Sequence analysis

dPattern: transcription factor binding site (TFBS) discovery in human genome using a discriminative pattern analysis

Seung-Hee Bae¹, Haixu Tang², Jing Wu⁴, Jun Xie⁴ and Sun Kim²,3,*

¹Department of Computer Science, ²School of Informatics, ³Center for Genomics and Bioinformatics, Indiana University – Bloomington, IN 47404 and ⁴Department of Statistics, Purdue University, West Lafayette, IN, USA

Received on January 9, 2007; revised on May 20, 2007; accepted on May 21, 2007

Advance Access publication June 5, 2007

Associate Editor: Burkhand Rost

ABSTRACT

Motivation: Transcription factor binding sites (TFBSs) are typically short in length, thus search with a profile model from known TFBSs produces many false positives. When combined with additional information, gene expression data in this article, sensitivity and specificity of TFBS search can be improved significantly.

Results: By modifying our previous REFINEMENT approach, we developed dPattern that searches for occurrences of TFBSs in the promoter regions of up/down regulated or random genes.

Availability: http://platcom.org/projects/dpattern

Contact: sun.kim@acm.org or sunkim2@indiana.edu

1 INTRODUCTION

The problem we are considering is how to increase the specificity of transcription factor binding site (TFBS) prediction by incorporating gene expression information. From gene expression data with an experimental design to study the effect of a certain condition, e.g. interferon stimulated genes, we can obtain up/down-regulated genes. Promoters in the up-regulated genes, say by interferon, constitute a positive sequence set where more TFBS occurrences are expected. Promoters in the down-regulated or random genes constitute a negative sequence set where less TFBS occurrences are expected. Given a set of known TFBSs that we want to search for, e.g. interferon stimulated regulatory elements (ISREs), we have developed a computational method, called REFINEMENT (Tsukahara et al., 2006), that refines the original motif model iteratively to discriminate occurrences in the positive set from those in the negative set. This discriminative method was successful in reducing false positives in the promoter regions of up-down regulated genes. Here we developed a new program, called dPattern and a web server. The main differences between REFINEMENT and dPattern are: (1) dPattern uses Patser (Hertz and Stormo, 1999) instead of a profile hidden Markov model, assuming no gaps in the TFBS, which improved the computational efficiency significantly so that a general web server can be provided, (2) dPattern uses a mixture model approach. It first computes a background model using human promoter sequences identified from alignments between human and mouse/rat/chimpanzee from the UCSC web site at http://genome.ucsc.edu and combines the background model and a TFBS model (user input) to produce an initial mixture model that is general enough to predict de novo TFBSs and (3) dPattern also use a simple rank-based discriminative model refinement procedure.

Input to dPattern are: (1) a set of known TFBSs, (2) a set of Ensembl gene identifiers for up-regulated genes and (3) a set of Ensembl gene identifiers for down-regulated or random genes. Then dPattern searches for occurrences of the TFBS in the promoter regions of the genes.

TFBS search is performed in two steps: mixture model generation and iterative model refinement.

2 MIXTURE MODEL GENERATION

The goal is to make the motif model general enough to detect TFBSs that are not included in the set of known TFBSs. Given a set of m known TFBSs of length n, \(\{x_1, \ldots, x_m\}\), we can build a profile \(P_y\) for \(1 \leq i \leq n\) and \(y \in \{A, T, G, C\}\). Using human promoter sequences identified from the alignments at UCSC, a kth order Markov model \(K_{y|x_1,\ldots,x_k}\) is computed. Then a mixture model \(M_y\) is computed by combining \(P_y\) and \(K_{y|x_1,\ldots,x_k}\). The values for a column i of \(M_y\) are computed as follows.

1. \((1) (1 \leq i \leq k)\)
   
   A background stationary-profile \(S_y\) of length k is computed using the weighted character frequencies of a prefix of length k of each known TFBS \(x_i\). The weight of each prefix is determined by its stationary probability from \(K_{y|x_1,\ldots,x_k}\) using \(\text{estim}_m\), a tool from the seq++ package (Miele et al., 2005). Then
   
   \[ M_{iy} = a P_{iy} + (1 - a) S_{iy} \]
   
   where, \(a\) is a preset value between 0 and 1 (0.75 by default) and all three matrices, \(M_{iy}\), \(P_{iy}\), and \(S_{iy}\), are of \(4 \times k\).

2. \((k+1 \leq i \leq n)\)
   
   The probability of \(y, B_y\), is computed by averaging the Markov model probabilities \(K_{y|x_1,\ldots,x_k}\) for m sequences, \(x_1,\ldots,x_m\). Then
   
   \[ M_{iy} = a P_{iy} + (1 - a) B_{iy} \]

*To whom correspondence should be addressed.
The Markov model $K_{x_{i\ldots j}}$ is computed using a set of alignments between human and mouse/rat/chimpanzee from the UCSC web site, and a background model $B$ is computed as described earlier. The background model $B$ is from conserved promoter regions between human and a reference genome, say mouse genome, thus we expect that $B$ performs better for TFBS prediction than a simple background model with pseudo counts since TFBSs are typically searched for in the promoter regions. The order $k$ of the Markov model is a parameter to be specified by the user. We recommend using the smallest value of $k$ that discriminates between the alignment of promoters of up-regulated genes (a positive set) and the alignment of promoters of down-regulated or random genes (a negative set); users can test and see the model parameter differences using gnuplot (http://gnuplot.org) at our web server.

