Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

An efficient method for statistical significance calculation of transcription factor binding sites

Ziliang Qian[1, 2, $], Lingyi Lu[1, 2, $], Liu Qi[3], Yixue Li[1, 3, 4,*]

[1]Bioinformatics Center, Key Laboratory of Molecular System Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China; [2] Graduate School of the Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100039, China; [3]School of Life Science and Biotechnology, Shanghai Jiao Tong University; [4]Shanghai Center for Bioinformatics Technology, 100 Qinzhou Road, 200235 Shanghai, China; [$]These authors contributed equally to this work; Yixue Li* - E-mail: yxli@sibs.ac.cn; *Corresponding author

received December 13, 2007; accepted December 31, 2007; published online December 30, 2007

Abstract:
Various statistical models have been developed to describe the DNA binding preference of transcription factors, by which putative transcription factor binding sites (TFBS) can be identified according to scores assigned. Statistical significance of these scores, usually known as the p-value, play a critical role in identification. We developed an efficient algorithm to provide precise calculation of the statistical significance, remarkably enhancing the calculation efficiency by reducing the time complexity from an exponent scale to a linear scale, and successfully extended the application of this algorithm to a wide range of models, from the commonly used position weight matrix models to the complicated Bayesian Network models. Further, we calculated p-values of all transcription factor DNA binding sites recorded in the database, JASPAR, and based on these, we investigated some unseen properties of p-values as a whole, such as the p-value distribution of different models and the p-value variance according to changed scoring schemes. We hope that our algorithm and the result of computational experiments would offer an improved solution to the statistical significance of transcription factor binding sites. The software to implement our method can be downloaded from http://pcal.biosino.org/pCal.html.

Keywords: transcription factor; DNA; binding sites; Bayesian network

Background:
Transcription factor (TF), termed as the major regulator of tissue and environment specific gene expression, binds to sequence specific sites in the regulatory region to control the transcription of its target gene. Extensive efforts have been made in developing statistical methods to describe DNA binding specificity of transcription factors **[2-8]** based on known transcription factors and their binding sites **[1, 9]**. Early in the 1980s, Gary D. Stormo **[10]** proposed the widely used Position Weight Matrix (PWM) model to characterize the binding preference, which rests on the main assumption that the probability of any nucleotide that occurs at a certain position of the binding site is independent of those occurring at all other positions. However, this assumption of independency remains disputed in some circumstances. Recent work demonstrated the necessity of the inner dependency among nucleotide positions **[5, 11]** and employed new methods to capture the inner dependency **[4, 6, 8]**, such as the Bayesian Network Model **[2]**, serving as a more natural approach to uncover the substance of transcription factor binding site (TFBS).

Given the model which characterizes DNA binding preference, we can perform a genome-wide scan to identify putative TFBSa. Each putative TFBS will be assigned a score by a certain score scheme to evaluate the binding potential. But it remains a challenge since controlling the false positive/negative prediction is urgently needed in performing such a scan, especially for eukaryotic genomes where binding sites appear in extremely long intergenic regions. Calculating statistical significance of these scores provides a conventional method to reduce errors in prediction. Formally, for a statistical model M to predict TFBS of the fixed length L, the p-value of a putative binding site with score S is calculated using equation (1) (given in supplementary material).

A simple idea is to exhaustingly enumerate the total nucleotide sequences with length of L, named as the sequence set hereinafter. Unfortunately, this naive method takes a time complexity of $O(4^L)$, an undesirable task when L is larger than 10. In 2005, Barash Y. improve this naïve method by introducing importance sampling technique **[2]**. However, due to the nature of the sampling technique, it produces an approximate result rather than an accurate solution.

To overcome the time complexity raised by directly enumerating the sequence set, we turn to enumerate the total scores, named as the score set hereinafter, of all possible nucleotide sequences with length of L. It is a compressed set much smaller than the sequence set since an element of the score set may have several mappings

into the sequence set, that is to say, some different sequences often have the same score. Based on this idea, in this contribution, we propose an accurate solution with time complexity of $O(4xLx\Omega)<<4^L$ to calculate p-values, where Ω is a constant about 10^4.

