optimization; Adaptive Resonance Associative Map (ARAM) [19] and Support Vector Machines (SVM) [15] which are based on heuristic learning. We adopted SVM in our study as it was reported to outperform other classification methods in some previous empirical studies [25], [13]. The concept of a two-class, linear-kernel SVM is summarized below.

Given a set of linearly separable points $\mathbf{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, each belonging to one of the two classes labeled as $y_i \in \{-1, +1\}$, a *separating hyper-plane* divides $\mathbf{S}$ into two sides, each containing points with the same class label only. The separating hyper-plane can be identified by the pair $(\mathbf{w}, b)$ that satisfies

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{5}$$

for any data point $\mathbf{x}$ on the hyper-plane, and

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \tag{6}$$

for any training sample $\mathbf{x}_i \in \mathbf{S}$. The goal of the SVM learning is to find the *optimal separating hyper-plane* (OSH) that has the maximal margin to both sides. This can be formularized as:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}||\mathbf{w}||^2 \\ &\text{subject to} && y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \end{aligned} \tag{7}$$

The points that are closest to the OSH are termed *support vectors* (Figure 1).

During classification, SVM makes decision based on the OSH. It finds out on which side of the OSH the unknown

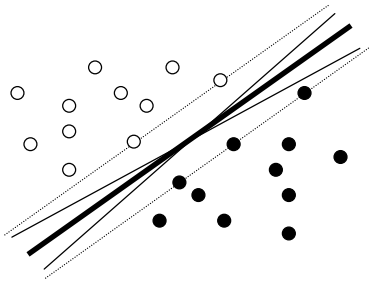

Fig. 1. The *separating hyper-planes* (the set of solid lines), the *optimal separating hyper-plane* (the bold solid line), and the *support vectors* (data points on the dashed lines) in the SVM learning model. The dashed lines identify the maximum margin.

instance is located and assign the class label to the unknown instance correspondingly. In our experiments, since the SVM classifier is trained with data in two classes, i.e. TG and non-TG, SVM will make prediction on whether a gene falls in the TG class or non-TG class.

There exists a large number of SVM variances, with different non-linear kernels for handling non-linearly separable data. Examples include the polynomial kernel, the radial basis function (RBF) kernel, and the sigmoid kernel. These kernels were also studied in our experiments [1].

IV. Prediction of Auxin Response Factor Target Genes in Arabidopsis

We applied the proposed systematic approach to the prediction of TGs of the Auxin Response Factors (ARF) in *Arabidopsis thaliana*. This section reports our experiment in details.

A. Data Set

The Auxin Response Factors (ARFs) are a family of transcription factors that regulate target genes after a plant is treated with auxin. Biological studies have shown that an ARF binds to the conserved motif TGTCTC under certain conditions [17]. However, not all genes whose upstream regions contain TGTCTC may be regulated by an ARF. Traditionally, whether a gene with TGTCTC in its upstream region may be regulated by ARF has to be verified through biological experiments. In 2004, Goda et. al. [10] treated *Arabidopsis thaliana* plants with auxin and brassinosteroid, and investigated the gene expressions using Affymetrix Genechip. Their experiments verified that among the 3137 genes whose upstream contains the motif TGTCTC, 263 are ARF target genes (TGs), and 2874 genes are not ARF target genes (non-TGs).

Our understanding to the problem are twofold. Firstly, besides the primary binding site TGTCTC, there may exist some other ARF "weak binding sites" that work together with TGTCTC during the gene regulatory process. On the other hand, some sequential characteristics, which in turn leads to various biochemical and protein structural features, may prevent an ARF to bind on the TGTCTC motif. Regardless the complexity of the biological process, we may build the computational classification model that learns the sequential differences between the reported TGs and non-TGs, and make prediction accordingly.

To prepare the primary sequence data, we obtained the accession IDs of the reported 3137 genes, then identified their location information from TAIR6 Arabidopsis Information Resource (available via ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR6_genome_release). The location information of each gene includes its corresponding chromosome ID, the transcription start point and end point. With these, we extracted the upstream region of each gene through a search from the *Arabidopsis thaliana* whole genome sequence (available via ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/whole_chromosomes). We limited the upstream regions to be within 1000 bases as previous studies have shown most binding sites fall within this frame.