After constructing a mixture model, dPattern searches for statistically significant occurrences of the new model in the promoter regions of two target sequence sets using Patser program (Hertz and Stormo, 1999) with a prior of each nucleotide being equally likely. The system also makes a Web-Logo (Crooks et al., 2004) with respect to each occurrences set.

3 MODEL REFINEMENT PROCEDURE

The goal is to refine the model $M_{iy}$ to a new model $M'_{iy}$ so that the number of the model occurrences in the negative sequence set is reduced while the number of occurrences in the positive sequence set is increased. We modified the refinement procedure of REFINEMENT (Tsukahara et al., 2006) as follows.

1. Ranking candidates: dPattern ranks occurrences of $M_{iy}$ using the score from Patser. We use a negative ranking strategy to leverage the discriminative information (up versus down/random genes), i.e. the profile model of negative sequences is used to rank binding site occurrences in the positive sequence set. More formally, let $O^+$ be occurrences in the positive set and $O^-$ be occurrences in the negative set. We build a new profile using $O^-$ and rank $O^+$ based on the scores by the profile, thus we use a positive model to collect candidates and use a negative model to rank the candidates.

2. Selecting candidates: given a list of TFBS occurrences in the positive sequence set, the user can select candidates either manually using the checkbox on the web page or simply specify top $N$ sequences.

3. Constructing new model: a new refined model $M'_{iy}$ is constructed with the selected candidate sequences. Then TFBS search is performed again with the new model and the search result is displayed on the web with a summary and logos for the model and its occurrences in the positive and negative sequence sets. The entire steps can be repeated as many times as the user wants.

Figure 1 shows an experiment with ISREs in human genome after a single refinement step. dPattern increased the number of predicted TFBSs in the genes up-regulated by interferon while reducing the number of predicted TFBSs in the down-regulated genes. As a result, the occurrence ratio (up to down) increased significantly to 4.33 from 3.14.

4 COMPARISON WITH REFINEMENT

For the comparison, we used REFINEMENT web server and the dPattern web server with a Markov model of order $k = 6$. We tried one iteration step for both REFINEMENT and dPattern web servers, and then we measured running times of both web servers and counted the number of predicted ISRE elements in both the up-regulated promoter region sequence set and down-regulated promoter region sequence set.

For the running time, REFINEMENT web server took 116 s to get the result with one refinement iteration step, and the dPattern web server took only 33 s to get the predicted result. In terms of predicted binding sites, REFINEMENT web server predicted 87 ISRE sites in the promoter regions of genes up-regulated by interferon and 68 sites in the promoter regions of genes down-regulated by interferon. On the other hand, dPattern predicted 91 sites in the up-regulated promoter region sequences and only 21 sites in the down-regulated promoter region sequence set.

5 MODEL REFINEMENT VERSUS DE NOVO PREDICTION

dPattern is to refine an original model that was constructed using known transcription binding sites. Prediction of de novo binding sites is a challenging problem which is different from the refinement problem in this article. Recently, there has been interesting development on the de novo binding site problem using discriminative approaches (Kawada and Sakakibara, 2005; Sinha, 2006). Use of discriminative approaches can make the binding site prediction algorithms less sensitive to model parameter settings as shown in Sinha, (2006).

We ran DIPS, Discriminative PWM Search, (Sinha, 2006) on our ISRE sequence data to measure how good the de novo binding site prediction would be. DIPS was not able to find the ISRE bind sites correctly after running more than 2 days on a Linux machine (3.0 GHz Intel processor and 4 GB main memory). The main reason is that our sequence data (each of 7 kb in length) is significantly longer than the data used in Sinha (2006).

There is a growing evidence that some binding sites can be very far away, say >100 kb, from the transcription start site. For example, a genome wide analysis of binding sites of estrogen receptor, the master transcriptional regulator of breast cancer phenotype and the archetype of a molecular therapeutic target, showed that binding sites were detected as far as 206 kb from TSS (Carroll et al., 2006).

Thus, a discriminative method that combines de novo binding site prediction and model refinement techniques may be an interesting approach to deal with binding sites that are far away from TSS.
6 CONCLUSION

dPattern is a web server where users can search for occurrences of a specific TFBS in the human genome with gene expression data. The accuracy of TFBS search, which often produces many false positives and negatives, can be improved significantly by incorporating gene expression data. Since each gene expression experiment is designed with a specific goal, dPattern will be a valuable tool for TFBS search with gene expression data.

ACKNOWLEDGEMENTS

This work is supported by a US National Science Foundation Career DBI-0237901 grant (to Kim), US National Institutes of Health National Cancer Institute Grant CA113001 (to Kim) and a Collaboration in Life Sciences and Informatics Research (CLSIR) grant from the state of Indiana (to Xie, Wu, Tang and Kim).

Conflict of Interest: none declared.

REFERENCES