JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 27, 855-868 (2011)

Similarity Analysis between Transcription Factor Binding Sites by Bayesian Hypothesis Test*

QIAN LIU+, SAN-YANG LIU AND LI-FANG LIU
+*Department of Mathematics*
School of Computer Science and Technology
Xidian University
Xi'an, 710071 P.R. China

Transcription factor binding sites (TFBS) in promoter sequences of higher eukaryotes are commonly modeled using position frequency matrices (PFM). The ability to compare PFMs representing binding sites is especially important for de novo sequence motif discovery, where it is desirable to compare putative matrices to one another and to known matrices. We propose to identify and group similar profiles using Bayesian hypothesis test between PFMs, describing a column-by-column method for PFM similarity quantification based on Bayes factor and posterior probability of null model that aligned columns are independent and identically distributed observation from the same multinomial distribution. We group TFBS frequency matrices from less redundant JASPAR into matrix families by cluster analysis according to Bayes factors and posterior probability of similar PFMs. Clusters of highly similar matrices are identified. We further compare the performance of this method to Pearson χ^2 test on simulated data. The proposed method is very simple, easily implemented and outperforms the other method in our test. Taking Pearson product moment correlation coefficient as an objective criterion of the performance, results indicate that Bayesian test performs better than the classical methods on average.

Keywords: transcription factor binding site, position frequency matrices, similarity, Bayes factor, posterior probability, cluster analysis

1. INTRODUCTION

Transcription regulation is carried out by the interactions between transcription factors and their binding sites in DNA. Transcription factor binding sites (TFBS) are short, degenerate nucleotide sequences, and usually are 6-20 bp long. TFBS discovery in promoter sequences is important for predicting transcription regulation. Candidate binding sites of known transcription factors are located by consensus sequence search or binding scores calculated from position weight matrices (PWMs) [1]. These matrices are derived from position frequency matrices (PFMs) obtained by aligning binding sites for a given transcription factor. PFMs contain the observed nucleotide frequencies at each position of the alignment. A popular collection of eukaryotic PFMs is given by the TRANSFAC database [2]. Furthermore, an open-access database, JASPAR database which is less redundant database [3], has been compiled recently. Table 1 shows the position frequency matrix for motif MA0026 from the JASPAR database.

Received October 30, 2009; revised May 5 & July 7, 2010; accepted August 23, 2010.
Communicated by Jorng-Tzong Horng.
* This work was supported by the National Natural Science Foundation of China (Grants No. 60705004) and the Fundamental Research Funds for the Central Universities (k50510030004).

Table 1. Position frequency matrix for motif MA0026.

Pos	1	2	3	4	5	6	7
A	2	5	0	0	17	17	5
C	12	12	0	0	0	0	1
G	2	0	17	17	0	0	10
T	1	0	0	0	0	0	1

Existing approaches for quantifying PFM similarity include the Pearson correlation coefficient (PCC) method [4, 5], the tool CompareACE [6] based on PCC method, the average log-likelihood ratio (ALIR) method [7], a method based on the constraints imposed by the structures of DNA binding proteins introduced by Sandelin and Wasserman [8] which allowed for gapped PFM alignment. More recently, a statistical test method using approximated Pearson χ^2 test and Fisher-Irwin exact test [9] was proposed. They showed that these two new functions have better discriminative power than the ALIR and PCC similarity functions.

In this work we develop a statistical method for comparing two similar PFMs with one another. The type of comparison is valuable within the context of TFBS discovery. We describe a column-by-column method for PFM similarity quantification based on Bayes factor and posterior probability of null model that aligned columns are independent and identically distributed observation from the same multinomial distribution. We compared the performance of this method to the Pearson χ^2 test method on simulated data. Taking Pearson product moment correlation coefficient as an objective criterion of the performance, results indicate that Bayesian test performed better on average.

We used this PFM similarity quantification to classify TFBS frequency matrices into matrix families by cluster analysis according to Bayes factors and posterior probability of similar PFMs. We grouped PFMs in JASPAR into PFM-families. We also found that PFM-families are likely to include TFBS PFMs for related transcription factors. Some clusters not only reflect pronounced similarities of the same transcription factor with different degrees of stringency, but also identify different transcription factors of certain families having almost identical binding motifs. For example, a PFM family includes MA0026, MA0028, MA0062, MA0076, MA0080, and MA0081 for binding sites of ETS transcription factor and another PFM family includes MA0014 and MA0027 for binding site of PAIRED transcription factor and HOMEO transcription factor.

