W438–W441 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki590

T-Reg Comparator: an analysis tool for the comparison of position weight matrices

Stefan Roepcke*, Steffen Grossmann, Sven Rahmann[1] and Martin Vingron

Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany and
[1]Department of Genome Informatics, University of Bielefeld, 33594 Bielefeld, Germany

Received February 14, 2005; Revised April 19, 2005; Accepted May 3, 2005

ABSTRACT

T-Reg Comparator is a novel software tool designed to support research into transcriptional regulation. Sequence motifs representing transcription factor binding sites are usually encoded as position weight matrices. The user inputs a set of such weight matrices or binding site sequences and our program matches them against the T-Reg database, which is presently built on data from the Transfac [E. Wingender (2004) *In Silico Biol.*, 4, 55–61] and Jaspar [A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman and B. Lenhard (2004) *Nucleic Acids Res.*, 32, D91–D94]. Our tool delivers a detailed report on similarities between user-supplied motifs and motifs in the database. Apart from simple one-to-one relationships, T-Reg Comparator is also able to detect similarities between submatrices. In addition, we provide a user interface to a program for sequence scanning with weight matrices. Typical areas of application for T-Reg Comparator are motif and regulatory module finding and annotation of regulatory genomic regions. T-Reg Comparator is available at http://treg.molgen.mpg.de.

INTRODUCTION

The binding of transcription factors to target DNA in a sequence-specific manner is a key step in transcriptional regulation. Binding affinities of transcription factors are often described by position weight matrices (PWMs). These matrices specify the sequence motif by giving a base distribution for each of its positions. Transfac and Jaspar (1,2) are two well-recognized projects that aim at giving comprehensive collections of eukaryotic transcription factors and descriptions of their respective binding affinities in terms of PWMs. In order to understand the binding specificity of transcription factors, it is essential to be able to relate newly derived motifs to the existing collections. Our work supports the comparison of weight matrices.

Various large-scale approaches are nowadays available for identifying novel binding sites. A typical bioinformatics analysis might encompass a search for over-represented sequence motifs in the promoters of co-regulated genes, e.g. based on a software package such as MEME (3). Likewise, wet-lab techniques including SELEX (4) and chromatin immunoprecipitation (5,6) may provide data from which PWMs are derived. The growing number of PWMs leads to the problem of distinguishing new from old: for a newly derived PWM it is not immediately clear whether it describes an already known binding site or whether it is actually new information. This problem is further aggravated by the fact that the binding specificity of many transcription factors is not very pronounced. In practice, available PWMs can be very short (4–6 positions) or can contain a substantial number of uninformative positions. For some well-studied transcription factors several highly similar PWMs are reported in Transfac.

Recent papers describe a number of comparison methods for weight matrices (7–10). Schones and co-authors focus on the question of determining the most adequate measure of similarity (7). They restrict each matrix to a core part and base their method on the product-multinomial distribution. Sandelin and Wasserman cluster the weight matrices of Jaspar and propose a grouping into familial binding profiles (8). In two papers by the group of De Moor a similarity measure based on the Kullback–Leiber distance has been used to compare weight matrices (9,10).

Our web application T-Reg Comparator is dedicated to supporting studies into transcriptional regulation. Lists of motifs can be compared against weight matrices in the T-Reg database, which currently contains all PWMs from Transfac and Jaspar (1,2). For convenience we facilitate the input of sets of binding sites or alignments as well. An important feature of our tool is that partially overlapping but high-scoring matches are recognized as well. A detailed report tells the user whether the identified motif is novel or resembles the binding site of some known transcription factor. In addition,

*To whom correspondence should be addressed. Tel: +49 30 84131159; Fax: +49 30 84131152; Email: roepcke@molgen.mpg.de

T-Reg Comparator provides an interface to a program developed in our group for the scanning of sequences with weight matrices (11).

MATERIALS AND METHODS

For weight matrix comparison we use a method that has already been introduced by De Moor and colleagues (9,10). It is based on a symmetrized, position-averaged Kullback–Leibler distance. To compare two weight matrices: the shorter one is moved along the other and all shifted positions that satisfy the following three conditions are considered. First, at least half of the shorter matrix has to overlap with the longer matrix. Second, this overlap has to be at least four positions long. And third, the overlapping part of at least one matrix has to have a position-averaged entropy below one with respect to the natural logarithm. For each such shift, a position-normalized dissimilarity score is calculated for the overlapping part, and the smallest dissimilarity score is used to measure the overall similarity between the two matrices.

Our web service works on the T-Reg database. The T-Reg database is an in-house relational database on transcriptional regulation that currently contains data from Jaspar and Transfac version 8.4 (1,2). Data from Jaspar and Transfac Public is freely accessible on our web site. However, T-Reg cannot be made publicly available for download because it contains data from Transfac that we are not allowed to redistribute. The Jaspar data was obtained from the project web site, http://jaspar.cgb.ki.se.