Methodology:

Consider a binding site $R = R_1R_2\ldots R_l\ldots R_L$ with L nucleotides, where $L\in Z^+$. Obviously, each site R_l could be any one of the four nucleotides 'A', 'C', 'G' and 'T', where $1\le l\le L$, $l\in Z^+$. And, the probability of R_l being 'A' is denoted as $p(l, R_l = \text{'A'})$. Generally, the nucleotide selecting preference of binding sites R can be expressed in terms of probability matrix, using equation (2) in supplementary material.

The independency assumption is not always reliable as it is common that a single amino acid residue contacts with more than one nucleotide and vice versa. So, to some extent, a nucleotide R_l is correlated with other nucleotides. In this case, the probability of R_l being 'A' is expressed in terms of the conditional probability using equation (3) (under supplementary material).

Generally, a more reasonable statistical model for describing the nucleotide selection preference of transcription factor DNA binding sites can be denoted as in equation (4) (shown in supplementary material).

Next we scan regulatory regions followed by assigning each sites a score in the light of the following score scheme, by using formula (5) (given in supplementary material).

Then potential TFBS, R, can be identified according to the score cutoff, where binding sites with scores larger than the threshold or with adequate statistical significance are considered as potential TFBSs. Calculating the statistical significance, usually known as the p-value, is a conventional way to define cutoffs. Since the null-distribution of total scores is known, the p-value of any potential TFBS can be easily obtained according to equation (1) (supplementary material). Therefore, our task is to offer an effective and accurate solution for calculating the null-distribution of all scores, which is regarded as the prerequisite to p-values.

Assume that the sequence set is composed of all short nucleotide sequence with length of L, denoted as in equation (6) shown in the supplementary material.

Intuitively, the size of S_L is smaller than R_L, since some of the different sequences in R_l result the identical score in S_L. Hence, enumerating set S_L can be much more efficient than enumerating R_L, leading to considerable reduction of computation time. In fact, our improvement in time economy goes far beyond this extent. Since the exact size of S_l contributes remarkably to the efficiency of our method as we mention above, how we control this factor is of great importance. Here we estimate that the size of S_l, is less than 10^{m+n}, where m, $n\in Z^+$ are parameters that represent the count of effective digits of $S(R_l)$, $S(R_l) \in S_l, l=1, 2\ldots,L$ before the decimal point and behind the decimal point respectively (equation (4) see supplementary material). Here, we summarize our method into pseudo codes presented in figure 1.

Main Data Structure:

Store the score distribution in a hash $M\{s\} = c$ (or $\langle s, c\rangle \in M$), where c denotes the count of score value s.

Initiation: Fill M_0 with $\langle 0,1\rangle$

Iteration:

```
FOR i = 0,1,...,L-1
    FOR EACH <s_k, c_k> IN M_i
        FOR R_{i+1} IN {'A','C','G','T'}
            Calculate score s' and corresponding count c'
                { s' += s(i+1, R_{i+1}, pa_{i+1})
                { c' ++
            IF s' exists in M_{i+1}
                M_{i+1}{s'} += c'
            ELSE
                M_{i+1}{s'} = c'
        END FOR
    END FOR
END FOR
```

Calculate p-Value Based on the Null-distribution of Total Scores:

The p-value for a putative TFBS with score s'

$$p\left(s\left(\text{random } L \text{ bp nucleotide sequence}\right) > s'\right) = \frac{\sum_{z_k \ge s', \langle z_k, c_k\rangle \in M_L} c_k}{\sum_{\langle z_j, c_j\rangle \in M_L} c_j}, 1 \le k, j \le \#M_L$$

where, $\#M_L$, the size of M_L.