Comparison tools for TFBS PFMs are important for test newly discovered TFBSs against existing matrices, reducing redundancy in databases and increasing the quality of the matrices. Our motivations for applying this similarity measure to established database, such as JASPAR are to eliminate "redundant" matrices within this database and to understand the similarities as well as differences among the different transcription factors. If two (or more) TFBSs have nearly identical core matrices, we can consider them to be redundant because they do not contain unique information in terms of appearance. The presence of redundant matrices complicates the use of these databases for further TFBS discovery, because searches that involve several nearly identical matrices will lead to an excessive number of predicted sites. So clusters of highly similar matrices are identified and can be used to optimize the search for TFBSs.

2. METHODS

Fig. 1 shows the sequence logos for four different motifs from the JASPAR database, which will be analyzed in section 3. It is clear that these motifs all show differences in width and appearance, but there exists some similarity or common structure within this set. The statistical problem of interest here is to identify the similarity between these different TFBSs and group them together based on their similarity.

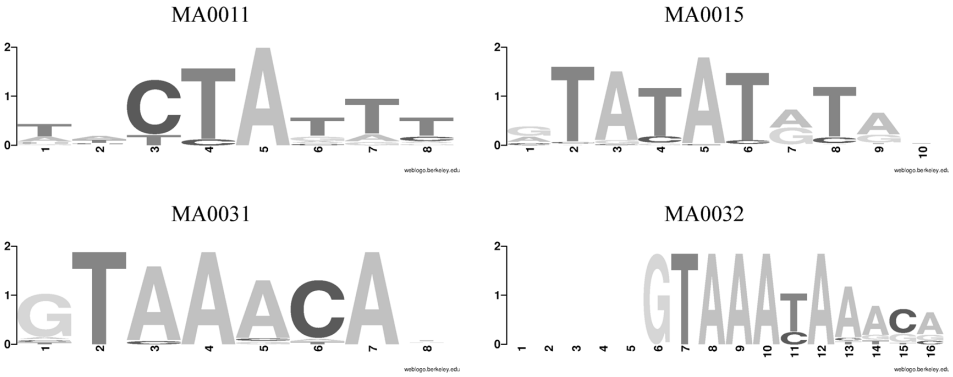


Fig. 1. Four different motifs from the JASPAR database.

In this section, we first present existing four column comparison functions for comparing PFMs: the Pearson correlation coefficient, the average log-likelihood ratio, the Pearson χ^2 test and the Fisher-Irwin exact test. Then we present skeleton algorithm for Bayesian hypothesis test. We use a statistical test for determining the likelihood that two columns are generated from the same multinomial distribution. This likelihood can be computed using Bayesian inference.

We assume that PFMs follow a product multinomial distribution [10]. Each column is a set of independent and identically distributed observations, and matrix comparisons reduce to column-by-column comparisons. The overall similarity score for a matrix pair is derived from the individual column scores.

2.1 Existing PFM Column Comparison Functions

In the following discussion, $\boldsymbol{X}$ refers to a column of one PFM and is a multinomial frequency vector. The quantity X_a refers to the probability of letter $a \in A$ in $\boldsymbol{X}$. We use N_{Xa} to refer to the count of letter a in column $\boldsymbol{X}$. Similar definitions apply for $\boldsymbol{Y}$, a column from the other PFM. The quantity $|A|$ refers to the length of the alphabet (4 for DNA, 20 for proteins).

Pearson Correlation Coefficient (PCC) The PCC was first introduced for computing PFM similarity by Pietrokovski [4]. For two columns $\boldsymbol{X}$ and $\boldsymbol{Y}$, PCC is computed using the following formula:

$$\mathrm{PCC}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{a\in A}(X_a-\bar{X})(Y_a-\bar{Y})}{\sqrt{\sum_{a\in A}(X_a-\bar{X})^2\sum_{a\in A}(Y_a-\bar{Y})^2}},\ \bar{X}=\frac{1}{|A|}\sum_{a\in A}X_a,\ \bar{Y}=\frac{1}{|A|}\sum_{a\in A}Y_a. \tag{1}$$

To compare matrices consisting of multiple columns, the scores of the individual column comparisons are summed.

Average Log-Likelihood Ratio (ALLR) The ALLR formula described by Wang and Stormo [7] to quantify similarity between columns $\boldsymbol{X}$ and $\boldsymbol{Y}$ for position frequency matrices is a weighted sum of two log-likelihood ratios:

$$\mathrm{ALLR}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\sum_{a\in A}N_{Xa}\log(\frac{Y_a}{p_a})+\sum_{a\in A}N_{Ya}\log(\frac{X_a}{p_a})}{\sum_{a\in A}(N_{Xa}+N_{Ya})} \tag{2}$$

where p_a is the background (prior) frequency of letter a. Again, to compare matrices consisting of multiple columns, the scores of the individual column comparisons are summed.

