Bioinformatics and Biosystems 2007, Vol. 2, No. 2, pp. 52-57

Discovering cis-regulatory motifs by combining multiple predictors

Hye-Shik Chang, Kyuwoong Hwang and Dongsup Kim

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology

Abstract

The computational discovery of transcription factor binding site is one of the important tools in the genetic and genomic analysis. Rough prediction of gene regulation network and finding possible co-regulated genes are typical applications of the technique. Countless motif-discovery algorithms have been proposed for the past years. However, there is no dominant algorithm yet. Each algorithm does not give enough accuracy without extensive information. In this paper, we explore the possibility of combining multiple algorithms for the one integrated result in order to improve the performance and the convenience of researchers. Moreover, we apply new high order information that is reorganized from the set of basis predictions to the final prediction.

Keywords: cis-regulatory element, motif discovery, transcription factor binding site

From the beginning of the computational analysis of biological data, discovering the biologically meaningful over-expressed pattern has been a key technique in the field. The patterns mean variety of functions depending on its basis. It can be a structural motif for protein structures or a transcription factor binding site for DNA sequences. The transcription factor binding site is a conserved and stable pattern found in genes that are regulated by a transcription factor, usually a protein. Because most induction and suppression mechanisms in gene regulation are controlled by protein's binding to a binding site, it is very important to identify the site in order to understand the biological mechanism. (D'haeseleer, 2006a)

Many algorithms with different approaches have been introduced for the discovery of binding sites. They utilize various types of evidences from the statistical analysis of sequence elements to the evolutionary information from orthologous genes. The ideas of such algorithms are distinctive even if the information source is same. The most straightforward approach is an enumeration. The approach exhaustively evaluates all the possible motifs. It sometime uses IUPAC codes for 2- or 3-nucleotides represent a position while some algorithms find all consensus sequences with an allowed maximum mismatches. Another popular approach is a deterministic optimization. MEME (Bailey and Elkan, 1995) is the most well-known implementation in this discipline. It repeats two steps called expectation and maximization and it converges into a local minimum after all. EM algorithm generates a position weight matrix (PWM) which is used for calculating probability of specific DNA sequence being a motif. Gibbs sampling is another optimization approach that uses stochastic sampling for subsequences. Starting from randomly selected sites, it focuses on the model of sample with high probability after the number of iterations. While many algorithms can be categorized in the three approaches mentioned above, there are still few alternative approaches that have different features such as graphical optimization (Reddy et al, 2007) and phylogenetic footprinting (Blanchette et al, 2003).

Most existing algorithms suggest a number of predicted motifs. However, their accuracy has been quite low without manual interpretation of an experienced computational biologist. Experts suggest to researchers that they should try many of different algorithms and interpret their biological meanings (D'haeseleer, 2006b). The major role of computational prediction is to save a labor and resource in experimental works. The discovery algorithm has to suggest a list of probable motifs without an extensive training and work on a computer console. The problem can be resolved by combining existing algorithms into an integrated set. Machine learning technology has been adopted by many applications in the analysis of biological data. In this paper, we propose a new method discovering transcription factor binding sites that proposes probable motifs based on predictions of existing algorithms. Moreover, integrating different algorithms makes high order data available such as a position preference among the algorithms and sequences. To confirm the usefulness of those data, we evaluate them with our own methods.

Methods

In this section, we outline our method which feeds data from various discovery algorithms to a learning algorithm. Robinson *et al.* (2006) once described a method for the similar purpose with taking quite different way from ours not only in the vectorization of algorithm results but also in representing high order data.

1. Algorithm

Performance of the combining algorithm can be maximized when the input features contain different information each other. We adopted few algorithms from different features to take the advantage of orthogonality. However, we excluded phylogenetic footprinting algorithms because they require evolutionary information which cannot be provided in a certain case. Finally, six well-known algorithms are chosen to be evaluated as candidates of the prediction vector. (Table 1)

To train the learning algorithms with the sequences, we need to annotate whether it is a binding site or not at each residue of sequence points. Because an input vector consists of a consistent number of values, we had to transform the results from each algorithm into a value set. As every algorithm gives different type of data and different number of motifs, same numbers of transformation methods were designed for the individual algorithms. Details of the methods are described later.