Figure 1: The pseudo code of our method

ISSN 0973-2063
Bioinformation 2(3): 97-100 (2007)

Results:

The dataset of all our computational experiments is taken from the database JASPAR **[1]**, including matrices that describe transcription factor DNA binding preference as well as their corresponding DNA binding sites. All the calculations were performed in a Dell Optiplex 270 machine, which has an Intel(R) Pentium(R) 4 CPU of 2.60GHz and 1.5GB RAM.

Selecting the optimal parameter

As we discussed in the method section, the adjustable parameter *c*, affects the size of score sets, and therefore affects the operation time greatly, where $c = m+n$, $c \in Z^+$ and m, $n \in Z^+$ represent the count of effective digits before the decimal point and behind the decimal point respectively of elements in score sets. We must find the optimal value of *c* so as to determine the upper limit of time complexity of our algorithm $O(Lx10^c)$. In our computational experiments, transcription factor DNA binding sites with various lengths in database JASPAR were used to compute the p-values and the result shows that c=4 is already good enough for calculating p-values even when the p-value is more strict than 10^{-5}.

Comparison among different scoring schemes

As we mentioned in the introduction, p-value calculation is also dependent on scoring schemes (see equation (1) in supplementary material). Many researchers proposed their own scoring schemes **[2, 4, 14]**, and whether different scoring schemes lead to different results of p-values has become of great interest to researchers. Authors of software package MATCH, proposed an information theory based method for calculating scores for potential TFBSs is shown in equation (7) under supplementary material.

How it outperforms the conventional cumulative probability based scoring scheme is given in equation (8) (see supplementary material).

Here we adopt two different score schemes to calculate the TFBSs from database JASPAR. The result shows there is no significant difference between each other in figure 2.

Besides, according to the p-value distribution in figure 3, we found that most of these p-values are around 10^{-4}, thereby supplying a reference point for credible p-values to transcription factor binding sites.

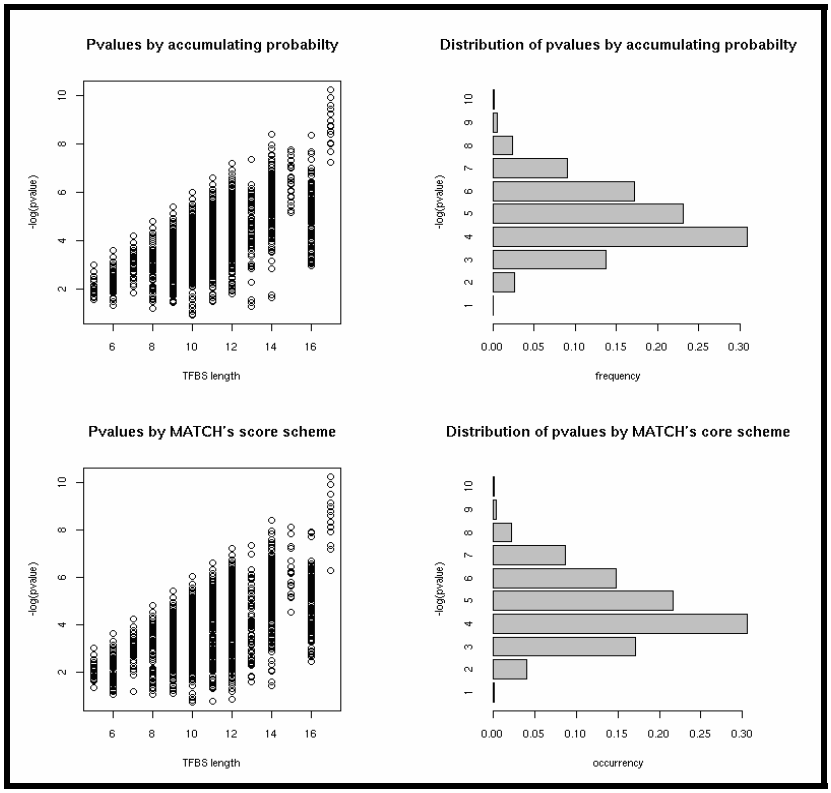

Figure 2: Comparison of p-values obtained based on different scoring schemes. The vertical axis in this figure depicts the −log(p-value) and the horizontal axis of left two sub graphs depicts length of TFBS, while the horizontal axis of two sub-graphs (right) depicts the frequency of different p-values. Each point in the two left sub-graphs represents the log of p-value corresponding to a certain TFBS. The two right sub-graphs are the distribution of p-values by accumulating probability and MATCH score scheme. Although different scoring schemes were used, p-value distributions not very different.