Pearson χ^2 Test The Pearson χ^2 test was introduced by Schones and coworkers [9] for comparing PFMs. The χ^2 P value is computed for the null hypothesis that the aligned columns are independent and identically distributed observations from the same multinomial distribution. The value of χ^2 is computed using the following equation:

$$\chi^2(\boldsymbol{X},\boldsymbol{Y}) = \sum_{j=X,Y}\sum_{a\in A}\frac{(N^e_{ja}-N^o_{ja})^2}{N^e_{ja}} \tag{3}$$

where $N^o_{ja} = N_{ja}$ is the observed count of letter a in column j, and $N^e_{ja} = N_jN_a/N$ is the expected count of letter a in column j.

The P value is calculated from this χ^2 score using $|A| - 1$ degrees of freedom. The P value for multiple columns is the geometric mean of the column P values of the individual columns.

Fisher-Irwin Exact Test (FIET) The FIET [9] is an analytical computation of the Pearson χ^2 P value. That is the χ^2 test is an approximation of Fisher-Irwin exact test. The marginal P value of the contingency table follows the multiple hyper geometric distributions [11]:

$$p = \frac{\binom{N_X}{N_{XA},N_{XC},N_{XG},N_{XT}}\binom{N_Y}{N_{YA},N_{YC},N_{YG},N_{YT}}}{\binom{N}{N_A,N_C,N_G,N_T}}. \tag{4}$$

The P value for multiple columns is the product of the P values of the individual columns.

2.2 Bayesian Hypothesis Test

Bayesian methods have already been used in algorithms for sequence alignment [10], but in our implementation we use Bayesian inference to get column comparison function. Bayesian inference encourages the use of predictive densities and evidence scores [12]. Based on Bayesian hypothesis test, first we will test the following hypothesis:

H_0: The column $\boldsymbol{X}$ and $\boldsymbol{Y}$ come from the same multinomial distribution,
H_1: otherwise. (5)

Although it can be used to test (5), the Pearson χ^2 test is an approximation test. The approximation does not hold when the marginal frequencies are small, specifically when at least one of the marginals is < 5, a condition that occurs often in PFMs of TFBSs [13].

This hypothesis test (5) can be evaluated directly in the form of a test for independence [12, 14]. It can be converted into a test for independence in the following way.

We introduce a random variable Q which takes values $\{X, Y\}$. When $Q = X$, it represents the distribution $\boldsymbol{X}$ and when $Q = Y$, it represents $\boldsymbol{Y}$. A second variable B takes its values from the DNA sequence alphabet A as shown in Table 2.

From this, we can test the following hypotheses (equivalent to (5)):

H_0: B and Q are independent random variables,
H_1: otherwise. (6)

Table 2. Relationships between random variables $\boldsymbol{B}$ and $\boldsymbol{Q}$.

Q \ B	$B = A$	$B = C$	$B = G$	$B = T$
$Q = X$	N_{XA}	N_{XC}	N_{XG}	N_{XT}
$Q = Y$	N_{YA}	N_{YC}	N_{YG}	N_{YT}

When B and Q are independent, $\boldsymbol{X}$ and $\boldsymbol{Y}$ will not share the same distribution. Under the hypothesis of independence, the probability of the data is

$$p(Q, B \mid H_0) = p(Q \mid \alpha_{j.})\, p(B \mid \alpha_{.k}) \tag{7}$$

where $\alpha_{j.}$ and $\alpha_{.k}$ are two different prior, respectively the hyper parameter of ($p(Q = X)$, $p(Q = Y)$) and ($p(B = A)$, $p(B = C)$, $p(B = G)$, $p(B = T)$).

The other hypothesis is that the variables are dependent and arise from a multinomial distribution on pairs (j, k). This distribution has eight values. Under this hypothesis, the probability of the data is

$$p(Q, B \mid H_1) = p((Q_X, B_A), (Q_X, B_C), \ldots, (Q_Y, B_T) \mid \alpha_{jk}) \tag{8}$$

where α_{jk} is yet a third prior, the hyper parameter of

$$(p(Q = X, B = A), p(Q = X, B = C), \ldots, p(Q = Y, B = T)).$$

Based on the Bayesian Theorem,

$$p(H_0 \mid Q, B) = \frac{p(H_0)p(Q, B \mid H_0)}{p(H_0)p(Q, B \mid H_0) + p(H_1)p(Q, B \mid H_1)} = \frac{1}{1 + \frac{p(H_1)p(Q, B \mid H_1)}{p(H_0)p(Q, B \mid H_0)}}. \tag{9}$$