Binding site annotation which serves as a correct answer in the training was calculated from the quality of binding value in the original experimental data because high quality of binding affinity means that the binding site is more likely to be close to the real motif. The annotation BA is defined as

$$BA_{ij} = M_{ij}\left(\frac{Q_{ij} - \mu_Q}{3\sigma_Q} + 1\right)$$

where i is a sequence, j is a position in the sequence, M_{ij} is 1 if BA_{ij} is a binding site or 0 otherwise, Q_{ij} is a binding quality of the site, μ_Q and σ_Q are an average and a standard deviation of all binding sites in the training set, respectively. This formula results 0 for non-binding sites and 1 for average quality binding sites.

Each algorithm predicts with just low accuracy. Many weak predictors with different random features may make the integration worse. We examined correlations between algorithms to select a subset that would predict with reasonable accuracy. Table 2 shows Pearson correlation coefficients between algorithm predictions and *BA*. No algorithm showed quite low correlation coefficient to *BA*. We chose AlignACE, MEME, MotifSampler and Weeder as input sources and excluded Projection and SeSiMCMC because they have low correlation to *BA* although they have relatively strong correlation to MEME.

2. High Order Data

If a transcription factor binds to a gene but does not bind to an orthologous gene, it will be very helpful to the prediction because their difference is a strong evidence of binding sites. It is relatively easy to perform an additional experiment that tests activity of a transcription factor for an orthologous gene. We added the prediction results from inactive orthologous genes into the set of input vector. It will work as a strong negative signal. We split the input vector set to three schemes; one includes results from positive samples on-

Table 1. List of evaluated existing algorithms

Algorithm	Type	Reference
AlignACE	Gibbs Sampling	Hughes *et al.* 2000
MEME	Deterministic	Bailey and Elkan, 1995
MotifSampler	Gibbs Sampling	Thijs *et al.* 2001
Projection	Random Projection	Buhler, 2003
SeSiMCMC	Gibbs Sampling	Favorov *et al.* 2005
Weeder	Enumerative	Pavesi *et al.* 2001

Table 2. Pearson correlation coefficients between each algorithms and binding site annotation (*BA*) in the training set of *Homo sapiens* data. Relatively high correlations are emboldened.

Algorithm	BA	AA	ME	MS	PR	SS	W
BA		.041	.060	.009	.017	.058	.053
AlignACE	.041		.015	.001	.001	.062	.068
MEME	.060	.015		.011	**.417**	**.278**	.115
MotifSampler	.009	.001	.011		.010	.011	.013
Projection	.017	.001	**.417**	.010		**.195**	.062
SeSiMCMC	.058	.062	**.278**	.011	**.195**		.076
Weeder	.053	.068	.115	.013	.062	.076	

ly ("posonly"), another includes results from positive and negative samples ("posneg"), and the last includes results from positive samples only and positive and negative samples together ("composite"). In this experiment, we chose sequences that had score ranging from 10 bit to 35 bit using BLASTn (Altschul *et al*. 1990).

Some kind of motif has highly conserved distance from the translation start site because of the physical property of DNA structure or a transcription factor. (Vardhanabhuti *et al*. 2007; Tomovic and Oakeley, 2007) If several algorithms predict a position as a binding site in most sequences, the position is likely a position-conserved binding site with relatively high confidence. This is information that is not represented in the calculated input vector because it just expresses information for a sequence. The collocated prediction score *CP* is defined as

$$CP_j = \frac{\sum_{a \in algs} \sum_{s \in seq} (PP_{asj} - PC_{asj})}{N_{seq}}$$

where *alg* is a set of all algorithms, *seq* is a set of sequences and PP_{asj} and PC_{asj} are 1 if the algorithm *a* predicts that position *j* on sequence *s* is in a binding site in "posonly" and "composite" predictions respectively or 0 otherwise. The Pearson correlation coefficient between *BA* and *CP* in the training set was 0.15; it is not bad as a weak classifier.

