Bioinformatics Advance Access published August 19, 2004

Similarity of Position Frequency Matrices for Transcription Factor Binding Sites

Dustin E. Schones[1, 2], Pavel Sumazin[1, 3] and Michael Q. Zhang[1]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, [2]Department of Physics and Astronomy, State University of New York, Stony Brook, NY 11794, USA and [3]Computer Science Department, Portland State University, P.O. Box 751, Portland, OR 97207, USA

ABSTRACT

Motivation: Transcription-factor binding sites in promoter sequences of higher eukaryotes are commonly modeled using position frequency matrices. The ability to compare position frequency matrices 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.

Results: We describe a position frequency matrix similarity quantification method based on product-multinomial distributions, demonstrate its ability to identify position frequency matrix similarity and show that it has a better false positive to false negative ratio compared to existing methods.

We group transcription factor binding site frequency matrices from two libraries into matrix families, and identify the matrices that are common and unique to these libraries. We identify similarities and differences between the skeletal-muscle-specific and non-muscle-specific frequency matrices for the binding sites of Mef-2, Myf, Sp-1, SRF and TEF of Wasserman and Fickett (1998). We further identify known frequency matrices and matrix families that are strongly similar to the matrices given by Wasserman and Fickett. We provide methodology and tools to compare and query libraries of frequency matrices for transcription factor binding sites.

Availability: Software is available to use over the web at http://rulai.cshl.edu/MatCompare

Contact: {dschones, sumazin, mzhang}@cshl.edu

Supplementary Information: Database and clustering statistics, matrix families, and representatives are available at http://rulai.cshl.edu/MatCompare/Supplementary

INTRODUCTION

Transcription-factor binding site (TFBS) discovery in promoter sequences is important for predicting transcription regulation. These binding sites are often represented as matrices, which are known in the literature under a variety of names: position weight matrices, position frequency matrices, alignment matrices, profiles, etc (Knuppel et al. (1994), Sandelin et al. (2004); Lenhard and Wasserman (2002)). We refer to a matrix consisting of nucleotide counts per position as a *position frequency matrix* (PFM). Schneider et al. (1982, 1986) and Staden (1984) were some of the first studies to use PFMs to characterize DNA binding site specificity. Berg and von Hippel (1987, 1988), Hertz et al. (1990); Hertz and Stormo (1999), and Stormo and Hartzell III (1989) refined the method to allow quantitative discrimination of sites, with calculated site scores approximating the binding energy of the profiled transcription factor.

Comparison tools for TFBS PFMs are important for testing newly discovered matrices against existing matrices, reducing redundancy in databases and increasing the quality of the matrices. Previous approaches for quantifying PFM similarity include: the average log likelihood ratio method proposed by Wang and Stormo (2003), the Pearson correlation coefficient method described by Pietrokovski (1996) and Hughes et al. (2000), and a method recently introduced by Sandelin and Wasserman (2004).

We describe a column-by-column method for PFM similarity quantification based on the likelihood that aligned columns are independent and identically distributed observations from the same multinomial distribution. We compare the performance of this method to the average log likelihood ratio method and the Pearson correlation coefficient method on simulated data. Our method outperforms the other methods in each of our tests. We do not compare with the method introduced by Sandelin and Wasserman (2004), because it is fundamentally different as they allow for gapped PFM alignment.

We use this PFM similarity quantification to classify TFBSs by PFM similarity. We group PFMs in TRANSFAC (Knuppel et al. (1994)) and JASPAR (Sandelin et al. (2004); Lenhard and Wasserman (2002)) into PFM-families and generate representatives for each family. We find that PFM-families are likely to include TFBS PFMs for related transcription factors. PFM-families and their representatives are useful for reducing the error when searching a PFM library. By comparing the similarity of a novel PFM to a PFM-family representative, as

opposed to all other PFMs, we lower the false positive rate while increasing the computational efficiency. Once a PFM-family is chosen, similarity between its family members and the novel PFM is computed with greater accuracy.

