Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W423–W426***
doi:10.1093/nar/gki413

PromoterPlot: a graphical display of promoter similarities by pattern recognition

Alessandro Di Cara, Karsten Schmidt, Brian A. Hemmings and Edward J. Oakeley*

Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH-4058 Basel, Switzerland

Received February 22, 2005; Revised and Accepted March 21, 2005

ABSTRACT

PromoterPlot (http://promoterplot.fmi.ch) is a web-based tool for simplifying the display and processing of transcription factor searches using either the commercial or free TransFac distributions. The input sequence is a TransFac search (public version) or FASTA/Affymetrix IDs (local install). It uses an intuitive pattern recognition algorithm for finding similarities between groups of promoters by dividing transcription factor predictions into conserved triplet models. To minimize the number of false-positive models, it can optionally exclude factors that are known to be unexpressed or inactive in the cells being studied based on microarray or proteomic expression data. The program will also estimate the likelihood of finding a pattern by chance based on the frequency observed in a control set of mammalian promoters we obtained from Genomatix. The results are stored as an interactive SVG web page on our server.

INTRODUCTION

The initial objective of this work is to develop a viewing tool to display the results of TransFac searches in a graphic form. In this paper, we will describe a software tool we have developed for combining expression data with promoter analysis.

Promoter analysis is a process that has historically been very difficult to perform in higher eukaryotes (1). The first question one must consider is how exactly to define a promoter? In the context of this article, we will be using the definition that the 'core promoter' should occupy a region 500 bp upstream of the start of transcription (2). The challenge with this definition is how should we obtain the starts of transcription? The variable and sometimes large 5′-untranslated regions of mammalian messages make any reference to the start of translation rather weak, and the poor sequence conservation between starts (3) makes computer prediction a complex and sometimes unreliable task. We purchased the promoter resources for human, mouse and rat from Genomatix (http://www.genomatix.de) which, as of spring 2004, contained ~156 000 starts of transcription. Many of these had been experimentally mapped using oligo capping technology (4). Non-commercial resources also exist, most notably the Eukaryotic Promoter Database (http://www.epd.isb-sib.ch/) (5) and the various genome sequence repositories.

There are two major schools of thought regarding promoter analysis at the present time: first, the sequence-based approach where short regions of sequence conservation between regulatory sequences are assembled in an attempt to predict regions of micro-conservation that might be important in the control of gene expression, e.g. the MEME motif discovery program (6) which is a tool for discovering motifs in a group of related sequences. MEME was one of the first such programs and it strives to develop position-dependent probability matrices for finding every possible letter at each position in a putative pattern. The motifs found do not contain gaps but can be rather short so that gaps are modeled by the occurrence of additional motifs with un-conserved relative spacings. The size of these motifs is automatically calculated by the program. One further program is MotifSampler, which uses Gibbs sampling to assign a probability distribution to the chance of finding apparently conserved regions of sequence (7). These models do not require any prior knowledge of the underlying biology and as such it can be difficult to assess the mechanistic significance of any pattern found (8), and even when sequence conservation does occur it does not necessarily imply a conserved regulatory function. The second major approach for promoter analysis is the knowledge-based search for known transcription factor binding sites [reviewed in (9)]. This process relies on the collection of information from the scientific literature about the known binding sites from which a consensus target site is estimated. This effort is largely the work of the German

*To whom correspondence should be addressed. Tel: +41 61 697 6986; Fax: +41 61 697 3976; Email: edward.oakeley@fmi.ch
Present address:
Karsten Schmidt, The Clayton Foundation Laboratories for Peptide Biology, The Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA

company Biobase GmbH (http://www.biobase.de) through their Match program associated with the TransFac database. Match provides extremely detailed reports of potential binding sites in target sequences; however, the complexity of the answers returned can be daunting. The challenge here is to filter the data so that we can extract biologically useful models for hypothesis formulation.

The objective of this work is to create a simple web application which would use expression data or proteomics data to filter the list of potential transcription factors predicted by TransFac. We then developed an advanced pattern recognition algorithm to extract patterns of conserved factors in the promoters under investigation. Looking for single transcription factors does not provide any measure of the significance of findings above that given already in TransFac, instead modules of multiple transcription factors in a defined order have been shown to be critical for modulating the expression of genes (10,11). Higher-order complexes may be considered as pairs, triplets or greater numbers of factors in a module. Models composed of three factors have already been shown to be more selective than those of only two (12). Models with even higher complexity are overly stringent and appear automatically when searching for patterns of three. If multiple overlapping binding sites are predicted for the same transcription factor, these are concatenated in the display process into a single big binding site. This helps to simplify the display process without having a negative effect on the pattern searching process. The output is presented as an interactive web page using the Adobe SVG plug-in (Adobe Systems Inc., San Jose, CA) for Internet Explorer. Our tool is distributed in two forms: one for local installation, which can automate the TransFac queries and promoter sequence extraction using in-house information resources (source code available on request); and a second which is freely available on the Internet (http://promoterplot.fmi.ch). The input required for the Internet-accessible version is a TransFac search, saved as a text file, for your promoters. A free version of TransFac is available from Biobase (http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi) which may be used for this purpose.

DESIGN AND IMPLEMENTATION

PromoterPlot was developed as a web-based application. It was designed using Visual Studio 2003 (ASP.NET 1.1) and uses IIS as a web server (Microsoft Corporation, Redland, WA). We recommend that users access our program using Microsoft Internet Explorer 6 or later on the Windows operating system as some users have reported problems with other browsers. An Alchemi grid is used for processing (www.alchemi.net).