The crucial quantity here is $\frac{p(Q, B \mid H_0)}{p(Q, B \mid H_1)}$, the evidence ratio in favor of independence, where $p(H_0) = p(H_1) = 0.5$. Hence the Bayes factor will be

$$BF(H_0, H_1) = \frac{p(Q, B \mid H_0)}{p(Q, B \mid H_1)} = \frac{p(Q \mid \alpha_{j.})p(B \mid \alpha_{.k})}{p((Q_1, B_1), (Q_1, B_2), \ldots, (Q_2, B_4) \mid \alpha_{jk})} \tag{10}$$

where $\alpha_{j.} = \sum_{k=A,C,G,T} \alpha_{jk}, \alpha_{.k} = \sum_{j=X,Y} \alpha_{jk}$. (11)

Assume a conjugate prior of a multinomial distribution with probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ is the Dirichlet distribution:

$$p(\boldsymbol{p} \mid \alpha) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_n) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{\alpha_i - 1} \tag{12}$$

where $p_i > 0$ (13)

$$\sum_{i=1}^{n} p_i = 1. \tag{14}$$

The hyper parameter α_i of p_i can be interpreted as a virtual count for value i. Large α_i correspond to strong prior knowledge about the distribution and small α_i correspond to ignorance.

Given a Dirichlet prior, the joint distribution of S independent identical distribution samples $\boldsymbol{X} = \{x_1, \ldots, x_S\}$ and probability vector $\boldsymbol{p}$ of a multinomial distribution is

$$p(\boldsymbol{X}, \boldsymbol{p} \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} p_i^{N_i + \alpha_i - 1} \tag{15}$$

where $N_i = \sum_{s=1}^{S} \delta(x_s = i).$ (16)

So the posterior is

$$p(\boldsymbol{p} \mid \alpha, \boldsymbol{X}) \sim D(N_i + \alpha_i). \tag{17}$$

Finally, we can calculate.

$$p(\boldsymbol{X} \mid \alpha) = \int_{\boldsymbol{p}} p(\boldsymbol{X}, \boldsymbol{p} \mid \alpha) d\boldsymbol{p} = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\Gamma(S + \sum_{i=1}^{n} \alpha_i)} \prod_{i=1}^{n} \frac{\Gamma(N_i + \alpha_i)}{\Gamma(\alpha_i)} \tag{18}$$

which is the probability that the data all come from one multinomial distribution.

Then based on Eq. (18), we get

$$p(Q \mid \alpha_{j.}) = \frac{\Gamma(\sum_{jk} \alpha_{jk})}{\Gamma(N + \sum_{jk} \alpha_{jk})} \prod_{j=X,Y} \frac{\Gamma(N_j + \alpha_{j.})}{\Gamma(\alpha_{j.})}, \tag{19}$$

$$p(B \mid \alpha_{.k}) = \frac{\Gamma(\sum_{jk} \alpha_{jk})}{\Gamma(N + \sum_{jk} \alpha_{jk})} \prod_{k=A,C,G,T} \frac{\Gamma(N_k + \alpha_{.k})}{\Gamma(\alpha_{.k})}, \tag{20}$$

$$p((Q = X, B = A), p(Q = X, B = C), \ldots, p(Q = Y, B = T) \mid \alpha_{jk})$$
$$= \frac{\Gamma(\sum_{jk} \alpha_{jk})}{\Gamma(N + \sum_{jk} \alpha_{jk})} \prod_{jk} \frac{\Gamma(N_{jk} + \alpha_{jk})}{\Gamma(\alpha_{jk})}, \tag{21}$$

where $N_j = \sum_{k=A,C,G,T} N_{jk}, j = X, Y,$ (22)

$$N_k = \sum_{j=X,Y} N_{jk}, k = A, C, G, T, \tag{23}$$

$$N = \sum_{k=A,C.G.T} (N_{Xk} + N_{Yk}). \tag{24}$$

Thus we could calculate the Bayes factor of the null hypothesis that column $\boldsymbol{X}$ and $\boldsymbol{Y}$ come from the same multinomial distribution:

$$BF(H_0, H_1) = \frac{\Gamma(\sum_{jk} \alpha_{jk})}{\Gamma(N + \sum_{jk} \alpha_{jk})} \prod_{j=X,Y} \frac{\Gamma(N_j + \alpha_{j.})}{\Gamma(\alpha_{j.})} \times \prod_{k=A,C,G,T} \frac{\Gamma(N_k + \alpha_{.k})}{\Gamma(\alpha_{.k})} \prod_{jk} \frac{\Gamma(\alpha_{jk})}{\Gamma(N_{jk} + \alpha_{jk})}. \tag{25}$$

We can define all priors: $\alpha_{jk} = 0.5$ as the so called Jeffreys' prior when we evaluated the method using correlation coefficient as a parameter of success.