3. Dataset

We used TRANSFAC Pro version 11.1 (Matys *et al*. 2006) for both of the training set and testing set. Gene sequences are retrieved from the EMBL database and binding sites that lacks EMBL database reference are abandoned. 4,368 binding sites were survived after removing binding sites that are too distant (>2,000nt) from translation start site or sequences with duplicated homologs. Then sequences do not overlap at least 300nt each other are removed for the alignment at translation start site. The final dataset for the experiment included 1,470 binding sites on 1,086 gene sequences for 163 transcription factors. One half of transcription factors were used for training and the second half were used in testing. Total sequence length was 1,643,190 and 32,636 binding site points were annotated in positive class. BLAST was run to pick negative samples with default options for DNA sequence.

4. Input Vectors

All algorithms were set to find multiple candidates of motifs with exactly 8nt long.

AlignACE: Although we ran AlignACE just with default option, it gave many possible motifs with variable length. We searched the given motifs in the sequences, and assigned scores for all positions found that consist the predicted binding site.

MEME: For each set of sequences, MEME was run to find 25 distinct motifs. Each motif represented as a position specific scoring matrix (PSSM) is applied to every windowed positions on all the sequences. For each position, maximum score in the window was taken to represent the position as a value. A normalized value of -log(*E-value*) is multiplied to the product of probabilities for the final score.

MotifSampler: MotifSampler found ten motifs each for set of sequences. The score from MotifSampler was applied as an input vector without any arbitrary calculation because they provide just a IUPAC code and its score for replacements.

Projection Genome Toolkit: Projection motif finder was ran to find ten different motifs. Instead of motif models, it gives instances found from given sequences. We assigned same scores to the positions that are predicted as a binding site by Projection.

SeSiMCMC: Like Projection Genome Toolkit, SeSiMCMC does not build motif models. Exactly same method was applied for SeSiMCMC.

Weeder: Weeder was run to find three distinct motifs. Found motifs are applied to the all the possible positions in the sequences. log(*4 - differences*) were calculated as score and Weeder confidence score is multiplied. As it did in MEME, maximum score in the window was taken to make continuous blocks.

5. Sampling and Training

Our bare data was unbalanced dataset because most positions are not binding sites. We undersampled majority class non-binding site positions in order to balance for training. The majority class samples were randomly selected as many as binding site positions. No oversampling technique was adopted.

Training was performed using LIBSVM (Chang and Lin, 2001) for SVM methods and Orange (Demsar *et al.* 2004) for random forest methods. SVM was trained using v-SVR core with RBF kernel (v=0.8). Random forest was grown with 500 trees for the training, and replacement was allowed.

Results

1. Individual Predictions

CASE 1. NF-AT2 is a human nuclear factor of activated T-cells. (Northrop *et al.* 1994) TRANSFAC reports 14 binding sites in 10 gene sequences that NF-AT2 can bind. Two binding sites on PTGS2 (COX2) have quality of 6. Many mouse genes and interleukin-2 (IL-2) have low quality to NF-AT2 binding. The best prediction for the transcription factor was done by MEME but it is 7th, which mean that we need to test at least 7 positions experimentally to confirm the true site. Although the other algorithms did not choose the true site early, they gave scores higher than average. As a result, our combined method could pick the real site in a high rank. As shown in Table 3, SVM-composite reported the site as the third and RF-composite reported the site as the fourth probable binding site.

CASE 2. C/EBPβ is a rat liver-enriched transcriptional activator. (Lichtsteiner *et al.* 1986) According to TRANSFAC, 55 human genes and 20 rat genes are recorded to be active for the transcription factor. All the algorithms failed to identify the binding site in human alcohol dehydrogenase 2 (ADH2). However, very adequate negative samples were chosen thanks to rich number of sequences. Predictions with positive samples recorded medium score for the true binding site but all composite predictions said "no" to that site. As a result, random forest could pick the site. Nonetheless, SVM seems that it had been overfit in training phase due to the numerous false negative samples.

Table 3. Number of false positives above the first true positives in the predictions of NF-AT2 and C/EBPβ.