We compare the matrices present in the TRANSFAC database to those in JASPAR, and vice versa. With a similarity threshold of 0.05, 16 of the PFMs from JASPAR are found to have no counterpart in TRANSFAC, including binding site matrices for EN-1, Elk-1, FREAC-3, GATA-2, Gfi, Gklf, HMG-1, MYB.ph3, Pax-2, SAP-1, SQUA, Tal1beta-E47S, c-FOS, c-MYB_1, p50, and Spz1. With a similarity threshold of 0.01, six of the PFMs from JASPAR have no counterpart in TRANSFAC, including binding site matrices for Elk-1, FREAC-3, GATA-2, HMG-1, SAP-1, and Tal1beta-E47S.

We compare the skeletal muscle binding site PFMs given by Wasserman and Fickett (1998) to the independently curated matrices, and to PFM-families and individual PFMs in TRANSFAC. We show that the muscle-specific and the non-muscle-specific binding site matrices for Mef-2 are strongly similar in eight core positions, and different in the remaining positions; the PFMs for Myf are similar in seven core positions; the the PFMs for TEF are similar; and the PFMs for SRF are weakly similar.

In the remainder of the paper we introduce methods for calculating PFM similarity distances, and compare these to PFM similarity measures described previously. We use the methods we introduce to build PFM families in TRANSFAC and JASPAR. We demonstrate the effectiveness of these techniques by producing conclusive comparisons of the PFMs given by Wasserman and Fickett (1998), and by identifying similar PFMs and PFM families in TRANSFAC and JASPAR.

SYSTEMS AND METHODS

In this section we present and compare the performance of four methods for comparing PFMs: the Pearson correlation coefficient, the average log likelihood ratio, the Pearson chi square test, and the Fisher-Irwin exact test. We conclude the section with a description of the clustering methodology used to build PFM-families, and with a description of the PFM libraries themselves. It is important to note that, as expected, all methods perform less effectively when PFMs are built from alignments of few sequences. The Pearson chi square test and Fisher-Irwin exact test allow for power quantification that can be used to determine the confidence level in the similarity of the PFMs.

Distance Measures

We adopt the methodology of Liu et al. (1995), where position frequency matrices follow a product multinomial distribution. 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.

Methods for comparing frequency matrices have been described by Pietrokovski (1996), Hughes et al. (2000), Wang and Stormo (2003), and Sandelin and Wasserman (2004). Pietrokovski tested four different methods for comparing multiple alignments of protein sequences and determined that the Pearson correlation coefficient is the most effective statistic of the four. Hughes et al. employ the Pearson correlation coefficient to compare PFMs. Wang and Stormo introduce the average log likelihood ratio statistic, based on the information content of the binding sites.

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 the Fisher-Irwin exact test or approximated using the Pearson chi square test.

Pearson Correlation Coefficient A general similarity measure between two columns X and Y can be written as in Eisen et al. (1998):

$$S(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - X_{\text{off}}}{\Phi_X}\right)\left(\frac{Y_i - Y_{\text{off}}}{\Phi_Y}\right) \quad (1)$$

where, $\Phi_Z = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_i - Z_{\text{off}})^2}$. For TFBS matrices, we have an alphabet of size four (N = 4). When Z_{off} is set to the mean of Z ($Z_{\text{off}} = \bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$), this similarity measure is the Pearson correlation coefficient (PCC) given in Equation 2. To compare matrices consisting of multiple columns, the scores of the individual column comparisons are summed.

$$PCC(X,Y) = \frac{\sum_{b=A}^{T}(X_b - \bar{X})(Y_b - \bar{Y})}{\sqrt{\sum_{b=A}^{T}(X_b - \bar{X})^2 \sum_{b=A}^{T}(Y_b - \bar{Y})^2}} \quad (2)$$