Table 3. Jeffrey's scale for the interpretation of Bayes factor (BF).

Bayes Factor (BF) range	Evidence
BF ≥ 1	Null hypothesis (model)* is supported
0.3 ≤ BF < 1	Minimal evidence against null hypothesis (model)*
0.1 ≤ BF < 0.3	Substantial evidence against null hypothesis (model)*
0.01 ≤ BF < 0.1	Strong evidence against null hypothesis (model)*
BF < 0.01	Decisive evidence against null hypothesis (model)*

* Null model/hypothesis: column ***X*** and ***Y*** come from the same multinomial distribution.

Acting as a measure of similarity between column ***X*** and ***Y***, a higher Bayes factor implies similarity according to Jeffreys' scale [15] given in Table 3.

The Bayes factor for multiple columns is the product of the Bayes factors of the individual columns. The null model is accepted when Bayes factor is much bigger than 1 and when the posterior probability $BF/(1 + BF)$ of the mull model is bigger than 0.8, assuming $p(H_0) = p(H_1) = 0.5$. We used the posterior probability of the null hypothesis as a final score for the similarity analysis between PFMs, because it is a more precise score value than Bayes factor [14].

3. RESULTS AND DISSCUSSION

In this section, we report out evaluation of the presented Bayesian hypothesis test method and its comparison to the method based on Pearson χ^2 test.

3.1 Real Database JASPAR

A publicly available JASPAR database includes 111 experimentally verified transcription factor binding sites associated mainly to vertebrates which differ substantially in appearance, number of counts, and motif width. Between different motifs, the number of binding sites used to construct the motif matrix (the total number of multinomial counts) varies from 6 to 389, with an average of around 35 counts per matrix. The range of matrix widths was from 4 to 30 base pairs (bp), although the matrices were generally short, with an average width of approximately 11 bp. We describe the clustering result of this real database. For the large-scale statistical analysis, we discarded all matrices with inconsistencies, for example matrices, where the number of sites aligned to construct the matrix (sample size) could not be determined. Furthermore, we excluded rather poor matrices with a length below 5 bases or a sample size below 5. After these consistency checks and filtering steps we arrived at 106 different matrices for JASPAR.

A matrix core is identified in each PFM as the five most-conserved contiguous columns (highest confidence) [9, 16] using information content [17]. Information content of an alignment matrix is also called Kullback-Leibler information and abbreviate it as I_{seq}:

$$I_{seq} = \sum_{j=1}^{L} \sum_{i=1}^{|A|} \frac{n_{ij}}{n} \log \frac{n_{ij}/n}{b_i} \tag{26}$$

where L is the column number of PFM, n_{ij} count of the ith letter from an alphabet A in the

jth column of alignment, $n = \sum_{i=1}^{|A|} n_{ij}$ and b_i the background frequency of the ith letter in A.

We used matrix cores to measure distances between PFMs in JASPAR. But the number of significantly different columns depends on the relative position of both matrices [18], so in our algorithm, we study all possible alignments with a minimum overlap of 3 bases for matrix cores of two PFMs, we count scores of corresponding columns which are statistically independent. We compared all PFM core pairs. The comparisons were ranked according to the scores, which are Bayes factor and posterior probability of the null model based on Bayesian hypothesis test. A similarity threshold was set so that two PFMs with a Bayes factor below the threshold are deemed incompatible and a Bayes factor above the threshold are considered similar. We chose the threshold 1 for Bayes factor according to Table 3 and 0.8 for posterior probability. The time complexity for the calculation of Bayes factor Eq. (25) and posterior probability is linear O (the width of PFM).

Clustering Analysis of Similar PFMs Here, agglomerative hierarchical clustering is used for PFMs clustering. The clustering procedure consists of five main steps. The clustering algorithm is described as follows.

Initialization: Compute all-against-all similarity based on matrix cores of PFMs in JASPAR database.

Repeat:
1. Select the two most similar PFMs with maximum posterior probability score.
2. Merge the two selected PFMs and update the similarity matrix scores by remaining the smallest similarity between the TFBSs in the newly created cluster and all other TFBSs.

Until: An appropriate condition is met (*e.g*. there are no members sharing sufficient similarity (average posterior probability < 0.8)).

Output: PFMs families.