B. Experiment

1) Data Processing: Following the proposed approach, we started with building reverse-complemental n-gram profile (RCNP) of each gene's upstream region. Noting the fact that TGTCTC is the primary binding site, and considering the relative stable structure of DNA helix, we believed the characteristics of the sequential region nearby the primary binding site mostly affect the transcriptional process. Thus we limit our feature extraction in a small window that evenly cover the surround regions of the primary binding site. We tested different window sizes and found that a relatively small size, 100 bases (i.e. 50 upstream bases and 50 downstream bases) yielded satisfactory results. This in turn proved our hypothesis.

We used a combination of $n = 4, 5, 6, 7$ (i.e. four-grams, five-grams, six-grams, and seven-grams) as they appeared to give empirically good results in our trial-and-error studies. 10901 unique n-grams were generated from the 3137 sequences. Among them, we selected the top 5000 candidate conserved motifs that have the highest sequence-frequency. Out of these candidate conserved motifs, we selected top 500 motifs as the feature set based on IG. They include 19 four-grams, 47 five-grams, 121 six-grams, and 313 seven-grams. Table I lists the top 20 motifs in terms of IG value.

The selected motif features were further used to convert each sequence into vector format, according to Equation 4. At last we obtained a set of 3137 500-dimensional vectors, among which, 263 were labeled as the positive class (i.e. TG) and 2874 were labeled as the negative class (i.e. non-TG) for our classification experiments.

2) Performance Evaluation: Following a standard procedure, we carried out the leave-one-out cross-validation (LOOCV) to examine the efficacy of the proposed approach. On a set of N instances, the leave-one-out test replicates the experiment for N times. Each time it uses $N - 1$ instances as the training set while leaving one different instance as the testing sample. It is widely considered as an effective practice

[1]The SVM classifier used in our experiments was downloaded via http://svmlight.joachims.org/.

TABLE I

THE TOP 20 N-GRAMS BEING SELECTED AS THE FEATURE MOTIFS, ACCORDING TO INFORMATION GAIN (IG) SCORE. N_{TG} AND $N_{\overline{TG}}$ CORRESPOND TO THE NUMBER OF TARGET GENE SEQUENCES AND NON-TARGET GENE SEQUENCES THAT CONTAIN THE THE MOTIF.

Rank	Motif/Reverse-complement	N_{TG}	$N_{\overline{TG}}$	IG Score
1	CGGAG/CTCCG	6	239	0.001119329
2	ACTCAAG/CTTGAGT	12	31	0.000990069
3	TATTAA/TTAATA	53	337	0.000953703
4	ACCGG/CCGGT	3	160	0.000932376
5	TATTAAA/TTTAATA	28	139	0.000900589
6	GTCTTAA/TTAAGAC	12	34	0.000894641
7	GTAGA/TCTAC	60	409	0.000866178
8	TGAGAAA/TTTCTCA	2	130	0.000851824
9	CTGACAC/GTGTCAG	8	16	0.000834962
10	CACTTAG/CTAAGTG	8	16	0.000834962
11	AAAGT/ACTTT	115	949	0.000823801
12	GGGCTGA/TCAGCCC	5	5	0.000821921
13	ATTAA/TTAAT	113	930	0.000815191
14	CAGATC/GATCTG	1	101	0.000806264
15	AGTTTG/CAAACT	11	288	0.000804118
16	GTCTAA/TTAGAC	22	103	0.000787142
17	AATATTC/GAATATT	16	63	0.000769756
18	AAGCTC/GAGCTT	3	142	0.000754201
19	ACTAAAT/ATTTAGT	16	64	0.000750307
20	ACGAGG/CCTCGT	0	61	0.000747079