Algorithm	NF-AT2	C/EBPβ	HNF-4α
AlignACE	37	91	43
MEME	7	71	26
MotifSampler	32	92	85
Weeder	18	82	37
SVM-composite	**3**	76	5
RF-composite	4	**1**	**3**

CASE 3. HNF-4α is an activator of liver-specific genes. (Sladek *et al.* 1990) 13 human genes are known as active to HNF-4α. A cytochrome P450 gene (CYP2A6) that contains many subsequences repeats that are similar to motif. Due to the trap, many algorithms were confused where is the binding site; most of them failed to identify the true site. Our combined methods could report the true site in the high rank (3rd for RF, 5th for SVM) because of the high score of the collocated prediction.

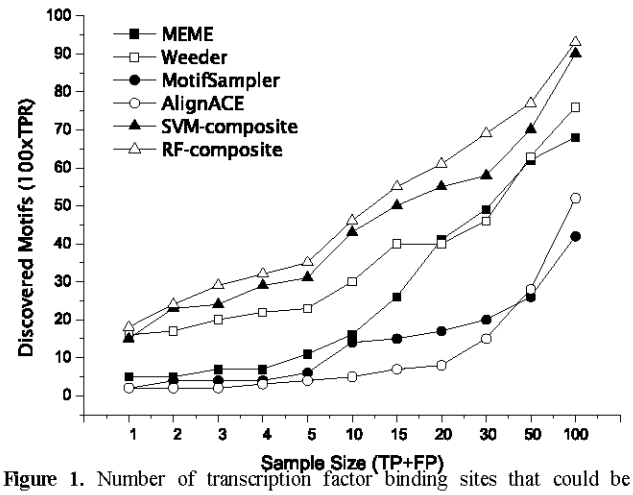

Figure 1. Number of transcription factor binding sites that could be found in each allowed number of predictions. Y axis represents 100-fold of the true prediction rate (TPR) which can be calculated as TP / (TP + FN). Both combined methods require fewer samples to be validated by the experiments in order to discover the motifs than the best of existing algorithm.

2. Analysis of Results

Our combined methods predicted with similar or better accuracy in major portion of cases, still there were few cases that better result was got from the single algorithm. Figure 1 shows the number of discovered true binding sites varying over the allowed number of false positives. It shows that combined methods steadily outperformed the best single algorithm in every size of allowed candidates. We compared the various options of the combined algorithm (data not shown). Random forest looks more appropriate for the combining on the whole because of its discriminative power for many weak classifiers. For the selection of samples, predictions from composite gives slightly better result compared to the other options.

Figure 2 shows statistical distributions of the area under ROC curves for each algorithm. MEME and Weeder had a

big gap between their best and worst performance. AlignACE showed a little bit better than random distribution after all. MotifSampler found few motifs very successfully while most of the rest were completely failed to discover. However, combined methods appeared superior over the existing algorithms. The big margin between combined methods and the others originates in the difference of their coverage. A single algorithm tended to miss several samples completely. Conversely, combined methods predicted reasonable candidates even when they could not figure out correct binding sites.

Discussion

We showed that the combined method of multiple discovery algorithms was useful to not only expand the coverage of prediction but also automate the interpretation of results from multiple algorithms. It gave fine predictions without loss of performance of the original algorithms.

Addition to the combining, we also evaluated two new high order information sources. One is the position conservation over a set of predictions, and the other is using two simultaneous motif discoveries with samples that are known as inactive for a transcription factor. Both of them did important roles in several examples.

Our approach has its limit on its being a combined method. It cannot be improved much without introducing new algorithm the input vector. It needs training for every time genome background changes that is not a cheap operation. Still it will be worth for certain cases because of its wide coverage and acceptable performance even without an expertise in the computational tools.

We avoided the information sources which require a limited condition such as phylogenetic tree or 3D structure of a transcription factor in this research. Such information sources can not be easily operated as a single independent algorithm. However, combined method will easily adopt the information as we successfully applied and merged multiple algorithms and additional information together. Future work will include the exploration of another information sources and the better handling of the input vectors and classifiers.