The clustering procedure described above was used to group PFMs into families of matrix similarity. Although these PFMs all show differences in width and appearance, but there exists some similarity or common structure within this set. We organized PFMs into transcriptions factor families for the JASPAR core PFM sets.

We identified 63 similar JASPAR PFM core pairs out of 5565 possible pairs and produced 11 PFM-families. These clusters reflect pronounced similarities in the matrix collections. A complete list of all groups of the similar PFMs in the set is available in the Fig. 2. Motifs that did not cluster with any other motifs are not show. An edge is drawn between motif i and j when Bayes factor is larger than 1 and posterior probability of the null hypothesis that they follow the same product multinomial distribution is larger than 0.8.

Our result shows several relationships, such as the class of NUCLEAR and the class of bHLH. The NUCLEAR family is represented by MA0002, MA0016, MA0066,

MA0071, MA0072 and MA0074 in the JASPAR cores set. The bHLH family is represented by MA0004, MA0006 MA0058, MA0059, MA0085, MA0093 and MA0104 in the JASPAR cores set. Although many of the clusters of TFBSs belong to the same protein family and same species, there are several interesting exceptions. We found that the NUCLEAR and bHLH families previously mentioned contain mostly motifs from Homo sapiens, but the NUCLEAR family also contains a motif (MA0016) from fruit fly Drosophila melanogaster and the bHLE group includes motifs (MA0004, MA0006 and MA0104) from Mus musculus.

We also found that some families contain TFBSs from within a single TF protein family. For example, the bZIP family includes MA0018, MA0096 and MA0097 of bZIP transcription factor. But there are exceptions, such as a family that contains MA0031 of FORKHEAD transcription factor and MA0034, MA0054 of TRP.CLUSTER transcription factor. Clearly, this family would not have been detected if TFBSs were grouped together based only on TF family. It is also interesting to note that most of the larger families contain similar motifs from different species, with motifs from human being grouped with motifs from mouse, fruit fly, and snapdragon flowers.

An interesting collection of structural classes of transcription factors has been compiled recently by Sandelin and Wasserman [8]. Consistent with their results we also found clusters of the NUCLEAR family, bHLH family, bZIP family, ETS family, REL family. There are PFMs of the same transcription factor with different degrees of stringency. For example, the ETS family with high similarity is given which includes MA0026, MA0028, MA0062, MA0076, MA0080, and MA0081 for binding sites of ETS transcription factor. But, we also find some difference with [8]. Many of the families of TFBSs belong to the same protein family and same species, but there are several exceptions such as the NUCLEAR and bHLH families. And most of the larger families contain similar motifs from different species. Different transcription factors of certain families have almost identical binding motifs. For example, a PFM family includes MA0014 and MA0027 for binding site of PAIRED and related HOMEO with Bayes factor 3.607947 and posterior probability 0.9730309.

In Fig. 2, the high similarity of these matrices can not be directly noticed by inspection of names or consensus sequences. Furthermore, subgroups might be detected using our Bayesian statistical approach. Therefore, one could construct "consensus matrices" as in [8] or one might select representative matrices in each cluster.

Our clustering results based on Bayesian hypothesis test identify several PFM-families that suggest that several transcription factor protein families are actually mixtures of several smaller groups of highly similar TFBSs, which provide substantially more refined information compared with the full set of TFBSs in the family. So, we provide a means by which to organize transcription factors based on PFM similarity and can be used to reduce motif redundancy within large database such as JASPAR.

3.2 Synthetic Data

Here, we wanted to test the effectiveness of Pearson χ^2 test and Bayesian hypothesis test in separating PFM pairs generated from the same distribution and PFM pairs generated from different distributions. Since large sets of experimentally verified similar matrix pairs are not available, artificial sets were prepared. We adopt the methodology of Liu