ISSN 0973-2063
Bioinformation 2(3): 97-100 (2007)

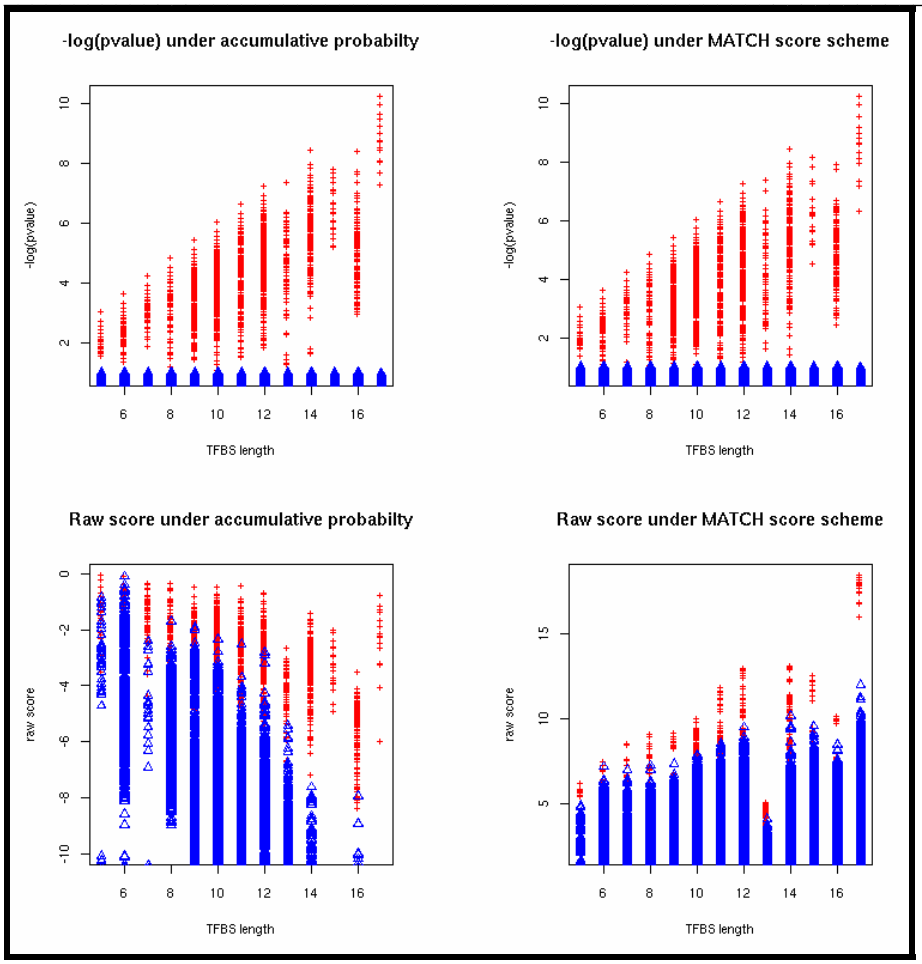

Figure 3: p-value is better than the raw score for identifying TFBS. The vertical axis of the two upper sub-graphs depicts the –log(p-value), while the vertical axis of the two lower sub-graphs depicts the raw score directly obtained from scoring schemes. The horizontal axis in all sub-graphs depicts the various lengths of TFBS. The blue dots represent p-values of true TFBSs provided by JASPAR whereas red dots represent p-values of DNA sequence fragments of various lengths from genome background. According to the two upper sub-graphs, there is a sharp distinction between true TFBSs and genomic background for both the cumulative probability based scoring scheme and the MATCH scoring scheme. However, in the two lower sub-graphs, blue dots and red dots appear fused together, indicating that raw scores are not appropriately used as the criterion to identify TFBS

