A Novel Transcription Factor Binding Sites Prediction Approach

Author
Affiliation
Address
email

Abstract

Transcription factors (TFs) and their DNA binding motifs, called transcription factor binding sites (TFBSs) play important roles in most biological processes. However, the list for TFBSs still remains largely unknown. Machine learning approaches have been intensively applied to predict TFBSs. In this paper, a novel prediction approach has been presented based on Markov Chain Monte Carlo (MCMC) method and latest discovery of TF-TFBS co-evolution. By defining and solving a problem modified from conventional TFBSs prediction problem, the paper provides a new way to predict TFBSs for poorly characterized TFs, which has been previously considered difficult. The performance of the proposed approach has been evaluated on real biological data.

1 Introduction

TFs are proteins which can regulate gene expression. TFs carry out their function through interacting with specific TFBSs [1]. TFBSs are DNA sequences that the TFs bind to. Different TFs bind to different TFBSs. And normally, a TF can bind to a set of TFBSs. So TFBSs are also called DNA motifs. The interaction of TFs and TFBSs regulates gene expression by promoting or repressing the speed and efficiency of gene transcription (**Figure1a**).

Using high-throughput experimental techniques, biologists have identified thousands of TFs. However, the corresponding TFBSs are largely unknown due to the experimental scale bottleneck. For example, over 1,200 of human and mouse TFs are annotated in the TFCat database [2] and the number is still increasing. However, only less than half of these TFs have binding sites mapped and annotated in public databases. Considering the possible combinations of DNA sequences, it is impossible for wet lab experiments to identify all the TFBSs. Currently, most of the TFBSs are predicted computationally, and only a small proportion will be validated by biological experiments.

This paper presents a novel machine learning approach based on the latest discovery of co-evolutionary relationship between TFs and TFBSs [3]. As the key component of this framework, the prediction approach uses a MCMC method as its core. This new approach is motivated to deal with the problem that predicting TFBSs for unknown or poorly studied TFs, which previous studies do not have solutions.

1.1 Related work

Since TFBSs are DNA sequence fragments composed by four types of bases {A, T, G, C}, the TFBSs prediction problem has been treated as motif finding problem by machine learning

scientists and bioinformaticians. In general, the current methods for TFBSs identification are designed to solve following motif finding problem:

Definition1*: Given a set S with N sequences, where each of the sequences is generated from the alphabet {A, T, G, C}, find out the subsequences set S' with N' sequences, where the subsequences have identical length l and share highest similarity with each other.*

From Definition1, we can see that the current methods assume that the DNA sequences which contain TFBS motifs have been given. And the goal is to identify those motifs. Some representative studies are briefly reviewed here. The TFBS prediction has been modeled as motif discovery problem in deterministic constraints methods, and solved by employing approximate string matching algorithms [4] However, due the diversity in TFBSs of a TF, over-predicting problem has been introduced and a large amount of false positives exist in this kind of method. Stochastic local search strategies, especially its representative genetic algorithms (GA) have also been applied in TFBSs finding to deal with local optima problems [5, 6]. However, GA suffered from its low speed when the problem size grows, and did not show significant improvement compared with some other methods such as heuristic based Gibbs sampling [7] and Hidden Markov Model [8]. Currently, the state-of-art machine learning method in TFBSs discovery is MEME [9] which implements an expectation-maximization (EM) algorithm as its core. Given the DNA sequences that are known to be bound by a TF, this EM will iteratively find the locations of the potential TFBSs fragments.

A very recent research in bioinformatics has revealed the co-evolutionary relationship between TFs and TFBSs [3]. It shows that during the evolution, a TF and its corresponding TFBSs are changing accordingly in order to maintain their interaction. For example, if there is a change (like, A to C) in the TFBS happened, and the TF does not change, then the TF may not be able to bind to that TFBS any more. For some important interactions in cell, such change may weaken TF-TFBSs interaction and thus result in abnormalities to the organisms. So organisms have developed a mechanism to mutation such interaction. As a result, we could observe the evolutions of many TFs and their TFBSs are significantly correlated across species. This paper also presents a way to measure the correlation by computing the correlation value of the evolutionary matrix (i.e. similarity matrix) of TFs and the matrix of TFBSs. This research provides the possibility of predicting TFBSs for poorly studied TF by looking at its neighbor TFs in the evolutionary tree.

1.2 Contribution

According to Definition1, previous methods can not deal with poorly studied TF since we do not have any prior TFBSs information. While based on the notion of co-evolution, the novel approach proposed here considers a modified problem according to Definition1 and then it could make prediction for poorly studied TFs.