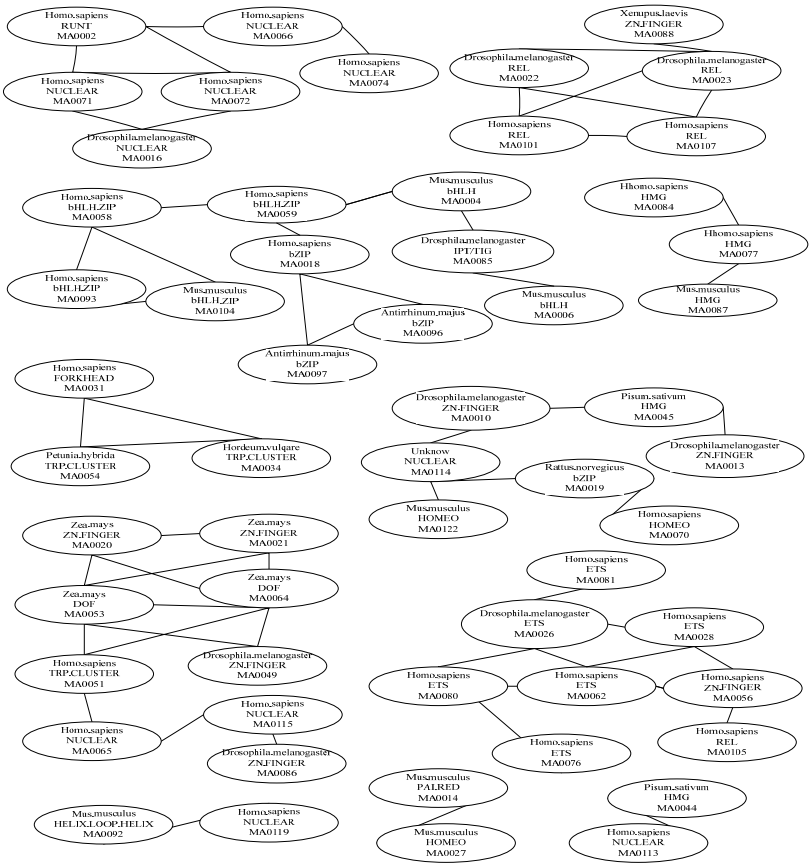


Fig. 2. Graph showing similar matrices: The motifs indicated in ellipses from JASPAR are labeled by the motif name as well as the protein family and species associated with that motif. Motifs that do not cluster with any other motifs are not shown. An edge is drawn between two motifs when Bayes factor is larger than 1 and posterior probability is larger than 0.8.

[10], where PFMs follow a product multinomial distribution. Each column is a set of independent and identically distributed observations. The generated datasets include 100 PFM PFM pairs generated from the same product multinomial distribution and 100 PFM pairs generated from different multinomial distribution. Each PFM was generated by sampling from a Dirichlet distribution with a sample size of 30 and five independent vectors by Matlab. Here, we chose distribution with five vectors and PFMs with 30 sequences in order to match with the average characteristic of JASPAR database.

Sensitivity and Specificity From the statistical point of view, this experiment was executed 1000 times. In Table 4, we summarized the mean results of 1000 runs in the experiment of separating random generated 200 PFM pairs based on the *P*-value provided by Pearson χ^2 test and posterior probability by Bayesian hypothesis test using Jeffreys' prior. The calculation of specificity and sensitivity was performed using the following formula:

$$\text{Specificity} = \frac{\text{True Negative(TN)}}{\text{True Negative(TN)} + \text{False Positive(FP)}},$$

$$\text{Sensitivity} = \frac{\text{True Positive(TP)}}{\text{True Positive(TP)} + \text{False Negative(FN)}}.$$

Finally, Pearson product-moment correlation coefficient [19, 20] was calculated using:

$$\text{Corr.Coef} = \frac{\text{TP} * \text{TN} - \text{FN} * \text{FP}}{\sqrt{(\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TP} + \text{FP}) * (\text{TN} + \text{FN})}}. \quad (27)$$

Correlation coefficient may take any value between − 1 (indicating perfect anti correlation) and 1 (indicating perfect correlation).

We conclude, based on Table 4, that our Bayesian approach is superior to the classical Pearson χ^2 test.

Table 4. Summary of mean results of 1000 runs for various column comparison functions in separating 200 PFM pairs.

Method	True positive	True negative	False positive	False negative	Sp	Sn	Corr.Coef
Pearson χ^2 test	100	89	11	0	0.89	1	0.89
Bayesian test	99	96	4	1	0.96	0.99	0.95

(100 generated from the same multinomial distribution and 100 generated from different multinomial distributions)

4. CONCLUSION

The method for using Bayesian hypothesis test to identify similarity of position frequency matrices for transcription factor binding sites is simple, easy to implement and has a linear time complexity. Our method shows high specificity and high Pearson product-moment correlation coefficient compared to Pearson χ^2 test in separating PFM pairs generated from the same distribution and PFM pairs generated from different distributions. So taking Pearson product moment correlation coefficient as an objective criterion of the performance, our method performs better than the classical methods on average.

We used our technique to classify PFMs in JASPAR into PFM-families. An examination of these families reveals a strong correlation between PFM similarity and the function of the corresponding transcription factors, but there are examples of similar PFMs that profile binding sites of transcription factors that are not likely to be related functionally. It allows for a statistical quantification of errors and is used to facilitate PFM queries in TFBS PFM libraries. In the near future, we will implement our techniques for classifying PFM-families and investigating novel TFBSs.

