Improvement of TRANSFAC Matrices Using Multiple Local Alignment of Transcription Factor Binding Site Sequences

Yutao Fu[1] bibin@bu.edu
Zhiping Weng[1,2] zhiping@bu.edu

[1] Bioinformatics Program, Boston University, Boston, MA 02215, USA
[2] Biomedical Engineering Department, Boston University, Boston, MA 02215, USA

Abstract

This paper describes a novel approach to constructing Position-Specific Weight Matrices (PWMs) based on the transcription factor binding site (TFBS) data provide by the TRANSFAC database and comparison of the newly generated PWMs with the original TRANSFAC matrices. Multiple local sequence alignment was performed on the TFBSs of each transcription factor. Several different alignment programs were tested and their matrices were compared to the original TRANSFAC matrices. One of the alignment programs, GLAM, produced comparable matrices in terms of the average ranking of true positive sites across the whole test set of sequences.

Keywords: GLAM, TRANSFAC, matrix

1 Introduction

Discovery and prediction of sequence motifs such as transcription factor binding sites (TFBSs) is an important topic in genomic research. Current research efforts in this area can be divided into two categories unsupervised and supervised methods. Unsupervised motif discovery, also known as the *ab initio* approach, searches for short local sequence alignment using multi-dimensional dynamic programming, expectation maximization or Gibbs sampling. Prior knowledge of the binding preferences of a large number of transcription factors (TFs) is represented in the form of position-specific weight matrices (PWMs) and can be applied in various supervised motif discovery, such as simple pattern scan and cis-element overrepresentation [4]. The JASPAR database [7] contains PWMs for many TFs. Another important collection is the TRANSFAC database [6] maintained by BIOBASE in Germany.

The detailed algorithms used by TRANSFAC to construct its PWMs have not been published. To provide tools for automatic construction of PWMs from TFBSs, we present here results on the comparison between TRANSFAC matrices and matrices constructed using several local sequence alignment programs. We tested four such programs, including GLAM (Gapless Local AlignMent; Dec 15 2003) [3], MotifSampler (v3.0) [8], MEME (Multiple Expectation maximization for Motif Elicitation, v3.0) [1] and AlignACE (Aligns Nucleic Acid Conserved Elements, v2.3) [5].

2 Method and Results

2.1 Data Sources

TRANSFAC® Professional 7.4.1 was downloaded from `http://www.biobase.de/`. The files involved are listed as follows: 1) gene.dat: accession numbers and references for 6692 sequences 2) matrix.dat: 695 TRANSFAC matrices and 3) site.dat: 13302 site locations.

We selected a subset of TRANSFAC data according to the following criteria: 1) The sequences can be retrieved using their RefSeq or EMBL accession numbers; 2) The TFBSs can be uniquely located within the sequences by string matching; and 3) For each matrix, there are at least three qualified TFBSs.

Through NCBI's fetch service, 5231 sequences were retrieved using RefSeq or EMBL accession numbers in the TRANSFAC gene.dat file. Among the 959 TFBSs uniquely mapped in these sequences, 757 TFBSs grouped into 96 datasets, each of which corresponds to a TF. TFBSs corresponding to other TFs were discarded because these TFs contained fewer than 3 TFBSs.

2.2 Matrix Construction and Performance Comparison

Sequences longer than 500bp were chopped to 500bp containing at least one TFBS (denoted as test sequences hereafter). Each TFBS was excised with flanking sequences on both sides, which are of the same length as the annotated binding site itself. These sequences were saved in FASTA format as input for multiple local alignment programs (denoted as running sequences hereafter). All programs were downloaded from the websites the authors provided (Table 1).

Table 1: Commands Used for Running Various Programs.

Program	Command line
GLAM	glam ⟨input.fasta⟩
MotifSampler[a]	MotifSampler -w ⟨W⟩ -b ⟨background⟩ -M1 -f ⟨input.fasta⟩ -o ⟨output⟩
MEME	meme ⟨input.fasta⟩ -dna -minw 4 -maxw 40 -nostatus -text -print_fasta -maxiter 10000 -mod oops -revcomp
AlignACE[a]	AlignACE -w ⟨W⟩ -i ⟨input.fasta⟩ -o ⟨output⟩
Possum[b]	Possum ⟨PWM⟩ -t-99 ⟨input.fasta⟩

[a] ⟨W⟩ indicates the length of the corresponding TRANSFAC matrix. ⟨background⟩ indicates a background model constructed using human genomic promoter sequences.