Definition2*: Given the initial set S with N sequences where the sequences have identical length l and generated from the alphabet {A, T, G, C}, and a target function f(**S**), find out the optimal set S' with N' sequences, where the sequences maximize f(**S**).*

The target function here is the correlation between the evolutions of TFs and TFBSs as defined in [3]. A MCMC algorithm is used to find the optimal TFBSs set.

The remainder of this paper is organized as following: it begins with an introduction of the proposed MCMC method (section2), then it describes the evaluation approaches for this new method and the results (section3), discussion are provided at last (section4).

2 Method

In this section, I will present how to use MCMC to optimize the TFBS set through maximizing the co-evolution between TFs and TFBSs, which is the key part of the proposed prediction framework.

2.1 Representation

Since one TF could bind to a number of TFBSs, the representation of these TFBSs is usually a matrix form. This matrix has l rows and 4 columns. Each row represents one position in the TFBS motifs, and each column represents one DNA base {A, T, G, C}. The TFBSs lengths ls could be different for different TFs. But normally l equals to 8 or 9. In this paper, l is considered as constant for all the TFs. The value of each cell a_{ij} in the matrix represents the probability that the i^{th} position in the TFBS is the j^{th} DNA base. Such matrix is called position weight matrix (PWM).

2.2 Data source

The TF data (name, sequence) could be obtained from Uniprot database, which is the most comprehensive protein database currently. The TFBS data (name, sequence, PWM) could be obtained from Jaspar database, which is the most widely used free TFBS database. Section 3 will describe the specific dataset used in this paper for evaluation.

2.3 Prediction approach

Firstly, as showed in **Figure1b**, for a poorly studied TF X, its closet *k* neighbor TFs (with known TFBSs) on the evolutionary tree are obtained (i.e. the TFs share the highest sequence similarities with X) using common bioinformatics tool ClustalW. According to [3], it is assumed that during the evolution, when a TF changes, its TFBSs change accordingly. So the TFBSs of X may generated from its neighbors' TFBSs.

1. These TFs compose our TF set Y, $Y=\{TF_1,TF_2,..\ TF_k\}$.
2. For all the TFs in Y, their similarity scores (computed by ClustalW) with X are treated as weights, $w=\{w_1,\ w_2,..\ w_k\}$.

Besides, since the TFBSs sets of all TFs in set Y are known, we write this as $S=\{S_1,S_2,..\ S_k\}$, each S_i contains identical N DNA sequences.

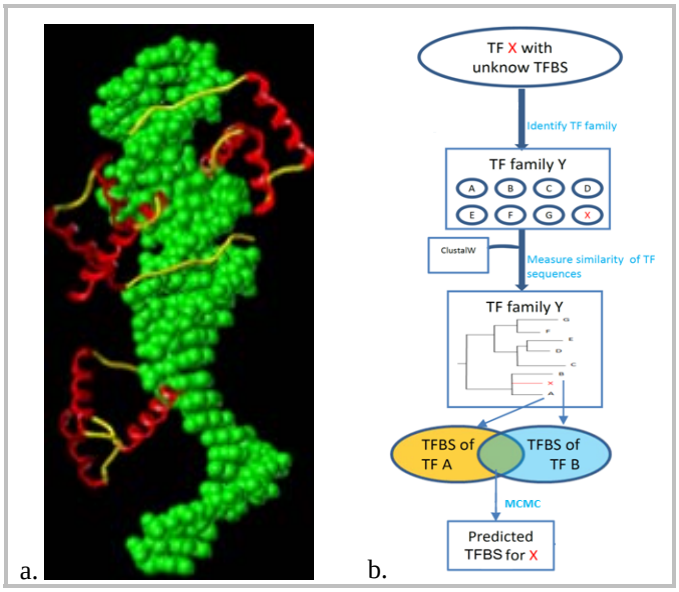

Figure 1: **a.**: 3D structures of TF-TFBS interaction complex. DNA is in green and the TF protein is in red and yellow. The picture shows a TF protein called MyoD, and is generated using Jmol visualization tool. **b.**: Schematic workflow to predict TFBSs. Suppose X is the poorly studied TF protein with unknown TFBSs information, it could be located to the closet known TF family Y (a set of TFs share highest similarity in sequences) by simply comparing the sequence similarity. The using MCMC, the TFBSs of X could be generated from its neighbors' TFBSs. For illustration, the figure shows the closet two neighbors of X, TF A and B.