The input consists of a set of weight matrices or sets of sequences, which can be entered into a text field or uploaded from a file. In the current version of the program we support several matrix formats: MEME, Transfac and Jaspar file formats and a raw data format. Sequence sets are given in FASTA format, or simply one sequence per line, and are then used to generate weight matrices. If the user inputs an alignment, weight matrices are constructed by counting the occurrences of bases in each position. If, instead, the user inputs unaligned binding site sequences, DiAlign (12) is utilized to compute a multiple alignment first. After specifying the input, the user chooses the set of matrices with which to make the comparison. For example, it is possible to restrict the comparison to the set of available vertebrate matrices in the public version of Transfac. Smaller sets of matrices lead to shorter computation times and results that can be easily interpreted. The dissimilarity score ranges from 0 to 5. We recommend a cutoff of 0.8 or 0.5, where the latter produces more specific results.

After all pairwise comparisons have been made, the application returns a table that contains the following information: the name and the consensus of the input matrix and a list of matrices with divergence smaller than the given cutoff. For these matrices, the name, the overlap, the orientation, the shift, the actual dissimilarity score and the consensus of the best match are provided. Further, a grouping of the transcription factors into coarse classes based on the structure of the DNA-binding domain is given in the annotation. Hyperlinks guide the user to the web pages of the source databases and additional information, Jaspar, Transfac Public, Transfac or TESS (13). If the matrix stems from the non-public part of Transfac, the hyperlink points to the Biobase website (http://www.biobase.de), where access is restricted to licensed users. A standalone version of T-Reg Comparator will be made available at our web site.

As a natural step in studies on transcriptional regulation, a user can start a sequence search with the input matrices. The annotation with the PWM hits on the sequence is based on an elaborate, statistically sound method that has previously been developed in our group (11). The software is freely available at http://genereg.molgen.mpg.de/ProfileStats/index.shtml. Two aspects of this method should be mentioned here. First, in the process of constructing a scoring matrix from the PWM, we use a regularization method that does not change the overall nucleotide composition of the profile and regularizes each position relative to its signal strength. Second, for the final scoring matrices, we calculate exact score distributions under a background and a signal model for motif sequences. This allows the determination of the scanning cutoff using statistical considerations, which is better than making an *ad hoc* decision such as setting the cutoff to 80% of the PWM's overall score range. In the current interface the user can choose an accepted false-negative rate. The output consists of a list of all the matches better than the cutoff. The orientation, the position and the false-positive estimate for each hit are printed out.

RESULTS AND DISCUSSION

We demonstrate the functionality of T-Reg Comparator on an example (Figures 1 and 2). We have constructed an example MEME file from a promoter analysis of ribosomal protein genes. The file is available in the Supplementary material or via the help page of T-Reg Comparator. The three weight matrices represent the typical cases that occur during motif discovery endeavours. Motif 1 is quite unspecific and matches other unspecific matrices best. It shows weak similarity to binding sites of the STAT family of transcription factors. When compared against Jaspar, the PWM Motif 8 matches MA0028 for the transcription factor Elk-1 best. The sequence logo for MA0028 is depicted in Figure 1 and the comparison is illustrated in Figure 2. Motif 8 resembles a typical binding site of a factor of the ETS family because it contains the characteristic core motif GGAA. The two positions preceding GGAA

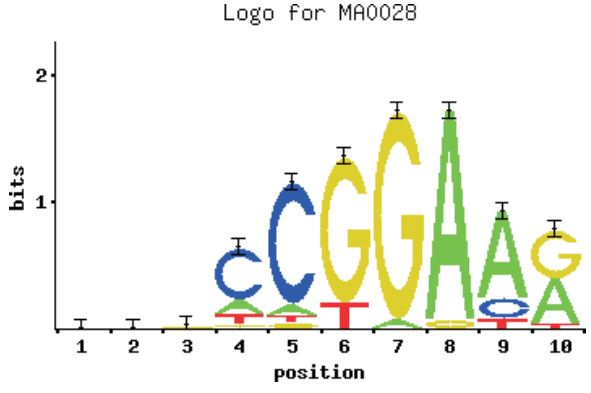

Figure 1. Sequence logo representation of the binding specificity of the transcription factor Elk-1, copied from the Jaspar web site, http://jaspar.cgb.ki.se (identifier MA0028). The height of each column indicates the information content of the corresponding position. The sizes of the characters represent the relative frequency of the corresponding bases.