Average Log Likelihood Ratio The average log likelihood ratio statistic (ALLR), introduced by Wang and Stormo (2003), is a weighted sum of two log likelihood ratios. The ALLR of two column vectors X and Y is given in Equation 3, where n_b is the number of occurrences, $f_b = n_b/\text{N}$ is the frequency, and p_b is the prior for base b. Again,

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

$$ALLR = \frac{\sum_{b=A}^{T} n_{bX} \log(\frac{f_{bY}}{p_b}) + \sum_{b=A}^{T} n_{bY} \log(\frac{f_{bX}}{p_b})}{\sum_{b=A}^{T}(n_{bX} + n_{bY})} \quad (3)$$

Pearson Chi Square The probability that two unnormalized frequency vectors of length 4 are selected from the same multinomial distribution can be described by the p-value of the 2×4 contingency table as seen in Table 1.

	A	C	G	T	
X	n_{xA}	n_{xC}	n_{xG}	n_{xT}	N_x
Y	n_{yA}	n_{yC}	n_{yG}	n_{yT}	N_y
	N_A	N_C	N_G	N_T	N

Table 1. 2×4 Contingency table used for column comparison, with margins.

The chi square statistic of Equation 4 can be used to test the hypothesis that the columns are samples from the same multinomial distribution, where n^o_{jb} is the observed number of base b at position j, and n^e_{jb} is the expected number of base b at position j, calculated as $n^e_{jb} = \frac{N_j N_b}{N}$ (Fleiss et al. (2003)). The p-value is calculated from this χ^2 value with 3 degrees of freedom, and the p-value for multiple columns is the product of the p-values of the individual columns. In our discussion we use the geometric mean of the column p-values, which allows for comparing different size matrices and setting column-based p-value thresholds.

$$\chi^2 = \sum_{j=X,Y} \sum_{b=A}^{T} \frac{(n^o_{jb} - n^e_{jb})^2}{n^e_{jb}} \quad (4)$$

Fisher-Irwin Exact Test The chi square test is an approximation of Fisher-Irwin exact test. The approximation does not hold when the marginal frequencies are small, specifically when at least one of the marginals is smaller than five – a condition that occurs often in PFMs of TFBSs (Fleiss et al. (2003)). The fixed marginal contingency table p-value follows the multiple hypergeometric distribution given in Equation 5 (Agresti (1992)). The two-sided p-value for the table is the sum of the probabilities of all tables that are at least as extreme. As in the χ^2 test, the p-value for multiple columns is the product of the p-values of the individual columns.

$$P = \frac{\binom{N_x}{n_{xA}, n_{xC}, n_{xG}, n_{xT}} \binom{N_y}{n_{yA}, n_{yC}, n_{yG}, n_{yT}}}{\binom{N}{N_A, N_C, N_G, N_T}} \quad (5)$$