Secondly, the proposed MCMC is applied using similar idea as the Metropolis–Hastings algorithm. The binding motifs of target TF X are sampled from its neighbors' TFBSs sets. The sampling process is a Markov process.

1. Initialization. Set the value of target function f=0. Randomly pick out one TFBSs set in S, say S_j. For each of the rest sets S_i in S, randomly sample n_i sequences out to represent this set. n_i is proportional to the corresponding weight w_i, i.e. $n_i = w_i N$. This will form an initial TFBSs set with $\sum_{-j} n_i$ sequences.
2. Generate a PWM matrix M by considering all the $\sum_{-j} n_i$ sequences. To calculate the matrix, we take a simple average of these sequences.
3. Then for all the sequences in S_j, use the PWM matrix M to score each of them. The score is simply a summing up of the base occurrence probability for each position.
4. The top n_j sequences in S_j with highest scores are selected as s_j. Then generate a new PWM matrix M' based on M by incorporating s_j. Use M' to compute the co-evolutionary value as described in [3].
5. The co-evolutionary value is then compared with f. If it is greater than f, assign it to f. And the algorithm proceeds to isolate another TFBS set $S_{j'}$ with sequences set s_j to represent S_j. If the value is not greater than f, keep f. And the algorithm proceeds to isolate another TFBS set $S_{j'}$ with sequences set randomly sampled to represent S_j.
6. Update all the other TFBSs set in S respectively.
7. Repeat 2 to 6 until convergence or maximum attempts reached.

The pseudo code of the MCMC algorithm is shown in **Figure2**

```
Procedure MCMC{
        t = 0;
        f=0;
        Initialize Set(t);
        While (Not Converge or t<Maximal_Attempt)
        {
                For Set_i(t) in Set(t)
                {
                        Update Set_i(t)
                        new_value = co_evo(Set_i (t));
                        If (new_value>f )
                        {
                                new_value = f;
                                Set_i(t+1) = Set_i(t);
                        }
                        t = t + 1;
                 }
        }
}
```

Figure 2: Pseudo-code of MCMC algorithm

3 Evaluation

To evaluate the performance of the proposed algorithm, real biological data is used. Since the algorithm is designed for poorly studied TFs with no prior information of its TFBSs, in order to evaluate it, TFs with known TFBSs are used without its TFBSs. The predicted TFBSs could then be evaluated by the real TFBSs. And this is done by performing leave-one-out-cross-validation on TFs of four well-studied TF families. At each time, one of the TFs is left out and considered to be the target TF. Its TFBSs are then predicted using the

above method.

3.1 Results

The performance are assessed by sensitivity, which measures the ratio of true predictions among all true TFBS; specificity, which measures the ratio of true predictions among all predictions; and Mathew's correlation coefficient (MCC) [10], which is a balance of sensitivity and specificity. Four TF families used in [3], including Homeo, HMG, TRP and bHLH families are used for evaluation. The data source has been described in section 2.2. The results are shown in **Table1**. The evaluation values in each cell are the average value across the whole family, with standard deviation in the bracket.

Table 1: Evaluation of TFBSs prediction in four real TF families

TF families	Species	TF numbers	Sensitivity (std)	Specificity (std)	MCC (std)
Homeo	CAEEL	100	0.66(0.04)	0.50(0.01)	0.52(0.01)
HMG	eukaryotes	15	0.37(0.02)	0.37(0.02)	0.35(0.01)
TRP	eukaryotes	11	0.25(0.18)	0.13(0.07)	0.16(0.10)
bHLH	eukaryotes	36	0.65(0.10)	0.46(0.02)	0.53(0.06)

In general, the performance is acceptable considering the difficulty for this prediction problem. And also here uses a stringent criterion to assess the prediction results: only when the predicted TFBS are exactly the same with the real TFBS, it makes a right prediction. Some relaxed criteria that commonly used (e.g. allow one or two wrong bases in TFBS) may result in a better look. The algorithm performed exceptionally well on the Homeo and bHLH TF set. And these two families have more TFs (100 and 36) been tested compared to HMG and TRP families (15 and 11), which makes the positive results more significant.

4 Discussion

Current TFBSs prediction methods are largely based on the conservation information of DNA sequences. This new method provides new insights by recruiting co-evolutionary information. It could serve as a supplementary approach to existing methods. However, in order to be really useful in practice and benefit the whole community in this field, the algorithm need to be further optimized. The randomization step, TFBSs set size and the length of the motif are a little bit arbitrary set in current version. Also, the performance of the algorithm is not stable and for some TF families the performances are poor, which may indicate some latent factors have not been taken into account.