[b] The '-t' option for Possum is the score threshold for reporting matches.

Each resulting local alignment was converted into a PWM by counting the number of different nucleotides in each position [4]. We used a simple matrix scan program, Possum [4], to generate a log-likelihood-ratio score for the PWM for each possible word in the test sequences. Among the words overlapping with the TRANSFAC-annotated TFBS in a sequence, the one with the highest score was determined to be the true-positive match. The score of the true-positive match in a test sequence was ranked against the ordered list of scores of all possible words, producing a PWM-specific rank score.

The sign test was performed using rank scores from TRANSFAC matrices and matrices obtained using different multiple local alignment algorithms. P-values of sign tests were calculated using the binomial distribution. This test provided a robust but simple way to compare the overall performance in the whole data set regardless of the magnitude of individual differences.

2.3 Results

Alignments of various lengths were obtained by applying multiple local sequence alignment programs to the running sequences for each dataset. The optimal alignment output by each program was used to construct a PWM for that dataset. Different running parameters for each program were tested and the combinations that produced the best rank scores were listed in Table 1.

GLAM, MotifSampler and MEME produced alignments for all datasets. Table 2 listed the width of the alignments found by each program. Note that MotifSampler used TRANSFAC matrix width

as its input ($\langle W \rangle$ in Table 1), therefore reproduced its width distribution. AlignACE also used this input, but reported "no alignment found" for 23 datasets, with the average alignment width for the rest of the datasets 1 bp shorter than TRANSFAC. GLAM and MEME automatically determined the alignment width corresponding to each TF.

The difference in matrix widths can be ignored given the nearly 1000 possible words for each test sequence (most of them were double-strand 500-bp-long sequences). Ranking true positive hits among all the words according to their log-likelihood-ratio scores demonstrated the relative performance of the matrices constructed using different alignments. A rank score of 999 was assigned to AlignACE when no alignment was produced and therefore no matrix was available. Table 3 summarized the results of pair wise comparisons between TRANSFAC and GLAM, TRANSFAC and MotifSampler, TRANSFAC and MEME, and TRANSFAC and AlignACE. Also shown are P-values for the pairwise comparison using the Sign Test. From the P-values, we can see that it is statistically significant that MotifSampler, MEME and AlignACE produced worse PWMs than TRANSFAC. GLAM outperformed TRANSFAC, although the difference is not statistically significant (P-value = 0.157).

Table 2: Lengths of Multiple Local Alignments.

Alignment	Min	Max	Average
MotifSampler (TRANSFAC)	6	26	12.4
GLAM	6	48	15.2
MEME	4	38	13.4
AlignACE (TRANSFAC)	6	26	11.4

Table 3: Rank Score Comparison against TRANSFAC Matrices.

Program	$+^a$	$-^b$	P-value (Sign Test)
GLAM	399	358	0.157
Motif Sampler	237	520	2.39e-25
MEME	316	441	5.36e-06
AlignACE	125	632	1.56e-82

[a]Number of constructed matrices that produced lower or equal rank scores (more desirable results) than the corresponding TRANSFAC matrices.

[b]Number of constructed matrices that produced higher rank scores (less desirable results) than the corresponding TRANSFAC matrices.

3 Discussion

Matrix construction is an important link between unsupervised and supervised motif discovery and is essential for achieving a satisfactory level of specificity. Although TRANSFAC is a widely useful database of PWMs, the method for its matrix construction has not yet been made publicly available, limiting its use in situations when novel TFBSs are available for constructing the PWMs.

The purpose of calculating log-likelihood-ratio scores using matrices scanning programs is to differentiate true motif hits from sequence background. The higher a word is ranked (lower rank scores) among an ordered list of scores that are associated with all available words in the test sequence, the more likely it is the true TFBS. From this point of view, matrix performance can be measured by the rank scores of true TFBSs in the running sequences. In our study, GLAM is the only program

that produced comparable rank scores to TRANSFAC matrices. Matrices generated by the other 3 programs introduced significantly higher rank scores, causing more false-positives or false-negatives depending on the scoring thresholds used.