P-values serve as a better criterion to identify TFBS than the raw scores do

In previous sections, we repeatedly mentioned that adopting p-value as a conventional way to define cutoffs for distinguishing the true TFBS from its background sequences. Why bother to do so rather than simply adopting raw scores directly obtained from scoring schemes, as showed in formula 5, 7 and 8 (under supplementary material), to define the cutoff? Another computational experiment gives a credible explanation to this question. We collected the true TFBS from database JASPAR as the positive dataset, and some DNA sequence with the same length of those TFBSs from genome background as negative control. We calculated raw scores as well as p-values of the type data. As results shown in figure 3, there is a blurred part between the true TFBS and the genome background when we adopt raw scores to distinguish them, whereas there is a sharp distinction between true TFBSs and genomic background when we adopt p-values. It is the blurred part that is very likely to cause the error in the identification. Therefore, we are able to conclude that p-values serve as a better criterion to identify TFBS than the raw scores do.

Discussion:

Various statistical models have been developed to describe the transcription factor DNA binding preference, by which we identify putative transcription factor binding sites according to scores assigned. Statistical significance of these scores play a critical role in assessing the efficiency

ISSN 0973-2063
Bioinformation 2(3): 97-100 (2007)

of prediction. We developed an efficient algorithm to provide precise calculation of the statistical significance. With regards to the time efficiency of our algorithm, our major improvement rests on two key points. First, by calculating the scores of the overlapping part of sequences foremost, we reduced the total time consumption considerably. Further, instead of enumerating elements in the sequence set, we performed our calculations with the more compressed score sets, thus we skillfully convert the time complexity of being a exponent in relation to TFBS length L to that of a linear relation with L, which is a remarkable improvement.

Moreover, since our algorithm is generally based on the enumerating approach, the p-value calculated by our method is a precise solution, different from the result of the sampling method, which is the approximate solution due to the nature of sampling strategy.

Beside the speediness and preciseness of this algorithm, another positive point lies in its applicability. As an alternative to Probability-Generating-Function-based methods, such as Staden's **[15]** and Huang's methods **[12]**, our method can be applied not only to the context of independent identical distribution of relevant nucleotides, like PWM models, but also to Bayesian Network models. In all, table 1 under supplementary material summarizes the properties of our method compared with others.

References:

[01] D. Vlieghe, *et al.*, *Nucl Acids Res.*, 34: D95 (2006) [PMID: 16381983]
[02] Y. Barash, *et al.*, *Comput Appl Biosci.*, 21: 596 (2005)
[03] I. Ben-Gal, *et al.*, *Comput Appl Biosci.*, 21: 2657 (2005)
[04] P. Hong, *et al.*, *Comput Appl Biosci.*, 21: 2636 (2005)
[05] Q. Zhou & J. S. Liu, *Comput Appl Biosci.*, 20: 909 (2004)
[06] O. D. King & F. P. Roth, *Nucl Acids Res.*, 31: e116 (2003) [PMID: 14500844]
[07] G. D. Stormo, *Comput Appl Biosci.*, 16: 16 (2000)
[08] N. I. Gershenzon, *et al.*, *Nucl Acids Res.*, 33: 2290 (2005) [PMID: 15849315]
[09] V. Matys, *et al.*, *Nucl Acids Res.*, 31: 374 (2003) [PMID: 12520026]
[10] G. D. Stormo, *et al.*, *Nucl Acids Res.*, 10: 2997 (1982) [PMID: 7048259]
[11] R. A. O'Flanagan, *et al.*, *Comput Appl Biosci.*, 21: 2254 (2005)
[12] H. Huang, *et al.*, *Journal of Computational Biology*, 11: 1 (2004) [PMID: 15072685]
[13] J. E. Donald & E. I. Shakhnovich, *Nucl Acids Res.*, 33: 4455 (2005)
[14] A. E. Kel, *et al.*, *Nucl Acids Res.*, 31: 3576 (2003) [PMID: 16085755]
[15] R. Staden, *Comput Appl Biosci*., 5: 89 (1989)