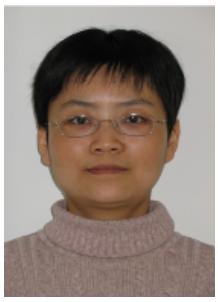
REFERENCES

1. G. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics*, Vol. 16, 2000, pp. 16-23.

2. V. Matys, *et al.*, "TRANSFAC: Transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, Vol. 31, 2003, pp. 374-378.
3. A. Sandelin, W. Alkema, P. Engström, W. Wasserman, and B. Lenhard, "JASPAR: An open-access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Research*, Vol. 32, 2004, pp. 91-94.
4. S. Pietrokovski, "Searching databases of conserved sequence regions by aligning protein multiple-alignments," *Nucleic Acids Research*, Vol. 24, 1996, pp. 3836-3845.
5. J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of Cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae," *Journal of Molecular Biology*, Vol. 296, 2000, pp. 1205-1214.
6. F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by wholegenome mRNA quantitation," *Nature Biotechnology*, Vol. 16, 1998, pp. 939-945.
7. T. Wang and G. Stormo, "Combining phylogenetic data with co-regulated genes to identify regulatory motifs," *Bioinformatics*, Vol. 19, 2003, pp. 2369-2380.
8. A. Sandelin and W. Wasserman, "Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics," *Journal of Molecular Biology*, Vol. 338, 2004, pp. 207-215.
9. D. E. Schones, P. Sumazin, and M. Q. Zhang, "Similarity of position frequency matrices for transcription factor binding sites," *Bioinformatics*, Vol. 21, 2005, pp. 307-313.
10. J. S. Liu, A. F. Neuwald, and C. E. Lawrence, "Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies," *Journal of American Statistics Association*, Vol. 90, 1995, pp. 1156-1170.
11. A. Agresti, "A survey of exact inference for contingency tables," *Statistic Science*, Vol. 7, 1992, pp. 131-177.
12. T. Minka, "Bayesian inference, entropy, and the multinomial distribution," Tutorial, Microsoft Research, University of Cambridge, http://research.microsoft.com/en-us/um/people/minka/papers/multinomial.html, 2003.
13. J. L. Fleiss, B. Levin, and M. C. Paik , *Statistical Methods for Rates and Proportions*, John Wiley and Sons, New York, 2003.
14. T. Andrija and J. O. Edward, "Quality estimation of multiple sequence alignments by Bayesian hypothesis testing," *Bioinformatics*, Vol. 22, 2007, pp. 2488-2490.
15. H. Jeffreys, *Theory of Probability*, Clarendon Press, Oxford, 1961.
16. R. Knuppel, P. Dietze, W. Lehnberg, K. Frech, and E. Wingender, "TRANSFAC retrieval program: A network model database of eukaryotic transcription regulating sequences and proteins," *Journal of Computer Biological*, Vol. 1, 1994, pp. 191-198.
17. G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, Vol. 15, 1999, pp. 563-577.
18. S. M. Kielbasa, D. Gonze, and H. Herzel, "Measuring similarities between transcription factor binding sites," *BMC Bioinformatics*, Vol. 6, 2005, pp. 237-248.
19. M. Burset and R. Guigo, "Evaluation of gene structure prediction programs," *Journal of Genomics*, Vol. 34, 1996, pp. 353-367.
20. M. Tompa, *et al.*, "Assessing computational tools for the discovery of transcription factor binding sites," *Journal of National Biotechnology*, Vol. 23, 2005, pp. 137-144.

Qian Liu (刘倩) received her M.S. and B.S. degrees in College of Mathematics of Information Science from Shaanxi Normal University, China, in 2004 and 2001, respectively. She is currently a doctoral candidate in Xidian University. Her current research interests are bioinformatics, statistical inference and algorithm design.

San-Yang Liu (刘三阳) received his Ph.D. in School of Science from Xi'an Jiaotong University, China, in 1989, M.S. degree in School of Science from Xidian University, China, in 1984 and B.S. degree in College of Mathematics of Information Science from Shaanxi Normal University, China, in 1982. He is both a professor and a doctoral supervisor of School of Science of Xidian University. His research interests are algorithm design, optimization theory, statistical inference and reliability theory.



Li-Fang Liu (刘立芳) received her Ph.D. in School of Computer Science and Technology from Xidian University, China, in 2006, her M.S. and B.S. degrees in School of Computer Science and Technology from the Academy of Equipment Command and Technology, China, in 1998 and 1995, respectively. She is an associate professor in School of Computer Science and Technology of Xidian University. Her current research interests include bioinformatics, machine learning, pattern recognition and intelligent computation.