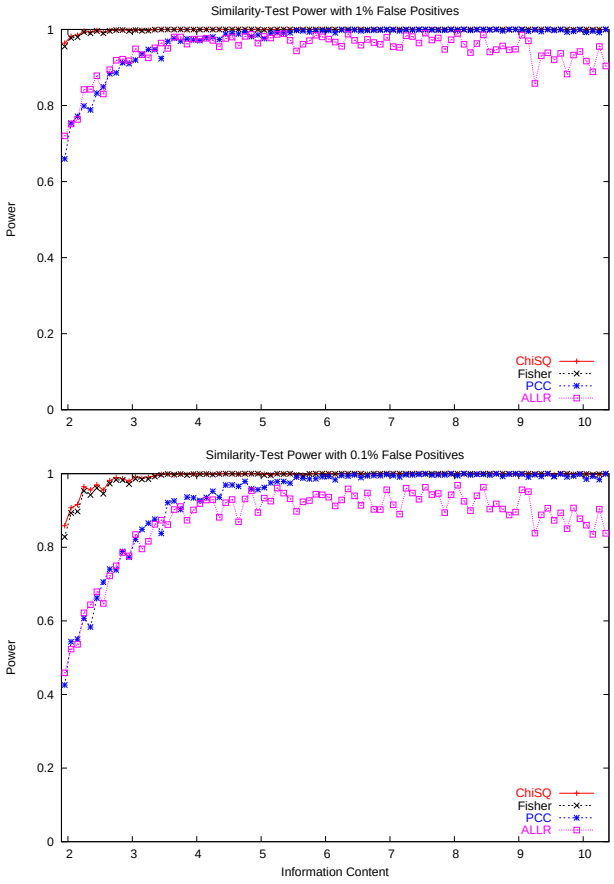

Fig. 1. Power of the methods to recognize PFMs generated from the same product multinomial distribution. Selectivity of ALLR and PCC is poor on low information PFMs.

Distance Measure Comparisons

We generated PFM libraries from product multinomial distributions of a given information content range, and tested the effectiveness of the four methods in separating PFM pairs generated from the same distribution and PFM pairs generated from different distributions; see Figure 1. Each library contains 20 PFMs generated from each of 10 distributions with 6 independent vectors with total information content ranging from 1.9 to 10.4 bits. Each PFM was generated by sampling from a Dirichlet distribution with sample size 30. We generated 220 libraries for each sample in order to achieve suitable power. We chose distributions with 6 vectors, and PFMs with 30 sequences to match with the average characteristics of the extended-core libraries of TRANSFAC and JASPAR. We controlled the false positive rate and compared the power (selectivity) of the four methods. When the false positive rate is set to 0.001, and information content is 3.5 or lower, the hypothesis that the power of the Pearson chi

square test and the Fisher-Irwin exact test is no greater than the power of the other two tests can be rejected with error probability $\alpha = 0.01$ and 99% power. Our experiments suggest that the chi square method is as good as the exact test method in detecting PFM similarity for the majority of PFMs in TRANSFAC, as can be seen in Figure 1.

PFM-family Construction

Two of the most widely used databases of transcription factor binding site matrices are the Transcription Factor Database (TRANSFAC) and the JASPAR database. JASPAR has a much smaller data set, and is manually curated with the goal of eliminating redundancy (Sandelin et al. (2004); Lenhard and Wasserman (2002)).

We describe the clustering of TRANSFAC; the clustering of JASPAR follows in similar lines. TRANSFAC version 7.2 includes 636 matrices, a small subset of which lack sufficient information for PFM construction. We selected all matrices for which we could estimate the correct frequency of each base at each position. This set consisted of 609 matrices.

A matrix core is identified in each PFM by TRANSFAC as the five most conserved contiguous columns (highest confidence) (Knuppel et al. (1994)). Extended cores were constructed to include columns that are adjacent to the cores and whose information content is greater than the information content of the highest entropy column in the core. We used matrix cores and extended cores to measure distances between PFMs.

We compared all PFM core pairs and all extended-core pairs using a sliding window of five columns. The comparisons were ranked according to p-value and a similarity threshold was set so that two PFMs with a p-value below threshold are deemed incompatible, and p-value above threshold are considered similar. We chose the threshold by estimating the associated rate of false positives and false negatives, where the expected number of false positives is the sum of the p-values of the incompatible pairs and the expected number of false negatives is the sum of the q-values ($q = 1 - p$) of the similar pairs.

We set the rate of false positive comparisons to 0.05 and used this to set a p-value threshold. The comparisons with similarity above threshold were then used as input for the *partitioning around medoids* (PAM) clustering algorithm of Kaufman and Rousseeuw (1990) in the S-PLUS software package.

Some of the clusters produced by PAM include pairs with similarity p-value lower than the threshold. These clusters were modified to eliminate pairs with probabilities below the similarity threshold, a process generally resulting in the breaking of a cluster into two or more smaller clusters. For the TRANSFAC cores, this process increased the number of clusters from 135 to 156. The resulting clusters can be described as cliques in the subgraph induced by edges with p-value greater than the threshold. PFM-families are given in the Supplementary Information.