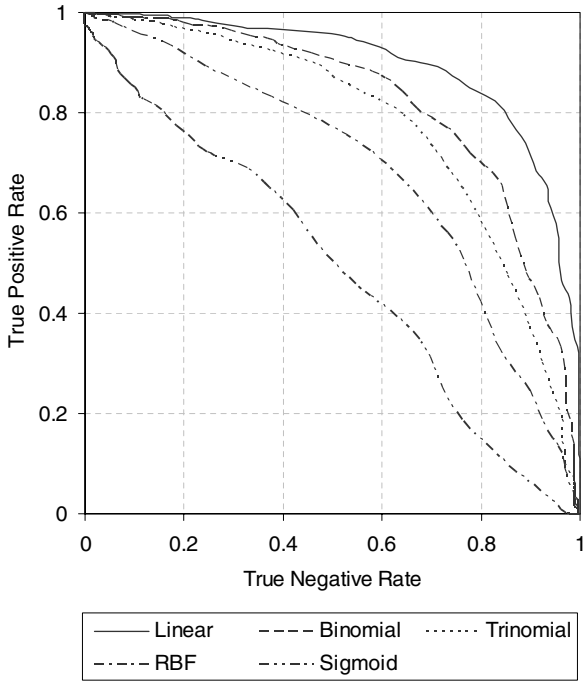

Fig. 2. The receiver operating characteristic (ROC) curves of the leave-one-out cross-validations (LOOCVs) on Auxin Response Factors (ARF) target gene (TG) prediction, using different support vector machine (SVM) kernels, namely linear, binomial, trinomial, radial basis function (RBF), and sigmoid.

to evaluate the generalized learning capability of an algorithm on a specific data set, with least statistical bias.

Based on the N prediction scores made by the leave-one-out test, we may draw various statistics. We firstly adopted the *receiver operating characteristic* (ROC) measure to evaluate our system's efficacy. A ROC curve plots the correlation of a classifier's true-positive rate (sensitivity) and true-negative rate (specificity, or the complemental false-positive rate in some other studies) according to the change of the prediction cut-off threshold. It is widely used to evaluate the classifier's systematic efficacy at all cut-offs. The area under the curve (AUC) is commonly used to summarize the results into a single score for the ease of comparison. In addition, since ARF TGs are our primary interest, we adopted a commonly used set of measures to evaluate the correctness of the prediction on the positive class (i.e. TG) with a default cut-off threshold 0.0. These measures are *precision* (P), *recall* (R), and F_1 [21], [25]. Let $\mathbf{S}$ be the positive sample set (actual positive) in cross-validation and $\mathbf{T}$ be the samples being predicted as positive by the classifier (positive prediction), the definitions of precision (P), recall (R) and F_1 are as follows:

$$P = \frac{|\mathbf{S} \bigcap \mathbf{T}|}{|\mathbf{T}|}, \tag{8}$$

$$R = \frac{|\mathbf{S} \bigcap \mathbf{T}|}{|\mathbf{S}|}, \tag{9}$$

and

$$F_1 = \frac{2|\mathbf{S} \bigcap \mathbf{T}|}{|\mathbf{S}| + |\mathbf{T}|}, \tag{10}$$

where the norm $|.|$ denotes the size of a data set; the intersection of the two sets $\mathbf{S} \bigcap \mathbf{T}$ identifies the correct prediction, i.e. those samples that are predicted as positive and are actual positive (true positive). It should be noted that the F_1 measure is intended to incorporate P and R measures into one for cross-comparison, and essentially

$$F_1 = \frac{2RP}{R+P}. \tag{11}$$

3) Results and Discussions: Figure 2 illustrates the ROC curves of the LOOCVs with different SVM kernels, namely linear, polynomial (specifically binomial and trinomial), RBF and sigmoid. The AUC scores and corresponding P, R, and $F1$ scores with a default cut-off threshold 0.0 are reported in Table II. Figure 2 shows that the linear SVM kernel outperformed other kernels in our controlled experiments, indicating that with the proposed feature selection and extraction process, the sequential data are best linearly separative. The binomial and trinomial kernels performed slightly worse than the linear kernel. The RBF kernel and sigmoid kernel performed worst, being close to a random prediction. It however should be noted that this reflects the mismatch between the RBF and sigmoid kernels and the actual data distribution, rather than the unsuccess of the proposed approach.

Figure 3 depicts the venn diagram on the prediction of the linear kernel SVM with a default cut-off 0.0. Vertical strip area denotes the TG training samples (i.e. actual TGs). Horizonal strip area denotes the predicted TGs. Their intersection, shown as in grids, denotes the true positive predictions. The rest, blank area, denotes the true negative predictions. Among the 635 positive predictions, 204 are correct; whereas only 59 out

TABLE II

THE AUXIN RESPONSE FACTORS (ARF) TARGET GENE (TG) PREDICTION RESULTS USING DIFFERENT SVM KERNELS, EVALUATED WITH AREA UNDER CURVE (AUC), AS WELL AS PRECISION (P), RECALL (R) AND F_1 ON TARGET GENES WITH DEFAULT CUT-OFF THRESHOLD 0.0.