GLAM is a Gibbs-sampling-based alignment program [3]. One of the most noteworthy features of GLAM is its ability to automatically determine the width of an alignment, unlike MotifSampler and AlignACE, which require user-designated motif width. For MotifSampler and AlignACE the rank scores improved when the sizes of the appropriate TRANSFAC matrices were used as input in contrast to using a fixed alignment width for all datasets, while for GLAM the performance was not significantly different whether such an input was specified or not (data not shown).

In addition to size effects, multiple alignment programs may construct improved alignments and therefore improve the rank scores, especially when a large number of running sequences are available. An example is the matrix for AP-1 binding sites (Figure 1 and Table 4). The information content was biased toward the 5' side for the TRANSFAC matrix and both of its end positions were highly degenerate. GLAM was able to partially correct these peculiarities and achieve lower rank scores.

Table 4: Rank Scores for AP-1 Binding Sites Calculated with Different PWMs.

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
TRANSFAC	1	56	1	2	2	3	2	2	1	6	1	1	5	1	1	17	4	1	2	2	1
GLAM	1	4	1	1	1	2	1	1	2	1	1	1	9	1	1	2	1	1	1	3	1

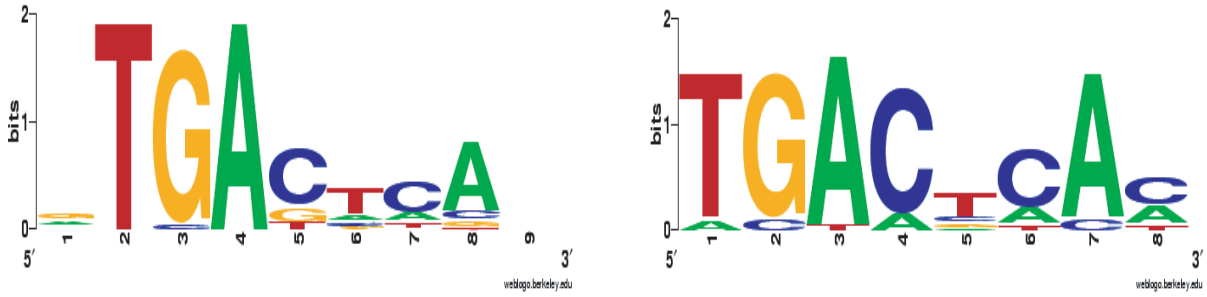

Figure 1: Sequence logos of TRANSFAC (9 bp, left) and GLAM (8 bp, right) matrices [2].

We point out several caveats in interpreting the results. The lack of TFBS mapping information in TRANSFAC sequences due to ambiguous reference points may cause confusion. As a result, only a small subset ($\sim 5\%$) of all known TFBSs were tested (corresponding to the sites for which unambiguous mapping information was available), therefore the PWMs obtained from alignment programs may be biased toward these sites. Cross validation was not performed due to the absence of the matrix construction algorithm of TRANSFAC. We also need to be cautious interpreting the comparison results, due to the complicated relationship between mathematical optimum and biological significance. Access to more genuine sequences annotated with accurate TFBS locations will be the key for future studies.

In conclusion, by examining rank scores of true TFBSs obtained using different PWMs, we showed that GLAM provided an alternative algorithm for matrix construction with comparable performance to TRANSFAC matrices. This may be useful for reducing human burden in curation of novel TFBS-containing sequences and to facilitate streamlined high-throughput analysis.

We thank Dr. Heather Burden for proofreading this paper.

References

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

[2] Crooks, G.E., Hon G., Chandonia J.M., and Brenner S.E., WebLogo: A sequence logo generator, *Genome Res.*, 14:1188–1190, 2004.

[3] Frith, M.C., Hansen, U., Spouge, J.L., and Weng, Z., Finding functional sequence elements by multiple local alignment, *Nucleic Acids Res.*, 32(1):189–200, 2004.

[4] Fu, Y., Frith, M.C., Haverty, P.M., and Weng, Z., MotifViz: An analysis and visualization tool for motif discovery, *Nucleic Acids Res.*, 32(Web Server issue):W420–W423, 2004.

[5] Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M., Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*, *J. Mol. Biol.*, 296(5):1205–1214, 2000.

[6] Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E., TRANSFAC: Transcriptional regulation, from patterns to profiles, *Nucleic Acids Res.*, 31(1):374–378, 2003.

[7] Sandelin, A., Alkema, W., Engström, P., Wasserman, W., and Lenhard, B., JASPAR: An open access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res.*, 32(Database issue):D91–D94, 2004.

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