Matrices in the JASPAR database are not annotated with a core section as the TRANSFAC matrices. In order to search in an unbiased manner for JASPAR matrices in TRANSFAC, we defined cores and extended cores in JASPAR matrices in a manner consistent with TRANSFAC. Statistics about these matrix sets are given in the Supplementary Information.

IMPLEMENTATION

In this section we describe the construction of PFM-families using core and extended core sections from matrices in both the TRANSFAC and JASPAR databases. A representative matrix is constructed for each PFM-family. We conclude with a study of the PFMs given by Wasserman and Fickett (1998), describing the similarities and differences between the collected muscle-specific PFMs and independently selected PFMs, and comparing these to TRANSFAC and JASPAR PFMs and PFM-families.

PFM-family Construction

The clustering procedure described above was used to group PFMs into families of matrix similarity. We organized PFMs into PFM-families for the TRANSFAC core, TRANSFAC extended core, JASPAR core and JASPAR extended core PFM sets. We outline the implementation for each of the sets. Statistics and a complete list of the PFM-families in each PFM set are available in the Supplementary Information.

We generate a representative matrix for each PFM-family by first aligning the matrices using a comparison window of five bases, and then summing all the elements across the aligned columns. The summing operation is consistent with the product multinomial model, where each column is a set of observations and the representative column is the combination of the categorical data sets (Liu et al. (1995)).

TRANSFAC Cores The largest PFM-family with high internal similarity is given in Table 2. This PFM-family contains 12 matrices that have an average similarity of 0.94. The ATF, CREB, bZIP910, and bZIP911 factors present in this PFM-family are all members of the bZIP family of proteins. Other CREB matrices exist in TRANSFAC, but are sufficiently distinct and do not appear in this cluster. Relaxed constraints leads to the inclusion of additional bZIP PFMs.

Another interesting result from the clustering of the TRANSFAC core matrices is the presence of multiple E2F binding site PFM-families, as shown in Table 3.

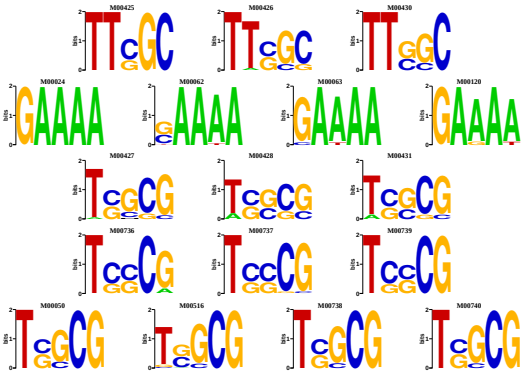

Fig. 2. Sequence Logos PFM-families that include matrices for binding sites of E2F transcription factors corresponding to Table 3.

PFM-families 124, 141 and 156 have identical consensus sequences, but there are differences in the relative strength of the signal at various positions. This can be seen from the sequence logos in Figure 2; the logos program is described in Schneider and Stephens (1990).

Matrix Id	Transcription Factor
M00017	ATF
M00338	ATF
M00483	ATF6
M00039	CREB
M00177	CREB
M00178	CREB
M00801	CREB
M00356	bZIP910
M00357	bZIP910
M00358	bZIP911
M00359	bZIP911
M00036	v-Jun

Table 2. The PFMs in this PFM-family have an extremely high average similarity. The binding site matrix for v-Jun is strongly similar to the other matrices, and is the only binding site matrix in the family for a transcription factor that is not annotated as a bZIP protein.

TRANSFAC Extended Cores We grouped the extended-core PFMs into 145 PFM-families. An example of PFM-families that are formed when using the extended cores and not formed when using the cores is given in Table 4. The factors MyoD, E47, E12, E2A, and myogenin belong to the bHLH (basic region + helix-loop-helix) factor class. MyoD, E47, E12, E2A, and myogenin are known to interact, and the Lmo2 complex transcription factor is known to bind to E2A and E47 (Mitsui et al. (1993)).