Kernel	AUC Score	P	R	F_1
Linear	0.8949	0.3213	0.7757	0.4543
Binomial	0.8232	0.3937	0.3802	0.3868
Trinomial	0.7757	0.3778	0.1939	0.2563
RBF	0.6840	0.2857	0.0076	0.0148
Sigmoid	0.4883	0.0530	0.0646	0.0582

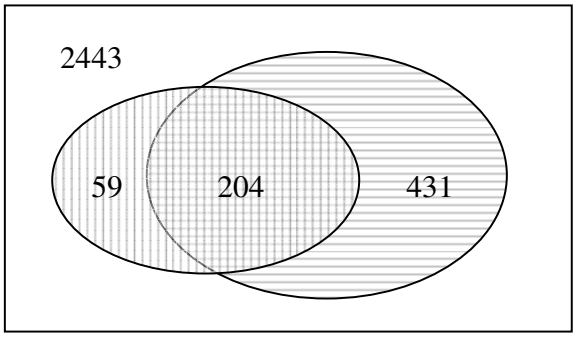

Fig. 3. The venn diagram of the leave-one-out cross-validation (LOOCV) on Auxin Response Factors (ARF) target gene (TG) prediction, using the linear support vector machine (SVM) kernel, with a default cut-off threshold 0.0. Vertical strip area denotes the actual target genes. Horizonal strip area denotes the predicted target genes.

of 263 actual target genes are missed in the prediction. This yields a 32.13% precision rate, a 77.57% recall rate, and a 0.4543 $F1$ score.

The overall results are quite encouraging to us, considering our studies reported here are quite preliminary and the parameters during each step are not optimized. One may understand the prediction results in this way: if we focus our biological experiments on the verification of the 635 genes that are predicted as ARF TGs, roughly one out of three inspected genes will be an actual ARF TG; and the experiments will only miss about 22% actual TGs. Thus the computational approach could greatly improve the efficiency of biological experiments. The satisfactory results suggest two facts. Firstly, the human inspected data are of relatively high consistency. In other words, there is a high correlation among the sequence upstream regions of the ARF TGs. And there really exist underlying differences between them and those of non-ARF TGs. Secondly, though the underlying differences are not notably viewable to the human, the proposed systematic approach indeed has the capability of identifying these differences through supervised learning. On the other hand, compared to prior reports on other information retrieval applications such as text classification [25], [13], the performance scores generated in our experiments are relatively low. This reflects the complexity of the biological process and the challenges to the computational prediction work, if merely based on sequential information.

V. SUMMARIES AND FUTURE WORK

It has been widely known that DNA sequences could be no more complicated than human's natural languages. Literally, they could be considered as in a language with limited alphabet (G, C, A, and T). Many well-studied natural language processing (NLP) and information retrieval (IR) algorithms have been successfully applied to sequence analysis. The proposed systematic approach incorporates a series of well-studied IR theories, namely n-gram profile for feature representation, information gain (IG) for feature selection, and support vector machine (SVM) for classification. While they appear well-known to the computational community, we took the initiative to extend and apply extend them to the challenging transcription factor target gene prediction problem. It should be noted that while n-gram profile has been applied for binding site prediction by some researchers, there is no reported work like ours that predicts target genes through classification model. By converting the prediction problem into classification task, our methods not only learns the common of positive samples (i.e. upstream regions of target genes), but also learns the differences between the positive samples and negative samples. In computational research, this practice has been shown to generally outperforms positive-sample-only learning methods such as the HMM model. Lastly while most importantly, this paper essentially proposes a systematic approach for representing DNA sequence features as data points in a higher-order vector space. With this, various vector-space based technologies may be applied to solve sequential problems. Hence, this approach is of great referential value to other sequence analysis studies.