Besides, the current version is written in Perl and it takes on average ~1hr to finish the prediction for one TF (including the whole pipeline instead of only the core MCMC though). The testing platform is on a typical desktop workstation with a 2.66 GHz Intel core 2 processor Q9400 and 16GB of RAM, and the system openSUSE 11.1. All programs run on a single thread.

Moreover, this MCMC method solves a modified problem compared with conventional one. A potential issue is that the TFBS sequences predicted by this method may not exist in the genome. But as the motif length in this paper is as short as 8 (compared with the genome length, 3.4 billion base for human), such issue may not be a problem in practice.

Current prediction framework is based on the condition that the novel TF X locates within a known TF family tree. However, the method proposed here could be applied to any new TF as showed in **Figure 3**. As long as we can get its protein sequence, it can be located on the evolutionary tree, and its neighboring TFs could be obtained based on the sequence similarity.

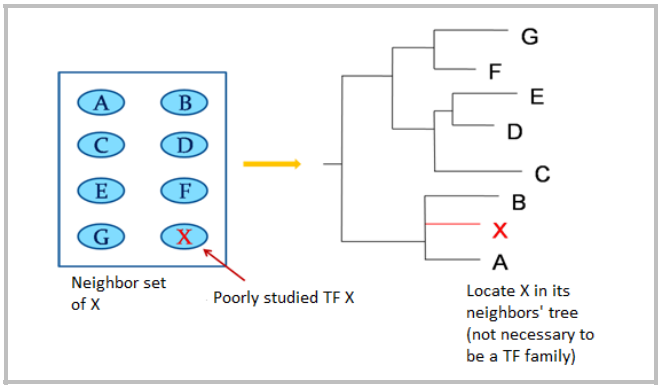

Figure 3: The relationship between target TF X and its neighbors in evolution. Any X will have neighbors with high sequence similarities. Its binding motifs could then be sampled from its neighbors' TFBSs sets.

Since currently there is no available computational approaches to predict TFBSs for poorly studied TFs, the evaluation does not involve comparison with other methods. Although this new method could be compared with some existing methods with minor adjustment, the key for this method is to deal with TFs with no prior information of its TFBSs which has been a gap in DNA motif finding field. Since the co-evolutionary relationship between TFs and TFBSs has just been discovered recently, this study may serve as an initial attempt and stimulate more researches in the future.

References

[1] D. S. Latchman, "Transcription factors: an overview," *Int J Biochem Cell Biol*, vol. 29, pp. 1305-12, Dec 1997.

[2] D. L. Fulton, *et al.*, "TFCat: the curated catalog of mouse and human transcription factors," *Genome Biol*, vol. 10, p. R29, 2009.

[3] S. Yang, *et al.*, "Correlated evolution of transcription factors and their binding sites," *Bioinformatics*, vol. 27, pp. 2972-2978, Nov 1 2011.

[4] J. R. P. Bieganski, J. V. Carlis, and E. Retzel, "Generalized suffix trees for biological sequence data: applications and implementations," in *In Proc. of the 27th Hawaii Int. Conf. on Systems Sci.*, 1994, pp. 35-44.

[5] Z. Wei and S. T. Jensen, "GAME: detecting cis-regulatory elements using a genetic algorithm," *Bioinformatics*, vol. 22, pp. 1577-1584, Jul 1 2006.

[6] M. A. L. a. A. M. Tyrrell, "The evolutionary computation approach to motif discovery in biological sequences," in *In GECCO '05: Proceedings of the 2005 workshops on Genetic and evolutionary computation*, 2005, pp. 1-11.

[7] C. E. Lawrence, *et al.*, "Detecting Subtle Sequence Signals - a Gibbs Sampling Strategy for Multiple Alignment," *Science*, vol. 262, pp. 208-214, Oct 8 1993.

[8] J. Wu and J. Xie, "Computation-based discovery of cis-regulatory modules by hidden Markov model," *J Comput Biol*, vol. 15, pp. 279-90, Apr 2008.

[9] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," *Proc Int Conf Intell Syst Mol Biol*, vol. 2, pp. 28-36, 1994.

[10] J. Wang and S. Hannenhalli, "Generalizations of Markov model to characterize biological sequences," *Bmc Bioinformatics*, vol. 6, Sep 6 2005.