Family	Matrix Id	Transcription Factor
40	M00430	E2F-1
40	M00426	E2F
40	M00425	E2F
94	M00024	E2F
94	M00062	IRF-1
94	M00063	IRF-2
94	M00120	d1
124	M00431	E2F-1
124	M00427	E2F
124	M00428	E2F
141	M00736	E2F-1/DP-1
141	M00739	E2F-4/DP-2
141	M00737	E2F-1/DP-2
156	M00050	E2F
156	M00516	E2F
156	M00738	E2F-4/DP-1
156	M00740	Rb/E2F-1/DP-1

Table 3. PFM-families that include matrices for binding sites of E2F transcription factors. Logos for the PFM cores are given in Figure 2.

Family	Matrix Id	Transcription Factor
108	M00001	MyoD
108	M00002	E47
108	M00071	E47
108	M00693	E12
108	M00412	AREB6
108	M00414	AREB6
109	M00184	MyoD
109	M00804	E2A
109	M00712	myogenin
109	M00277	Lmo2 complex

Table 4. PFM-families that include matrices for the binding sites of MyoD in the TRANSFAC extended core set. The relationship between the members of family 108 are lost when considering TRANSFAC cores only. Family 109 is represented by M00184 and M00804 in the TRANSFAC cores set. MyoD, E47, E12, E2A and myogenin are bHLH class factors (Mitsui et al. (1993)).

JASPAR Cores and Extended Cores We identified 61 similar JASPAR PFM core pairs and 80 similar JASPAR PFM extended-core pairs out of the 6431 possible pairs, and produced 23 and 36 PFM-families respectively. The

average similarity within clusters and the size of clusters is considerably smaller for the JASPAR database than for the TRANSFAC database. The clustering statistics are given in the Supplementary Information.

Representative Matrices

Through increasing the number of observations for each high information column and decreasing the number of PFMs in the initial search, representative matrices are used to increase the accuracy and the efficiency of similarity searches. An example of this is given in the following section.

Similarity designations are often difficult to make for TRANSFAC PFMs because PFMs are often constructed from the alignment of few sequences. PFM-family representatives allow for increased accuracy since they represent richer alignments.

Representative matrices can also be used to validate PFM-families. The representatives can be used to search the original database for related matrices. When doing this, matrices of the PFM-family corresponding to the representative matrix should be found with high similarity, followed by members of related families. The result of a query using the representative of the bZIP family introduced in Table 2 is shown in Table 5. Results for the other representative matrices created from the TRANSFAC extended cores are in the Supplementary Information.

Matrix ID	Factor	Score
M00483	ATF6	1
M00359	bZIP911	1
M00358	bZIP911	1
M00357	bZIP910	1
M00356	bZIP910	1
M00338	ATF	1
M00036	v-Jun	1
M00017	ATF	1
M00694	E4F1	1
M00179	CREB-BP1	1
M00115	Tax/CREB	1
M00694	E4F1	1
M00178	CREB	0.921834
M00177	CREB	0.891643
M00039	CREB	0.644119
M00697	HBP-1b	0.563122
M00113	CREB	0.376415
M00114	Tax/CREB	0.231502
M00513	ATF3	0.132264
M00514	ATF4	0.119509
M00801	CREB	0.0887971

Table 5. PFMs whose similarity with the representative of the PFM-family given in Table 2 is above threshold. PFMs with similarity lower than 0.64 are not family members.

Novel PFM Comparison

Wasserman and Fickett (1998) curated a set of PFMs for skeletal muscle-specific TFBSs and compared them to PFMs from independently selected promoter segments. They wanted to know if the two resulting PFMs in each pair differ substantially, and they offered observations about the differences. We describe the difference in quantifiable terms, and compare the PFMs to general PFMs from TRANSFAC.