The proposed approach contains a set of dynamics, such as the length of the gene upstream region, the length of the n-gram motif (n), the number of candidate conserved motifs being selected as the features (hence the dimension of the extracted feature vectors), and the various parameters in response to SVM's learning. These dynamics are not optimized due to the time constraint in preparing this manuscript. The optimization of these dynamics typically involves a manually trial-and-error or a computationally simulated annealing process, which remains a future work to us. Despite of this, the proposed approach has performed satisfactorily in the reported application, suggesting its promising efficacy.

Following this paradigm, our future work also include the comparison of different feature selection methods and different classifiers, as well as the cross comparison between the proposed approach and other target gene prediction algorithms reported in the literature, such as HMM.

Finally, as a side note, based on the reported data on *Arabidopsis thaliana*, we will apply the learnt classification model to predict ARF TGs in other related plant species, and further verify the predictions through biological experiments. These results are to be published in the future.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Samuel Roberts Noble Foundation for the financial support on the project. The

authors are grateful to the comments and suggestions raised by the anonymous reviewers of this manuscript.

REFERENCES

[1] S. Akutsu, T. Kuhara, O. Maruyama, and S. Minyano. Indentification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In *Proceedings of 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[2] T. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.

[3] Yonatan Bilu, Michal Linial, Noam Slonim, and Naftali Tishby. Locating transcription factors binding sites using a variable memory markov model. In *ISMB*, 2002.

[4] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994.

[5] T. Chen, V. Filkov, and S. Skiena. Identifying gene regulatory networks from experimental data. In *Proceedings of 3rd Annual RECOMB*, 1999.

[6] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[7] Grace Crowder and Charles Nicholas. Using statistical properties of text to create metadata. In *Proceedings of the 1st IEEE Metadata Conference*, 1996.

[8] B.V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Las Alamitos, California, 1991.

[9] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. In *RECOMB*, pages 127–135, 2000.

[10] Hideki Goda, Shinichiro Sawa, Tadao Asami, Shozo Fujioka, Yukihisa Shimada, and Shigeo Yoshida. Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in arabidopsis. *Plant Physiol.*, 134:1555–1573, 2004.

[11] I.J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, 1965.

[12] John Goutsias. A hidden markov model for transcriptional regulation in single cells. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(1):57–71, 2006.

[13] Ji He, Ah-Hwee Tan, and Chew-Lim Tan. On machine learning methods for chinese document classification. *Applied Intelligence*, 18:311–322, 2003.

[14] S. Huffman. Acquaintance: Language-independent document categorization by n-grams. In *TREC 4 Proceedings*, 1996.

[15] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, Springer, 1998.

[16] Wai Lam, Kon F. Low, and Chao Y. Ho. Using a Bayesian network induction approach for text categorization. In Martha E. Pollack, editor, *Proceedings of IJCAI-97, 15th International Joint Conference on Artificial Intelligence*, pages 745–750, Nagoya, JP, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[17] Zhan-Bin Liu, Cretchen Hagen, and Tom J. Cuilfoyle. A g-box-binding protein from soybean binds to the E1 auxin-response element in the soybean CH3 promoter and contains a proline-rich repression domain. *Plant Physiol*, 115:397–407, 1997.

[18] Claude Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[19] A.H. Tan. Adaptive Resonance Associative Map. *Neural Networks*, 8(3):437–446, 1995.

[20] Ben Taskar, Pieter Abbeel, and Daphne Koller. Discriminative probabilistic models of relational data. In *Proceedings of UAI-02, 18th Conference on Uncertainty in Artificial Intelligence*, pages 485–492, Edmonton, CA, 2002. Morgan Kaufmann Publishers, San Francisco, US.

[21] C.J. van Rijsbergen. *Information Retieval*. Butterworths, London, 1979.

[22] C.T. Workman and G.D. Stormo. ANN-Spec: A method for discoverying transcription factor binding sites with improved specificity. In *Pacific Symposium on Biocomputing*, volume 5, pages 464–475, 2000.

[23] Mandel-Gutfreund Y, Baron A, and Margalit H. A structure-based approach for prediction of protein binding sites in gene upstream regions. In *Pac Symp Biocomput*, pages 139–50, 2001.

[24] Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[25] Y. Yang and X. Liu. A re-examination of text categorization methods. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, 1999.

[26] Y. Yang and J.P. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 412–420, 1997.