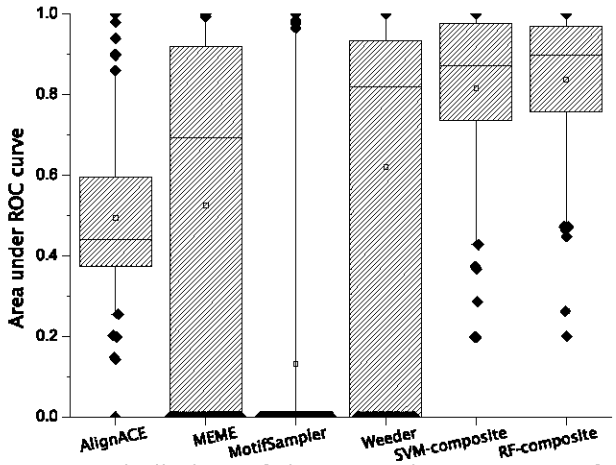

Figure 2. Distributions of the area under ROC curves for each algorithm. Boxes cover 50% of each set. Small open circles in the boxes represent the average values. Horizontal lines in the middle of boxes denote the medians. Vertical lines are drawn between 5% and 95% of each set. Diamond dots outside stand for outlier values which are in 5% from the top or bottom. Our combined methods, SVM-composite and RF-composite, show superior performance in this statistics thanks to their wider coverage.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol*. 215(3):403-410

[2] Bailey, T. L. & Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME. in *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. 21-29

[3] Blanchette, M. & Tompa, M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res* 31: 3840-3842

[4] Buhler, J. (2003) Provably sensitive indexing strategies for biosequence similarity search. *J Comput Biol*. 10(3/4):399-418

[5] Chang, C.-C., Lin, C.-J. (2001) LIBSVM: A library for support vector machines.

[6] Demsar, J., Zupan, B., Leban, G. (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, *White Paper*, Faculty of Computer and Information Science, University of Ljubljana.

[7] D'haeseleer, P. (2006) What are DNA sequence motifs? *Nature* 24, 423-425

[8] D'haeseleer, P. (2006) How does DNA sequence motif discovery work? *Nature* 24, 959-961

[9] Favorov, A. V., Gelfand, M. S., Gerasimova, A. V.,

Ravcheev, D. A., Mironov, A. A., Makeev, V. J. (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* 21(10):2240-5

[10] Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol.* 10;296(5) 1205-14

[11] Lichtsteiner, S., Wuarin J., Schibler U. (1987) The interplay of DNA-binding proteins on the promoter of the mouse albumin gene. *Cell.* 51:963-973

[12] Matys, V., Meinhardt, T., Prüß M, Reuter, I., Schacherer, F. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34, 108-110

[13] Northrop J. P., Ho S. N., Chen L., Thomas D. J., Timmerman L. A., Nolan G. P., Admon A., Crabtree G. R. (1994) NF-AT components define a family of transcription factors targeted in T-cell activation. *Nature.* 369:497-502

[14] Pavesi, G., Mauri, G., Pesole, G. (2001) An algorithm for finding signals of unknown length in unaligned DNA sequences. *Bioinformatics* 17, S207-S214

[16] Reddy, T. E., DeLisi, C. & Shakhnovich, B. E. (2007) Binding site graphs: a new graph theoretical framework for prediction of transcription factor binding sites. *PLoS Comput Biol* 3(5): e90

[17] Robinson, M., Sun, Y., Boekhorst, R. T., Kaye, P., Adams, R. & Davey, N. (2006) Improving computational predictions of cis-regulatory binding sites. in *Pacific Symposium on Biocomputing*

[18] Sladek, F. M., Zhong, W., Lai, E., Darnell jr, J. E. (1990) RT Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev.* 4:2353-2365

[19] Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., Moreau, Y. (2001) A higher order background model improves the detection of regulatory elements by Gibbs Sampling. *Bioinformatics* 17(12),1113-1122

[20] Tomovic, A. & Oakeley, E. J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics.* 23(8):933-941

[21] Vardhanabhuti, S., Wang, J. & Hannenhalli, S. (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res.* 1-11