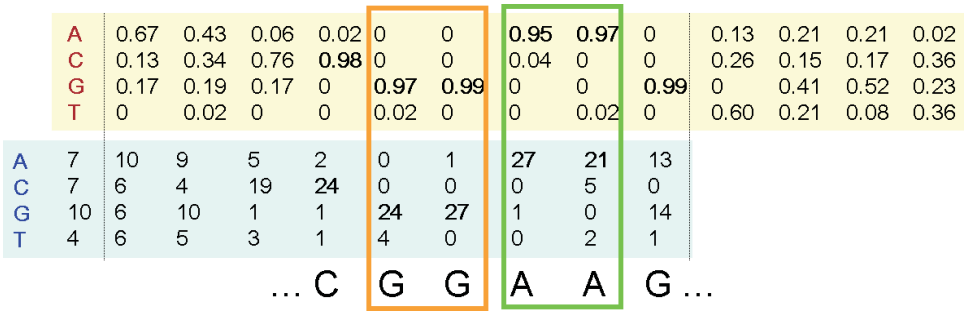

Figure 2. Representation of the comparison of two weight matrices. The upper, yellow one is a probability matrix taken from the example MEME file (Motif 8 in the example file on the help page). The lower, blue one is the Jaspar count matrix MA0028 for Elk-1. The divergence computed by T-Reg Comparator amounts to 0.644. Thin dashed lines indicate the overlapping part. The frames depict the core consensus GGAA of the ETS family of transcription factors.

are also highly informative and similar in both matrices. However, the position following GGAA is dissimilar in the two motifs: Motif 8 contains an unambiguous G whereas MA0028 has an A or a G. In summary, Motif 8 is probably a binding site for factors of the ETS family but not necessarily of Elk-1. Indeed, it has been shown previously that another ETS transcription factor, GABP, binds some ribosomal proteins' gene promoters (14). The third PWM in the example file, Motif 5, shows only poor similarity to other matrices. Hence, this motif can be regarded as novel, at least to the T-Reg database.

T-Reg Comparator is a tool designed to support researchers in identifying novel transcription factor binding sites. There are many situations in which researchers come up with weight matrices that describe the binding specificity of a set of transcription factors of interest. Identifying over-represented sequence patterns in sets of regulatory regions *in silico* (15) or performing in-depth analyses of the binding specificity of DNA-binding proteins *in vitro* (16) are just two of many examples.

In all these cases, there is a need to check whether some of the sequence patterns match the already described binding specificity of a transcription factor. To this end, the newly identified PWM must be compared with available data collections such as Transfac and Jaspar. However, these databases do not provide tools or data structures to address this question. In addition, there are further specific issues that arise when comparing weight matrices. First, PWMs stored in the databases are frequently very short or have many uninformative positions. Second, small PWMs can be parts of larger, modular PWMs (17). And third, single transcription factors can be associated with two or more PWMs, which sometimes differ substantially.

T-Reg Comparator is tailored to handle the situation described above. To achieve this, we use a dissimilarity score based on the symmetrized, position-averaged relative entropy, which has already been introduced by De Moor and colleagues for weight matrix comparison (9,10). All possible shifts of the matrices are considered in the comparison, and the one giving the lowest dissimilarity score is reported. In addition, uninformative comparison results are filtered out. We consider a comparison result to be uninformative when only small parts of the matrices have been compared or when the parts compared are highly uninformative.

Another important advantage of our method is that we provide a detailed description of the identified similarities. Reverse-complement or partially overlapping high-scoring matches are readily identified. Along with the PWMs, we also provide an interface to an elaborate sequence annotation method (11), which is very convenient for many typical analyses of transcription factor binding behaviour.

In the future we hope to make this resource the basis for a unified collection of experimentally and computationally derived PWMs.

ACKNOWLEDGEMENTS

We thank Stein Aerts for providing us with the names of the PWMs in Transfac Public and Amit Sinha for help with the manuscript. Funding to pay the Open Access publication charges for this article was provided by the Max Planck Institute for Molecular Genetics.

Conflict of interest statement. None declared.

REFERENCES

1. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
2. Wingender,E. (2004) TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol.*, **4**, 55–61.
3. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
4. Fitzwater,T. and Polisky,B. (1996) A SELEX primer. *Methods Enzymol.*, **267**, 275–301.
5. Hanlon,S.E. and Lieb,J.D. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr. Opin. Genet. Dev.*, **14**, 697–705.
6. Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.
7. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
8. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
9. Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, **9**, 447–464.
10. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of cis -regulatory modules. *Bioinformatics*, **19** (Suppl. 2), II5–II14.
11. Rahmann,S., Müller,T. and Vingron,M. (2003) On the power of profiles for transcription factor binding site detection. *Statist. Appl. Genet. Mol. Biol.*, **2** article 7.

12. Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**, 211–218.
13. Schug,J. and Overton,G.C. (1997) TESS: Transcription Element Search Software on the WWW Technical Report CBIL-TR-1997-1001-v0.0. Computational Biology and Informatics Laboratory School of Medicine University of Pennsylvania.
14. Genuario,R.R. and Perry,R.P. (1996) The GA-binding protein can serve as both an activator and repressor of ribosomal protein gene transcription. *J. Biol. Chem.*, **271**, 4388–4395.
15. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.
16. Bae,S.J., Oum,J.H., Sharma,S., Park,J. and Lee,S.W. (2002) *In vitro* selection of specific RNA inhibitors of NFATc. *Biochem. Biophys. Res. Commun.*, **298**, 486–492.
17. Tsai,R.Y. and Reed,R.R. (1998) Identification of DNA recognition sequences and protein interaction domains of the multiple-Zn-finger protein Roaz. *Mol. Cell. Biol.*, **18**, 6447–6456.