Edited by A. M. Khan, T. W. Tan & S. Ranganathan
Citation: Qian *et al.*, Bioinformation 2(5): 169-174 (2007)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

	PWM	Bayesian network model
Enumerating Sequence Set	Slow	Slow
Sampling Sequence Set (Barash, et al., 2005)	Fast but not accurate solution	Fast but not accurate solution
Probability Generating Function (Staden, 1989)	Accurate Solution	Not Available
Our Method	Accurate Solution	Accurate Solution

Table 1: Comparison with existing methods

Equations

$$p\left(S\left(\text{random } L \text{ bp nucleotide sequence}\right) > s\right) \quad \rightarrow \quad (1)$$

where function S is the scoring scheme. According to formula 1, the prerequisite to p-value is to obtain the distribution of scores of all possible L bp nucleotide sequences, which is usually called null-distribution.

$$\begin{pmatrix} p(1,R_1='A') & p(2,R_2='A') & \cdots & p(l,R_l='A') & \cdots & p(L,R_L='A') \\ p(1,R_1='C') & p(2,R_2='C') & \cdots & p(l,R_l='C') & \cdots & p(L,R_L='C') \\ p(1,R_1='G') & p(2,R_2='G') & \cdots & p(l,R_l='G') & \cdots & p(L,R_L='G') \\ p(1,R_1='T') & p(2,R_2='T') & \cdots & p(l,R_l='T') & \cdots & p(L,R_L='T') \end{pmatrix} \quad \rightarrow \quad (2)$$

$$p\left(l,R_l='A' \mid pa_l\right) \quad \rightarrow \quad (3)$$

where pa_l represents the configuration of parent nodes(direct dependent nucleotides) of the lth nucleotide. For example, if the lth nucleotide has two parent nucleotides and the configuration (appearance) of them are 'A' and 'T', than the probability of R_l being 'A' is $p(l,R_l=$ 'A' $\mid pa_l=$ 'AT').

$$p\left(l,R_l \mid pa_l\right) \quad \rightarrow \quad (4)$$

where, l = 1, 2, …, L, $R_l \in$ { 'A', 'C', 'G', 'T'}, and $pa_l \in \{R_1R_2\ldots R_i\ldots R_n \mid n \in Z^+, R_i \in$ { 'A', 'C', 'G', 'T'}. Insightful discussions of Bayesian Network description on TFBS can be referred to Nir Friedman and co-worker's work (Barash, et al., 2003).

$$S(R) = \sum_{l=1}^{L} S\left(l,R_l,pa_l\right) \quad \rightarrow \quad (5)$$

$$= \sum_{l=1}^{L} \log p\left(l,R_l \mid pa_l\right)$$

$$R_L = \left\{R \mid R = R_1R_2\cdots R_l\cdots R_L, R_l \in \left\{'A','C','G','T'\right\}, 1 \le l \le L, l, L \in Z^+\right\} \quad \rightarrow \quad (6)$$

The scores of each short sequence in R_L consist of another set $S_L=\{S(R) \mid R \in R_L, L \in Z^+\}$, namely the score set.

$$I(l) = \sum_{R_l \in \{'A','C','G','T'\}} p\left(l,R_l\right) \bullet \log\left(4 \bullet p\left(l,R_l\right)\right) \quad \rightarrow \quad (7)$$

$$S(R) = \sum_{l=1}^{l=L} p\left(l,R_l\right) \bullet I\left(l\right)$$

$$S(R) = \sum_{l=1}^{l=L} \log \frac{1}{p\left(l,R_l\right)} \quad \rightarrow \quad (8)$$

ISSN 0973-2063
Bioinformation 2(3): 97-100 (2007)