The reader is referred to Wasserman and Fickett (1998) for the PFMs classified as muscle-specific and independent. We compared the analogous PFMs from the muscle-specific and independent set. A summary of the results follows.

- *Mef-2 PFMs* – the muscle-specific and independent PFMs match well from position 4 to 11 (with similarity 0.21), and match weakly from position 1 to 4 (0.05).
- *Myf PFMs* – match well in 7 positions starting at position 4 of the muscle-specific PFM and 5 of the independent PFM.
- *SRF PFMs* – match weakly in 10 positions starting at position 3 of the muscle-specific PFM and 5 of the independent PFM (0.06).
- *Sp-1 PFMs* – match well in 10 positions starting at position 1 of the muscle-specific PFM and 2 of the independent PFM (0.49). However, the power of the comparison is lower than 90%.
- *TEF PFMs* – match well in 8 positions starting at position 2 of the muscle-specific PFM and 1 of the independent PFM (0.43). However, the power of the comparison is lower than 85%.

We compared the skeletal muscle-specific PFMs to the representative PFMs of TRANSFAC extended-core PFM-families. A summary of the results follows.

- *Mef-2 PFM* – matches best with the representative of the sixth PFM-family (M06), with a similarity of 0.45 for window of size 7 starting at position 4. M06 includes binding site matrices for aMEF-2, MADS-B and MEF-2. The PFM also matches M00006 (MEF2) in 10 positions, starting at its second position. However, M00006 is constructed from the alignment of five sequences and the similarity has considerably lower power than the similarity of the PFM with the representative of M06.
- *Myf PFM* – Myf PFMs in TRANSFAC did not meet our requirements and were removed from all analysis. The Myf PFM did not match any PFM-families.

- *SRF PFM* – matches the representative of PFM-family 75 (M075) with window of size 7 and similarity 0.14. The PFM-family includes a binding site matrix for BR-C_Z2 and two binding site matrices for SRF. The PFM matches, with window size 7, M00186 (SRF) with similarity 0.323518, M00404 (MADS-B) with similarity 0.249068, and M00810 (SRF) with similarity 0.248197. The match with M00404 is most likely false. M00404 is constructed from the alignment of seven sequences and its match with the SRF PFM begins at position 5 instead of position 3. Interestingly, the muscle-independent SRF PFM strongly matches M00152 (SRF) with a window size 10 and similarity 0.75. Thus, the muscle-specific and independent SRF PFMs match different SRF binding site matrices in TRANSFAC.
- *Sp-1 PFM* – matches the representative of PFM-family 28 (M028) with a window of size 6 and similarity 0.31. The PFM-family includes a binding site matrix for Muscle initiator sequences-19 and Muscle initiator sequences-20. The PFM matches, with window size 7, M00221 with similarity 0.30 and M00749 with similarity 0.24. Both these matches are likely to be false. M00221 is constructed from the alignment of seven sequences and M00749 from six sequences. Their alignments with the Sp-1 PFM start at different positions.
- *TEF PFM* – does not match any PFM-family.

DISCUSSION AND CONCLUSION

We present a technique to identify similarity between position frequency matrix profiles for transcription factor binding sites. This similarity method is deeply rooted in the theory of PFMs and is experimentally shown to outperform existing methods. It allows for a statistical quantification of errors, and is used to facilitate PFM queries in TFBS PFM libraries.

We used our technique to classify PFMs in TRANSFAC and JASPAR into PFM-families, which were then used to increase the accuracy of PFM queries. 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.

The analysis of the TRANSFAC and JASPAR databases reveals that the JASPAR database is less redundant, but almost all of the JASPAR matrices are represented in TRANSFAC. By grouping the TRANSFAC PFMs into PFM-families we build a higher quality PFM set that is also less redundant.

