



## PROMO: detection of known transcription regulatory elements using species-tailored searches

Xavier Messeguer<sup>1</sup>, Ruth Escudero<sup>1</sup>, Domènec Farré<sup>1</sup>,  
Oscar Núñez<sup>1</sup>, Javier Martínez<sup>1</sup> and M. Mar Albà<sup>2,\*</sup>

<sup>1</sup>Algorithmics and Genetics Group, Software Department, Universitat Politècnica de Catalunya, C/Jordi Girona Salgado, 1–3, 08034 Barcelona, Spain and <sup>2</sup>Virus Genomics and Bioinformatics, Wohl Virion Centre, Department of Immunology and Molecular Pathology, University College London, 46 Cleveland St, London W1T 4JF, UK

Received on June 5, 2001; revised on August 3, 2001; accepted on September 12, 2001

### ABSTRACT

**Summary:** We have developed a set of tools to construct positional weight matrices from known transcription factor binding sites in a species or taxon-specific manner, and to search for matches in DNA sequences.

**Availability:** PROMO can be accessed online at <http://www.lsi.upc.es/~alggen> under the research link.

**Supplementary information:** An example of the graphic interface (Figure 1) can be visualized at <http://www.lsi.upc.es/~alggen/recerca/promo/figuraBioinformatics.html>.

**Contact:** peypoch@lsi.upc.es; m.alba@ucl.ac.uk

One of the major challenges that follow the sequencing of genomes is to unravel the gene expression regulatory networks that operate in different types of cells. The transcription of a gene typically requires and is regulated by a number of cellular factors, that recognize and bind to short sequence motifs, in many cases located upstream of the gene coding sequence, in the so-called promoter and enhancer regions. Genes expressed in the same tissue or under similar conditions often share common regulatory motifs (Wasserman and Fickett, 1998), therefore the motifs found in a gene can be understood as a 'footprint' of its transcriptional regulatory mechanisms and to some extent gene function. The *in silico* prediction of potential regulatory sites is therefore a valuable tool to characterize new genes and to limit the amount of protein–DNA interactions to be tested experimentally.

Many binding sites for transcription factors have been experimentally identified and this information can be used to perform computational-based searches. A number of public databases store information on individual transcription factors and their binding sites, such as

TRANSFAC (Wingender *et al.*, 2001) or RegulonDB (Salgado *et al.*, 2001). Due to the intrinsic sequence variability of the motifs recognized by particular regulatory proteins appropriate representations of the sites are IUPAC consensi or positional weight matrices (Bucher, 1990). The latter store the frequency of the different nucleotides in the different positions of the motif and are generally considered superior as they are more specific and allow rating of the matches (Frech *et al.*, 1997). The TRANSFAC database (Wingender *et al.*, 2001) contains the largest available collection of eukaryotic factor-specific weight matrices, which can be used to search for potential matches in a DNA sequence of interest, for example by the MatInspector program (Quandt *et al.*, 1995).

The TRANSFAC collection of matrices is subdivided into very broad taxonomic groups (vertebrates, fungi, plants, insects and miscellaneous). The lack of flexibility in the taxonomy levels that can be considered may lead to problems in the interpretation of the results, specially when the binding sites have only been identified in a species which is distantly related to the one under study. In addition, searching with large collections of matrices may result in an increment in the number of false positives in the predictions, a general problem when attempting to identify short and variable sequences such as transcription factor binding sites. Bearing in mind these caveats we have developed a new approach to perform searches with weight matrices, which allows the user to tailor the searches to the species or group of species of interest. By selecting a particular species instead of a general group of organisms more specificity in the searches can be achieved. If few sites are known for the species under study selecting matrices from related species may still provide valuable information. Comparing different

\*To whom correspondence should be addressed.

species settings may be useful to analyze the cross-species conservation of particular known binding sites. Additional novel features of PROMO are the generation of the factor-specific matrices on the fly and the incorporation, as part of the output, of information on other genes which are known to be regulated by the subset of transcription factors that appear in the prediction.

PROMO has been written in C<sup>++</sup> and includes different modules, all available through a web server. The complete collection of TRANSFAC site, factor and gene entry files is used as a source of sequences and information. The species, or group of species, of interest is selected by the user. After the species selection weight matrices are automatically derived from at least three different binding sites per transcription factor, by anchoring the alignment of the relevant sequences on the completely conserved positions or 'core' of the binding site. An automata is then constructed which contains all the different possible subsequences that score above a given similarity threshold to any of the matrices. The similarity of a sequence to a matrix is calculated according to Quandt *et al.* (1995) and the default similarity threshold used by the program is 85% (or dissimilarity 15%). Exact matches of the query sequence to the automata represent putative transcription factor binding sites in the sequence. The two steps that the user is required to perform are: (1) 'SelectSpecies', select the species or taxonomic group of interest by using a taxonomic tree derived from the organism annotations in TRANSFAC site and factor entries and; (2) 'SearchSite', input a query sequence to search for matches to the matrices in any of the two strands. Other available options are 'ViewMatrices' and 'MatrixSpecificity'. The first one allows the visualization of the matrices that have been constructed including information on genes known to contain sites represented in the matrices. The second option is a Java applet for the comparison of the specificity of matrices corresponding to pre-defined taxonomic groups (see below for a definition of specificity). After steps 1 and 2 the program typically takes a few seconds to run and the results are presented online (example in supplementary material, Figure 1). The output includes the following: matches of the sequence to the factor-specific matrices in the corresponding sequence location, including the name of the factor that binds to the motif and dissimilarity percentage; expectation values of finding the different matches by chance alone, using a model with equiprobability of the four nucleotides, or a model with nucleotide frequency as in the query sequence and; information on the location of the predicted regulatory sites, either individually or combined, in other genes. The latter feature includes a graphical representation of the different sites in the regulatory regions of the genes, following the annotations in the TRANSFAC gene entries. The information on other genes may be very

useful as the observation of functional relatedness with the gene of interest may highlight particularly interesting hits.

When no species restriction is applied the number of matrices that PROMO generates is 452, derived from 4308 different sites. A variable number of matrices are created for different organism groups, for example 268 for animals, 26 for fungi, 19 for plants, 245 for vertebrates and 55 for humans. A simple way to measure and compare the specificity of different matrices is by defining the specificity for each position in the sequence as the distance between a vector representing the probability of each nucleotide and a vector where all nucleotides are equiprobable and the specificity of a matrix as a whole as the normalized average distance for all columns. Decreasing the taxonomic level under consideration leads to the expected increase in specificity when comparing matrices for the same transcription factor (same TRANSFAC entry), which will lead to a less noisy output. For example when we compare equivalent matrices from humans and vertebrates (55 different matrices), the human matrices are more specific ( $p < 10^{-2}$ ), the average specificity being 0.8 in contrast to 0.7 for the vertebrate matrices. Future developments we envisage are the explicit modelling of combinations of factor binding sites and the use of additional regulatory site databases.

## ACKNOWLEDGEMENT

This work was partially supported by project DGI BIO2001-2199.

## REFERENCES

- Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.
- Frech,K., Quandt,K. and Werner,T. (1997) Finding protein-binding sites in DNA sequences: the next generation? *Trends Biochem. Sci.*, **22**, 103–104.
- Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
- Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Díaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martínez,C. and Collado-Vives,J. (2001) RegulonDB (Version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
- Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
- Wingender,E., Chen,X., Fricke,R., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhäuser,R., Prüss,M., Schacherer,F., Thiele,S. and Urbach,S. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.