We also show that the cores in the TRANSFAC database do not always capture the whole signal. For example, the JASPAR PFM MA0003 is strongly related to the TRANSFAC M00075 and both are binding sites of an E2F transcription factor. However, the similarity between the PFMs cannot be detected when using the M00075 core alone. Another example is the similarities between the bHLH factors listed in Table 4. These PFMs only group together as similar when the extended cores are considered.

Sandelin and Wasserman (2004) use Needleman-Wunch (Needleman and Wunsch (1970)) to align matrices before comparing them. This method is attractive for comparing binding site PFMs that are composed of strongly conserved position clusters that are separated by non-conserved positions, such as binding sites for dimers like the leucine zippers. We chose to concentrate on the simpler configuration of adjacent, conserved positions as advocated by TRANSFAC. However, the extension to allow for gapped PFM alignments is possible and would be useful.

We compared the skeletal muscle-specific PFMs curated by Wasserman and Fickett (1998) to their corresponding independently curated PFMs. We show that the muscle-specific SRF binding site matrix is different from the independent SRF binding site matrix; these matrices match different binding site matrices in TRANSFAC. All other muscle-specific binding site matrices in the Wasserman and Fickett study are similar.

Finally, our major contribution is a methodology for comparing PFMs and for searching for PFMs in a PFM library. Our techniques can be used for classifying PFM-families, and for investigating novel PFM binding site matrices.

ACKNOWLEDGMENTS

This work was supported by NSF grants EIA-0324292 and DBI-0306152.

REFERENCES

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science 7*, 131–177.

Berg, O. G. and P. von Hippel (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology 193*(4), 723–750.

Berg, O. G. and P. von Hippel (1988). Selection of DNA binding sites by regulatory proteins II: The binding specifi city of cyclic amp receptor protein to recognition sites. *Journal of Molecular Biology 200*(4), 709–723.

Eisen, M., P. Spellman, P. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA 95*, 14863–14868.

Fleiss, J. L., B. Levin, and M. C. Paik (2003). *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Hertz, G., G. Hartzell III, and G. Stormo (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in the Biosciences 6*(2), 81–92.

Hertz, G. and G. Stormo (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics 15*(7), 563–577.

Hughes, J. D., P. W. Estep, S. Tavozoie, and G. M. Church (2000). Computational identification of *Cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology 296*, 1205–1214.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data - An Introduction to Cluster Analysis*. John Wiley & Sons.

Knuppel, R., P. Dietze, W. Lehnberg, K. Frech, and E. Wingender (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *Journal of Computational Biology 1*(3), 191–198.

Lenhard, B. and W. W. Wasserman (2002). TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics 18*(8), 1135–1136.

Liu, J. S., C. E. Lawrence, and A. Neuwald (1995). Bayesian models for multiple local sequence alignment and its Gibbs sampling strategies. *Journal of the American Statistical Association 90*, 1156–70.

Mitsui, K. K., M. Shirakata, and B. M. Paterson (1993). Phosphorylation inhibits the DNA-binding activity of MyoD homodimers but not MyoD-E12 heterodimers. *Journal of Biological Chemistry 268*(32), 24415–24420.

Needleman, S. and C. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48*, 443–453.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research 24*(19), 3836–3845.

Sandelin, A., W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard (2004). JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research 32*(1), D91–D94.

Sandelin, A. and W. W. Wasserman (2004). Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *Journal of Molecular Biology 338*, 207–215.

Schneider, T. D. and R. M. Stephens (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research 18*, 6097–6100.

Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research 10*(9), 2997–3011.

Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht (1986). Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology 188*(3), 415–31.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research 12*, 505–519.

Stormo, G. D. and G. Hartzell III (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the USA 86*, 1183–1187.

Wang, T. and G. D. Stormo (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics 19*(18), 2369–2380.

Wasserman, W. W. and J. W. Fickett (1998). Identification of regulatory regions which confer muscle-specific gene expression. *Journal of Molecular Biology 278*(1), 167–181.