Modeling factor patterns

Users can either use our tool as a simple TransFac viewer to simply display the results of a TransFac search without further processing or they can look for conserved potentially regulatory modules within the promoters. To identify these conserved modules, we developed a pattern-searching algorithm, which works by scanning the promoter sequence for patterns composed of three transcription factors (ABC). So as not to collect patterns of very distant factors, which are less likely to interact with each other, the patterns are selected according to the maximum base pair distance between the first and the third factor (C–A), in a user definable manner (default = 100 bp). All of the patterns discovered in this way are collected in a single list. The patterns which passed the following restrictions are retained: (i) the same order of transcription factors; (ii) an identical strand distribution of the factors; (iii) a conserved spacing C–A and B–A with a user-defined 'wobble' for the spacing conservation (default ± 10 bp); and (iv) the pattern must occur in more than one of the promoters analyzed (default = 2). This process is summarized in Figure 1.

In our initial analysis, we found that the TransFac Match search, using the 'minimize false positives' (FPs) setting, predicts on average one binding site every 20 bp. This high stringency search is very tempting because it does not generate

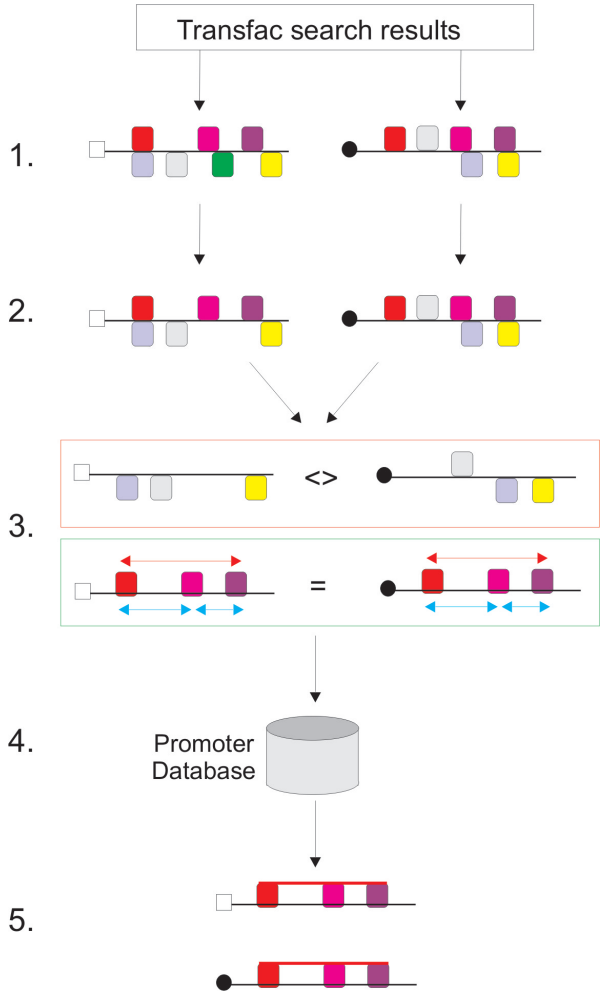

Figure 1. Pattern finding with PromoterPlot. 1: Transcription factor binding sites are predicted using the balance FP/FN option in TransFac. 2: Any factor for which there is evidence that it is not actually expressed or active in the tissue under investigation may be selectively removed from the analysis. 3: Patterns of three factors with conserved internal spacings and consistent binding strands which are found in two or more promoters are retained all others are discarded. 4: The frequency of the predicted patterns is compared with a database of mammalian promoters to estimate the probability of finding the observed results by chance. 5: The results are displayed as an interactive web page with matching genes from the database returned as Affymetrix IDs.

large numbers of binding sites; however, comparisons with published data are often not very good (13). Given that transcription factors bind cooperatively in nature it may be that sub-optimal sites are actually used *in vivo* which are stabilized by interaction with other proteins in a complex (13), but these sites often appear to be below the detection threshold for an FP search. The lower stringency search options of TransFac do a better job of finding the published interacting factors, with the minimize false negatives (FNs) option predicting as many as five binding sites per nucleotide and the balance FP/FN (SUM) option approximately one hit per nucleotide (if all vertebrate matrices are included in the search). However, when combined with our pattern discovery procedure, the SUM option appears to give a good balance between sensitivity and noise.

Clearly, the sequence of the promoter alone is unlikely to be sufficient for the effective modeling of transcription factor assembly because not every cell or developmental stage will necessarily use the same transcription factors to control expression in response to every possible stimulus (14). As not every transcription factor is in an active state in every cell, we can ask which transcription factors are expressed in the cell and what protein modifications might influence their activity? In our institute, we usually address this issue by looking directly at the expression of all transcription factors on Affymetrix microarrays. Any factor that is not expressed in any of the steps of the experiment is unlikely to be present to any significant quantity in regulatory complexes. Similarly, information may also be available from proteomics analysis of samples, which may reveal that certain factors are in inactive phosphorylation states or otherwise excluded from participation in the regulatory machinery. We decided that such information, when available, represents a useful resource for effective modeling and so users may optionally provide a file containing the names of transcription factors or TransFac matrix IDs which will be excluded from the analysis. Alternatively, during the course of expression data mining, we often find that transcription factors themselves are altered in their expression during an experiment. One might hypothesize that some of these changing factors have direct roles in the control of the other genes detected and a question that we are often presented with is 'can you show me what 'factor X' might be doing to my promoters?'. There are two solutions to this problem in our program: (i) if users have specific factors of interest in mind then they can provide a list of these and only patterns containing at least one of the named factors will be displayed; (ii) users can click on the factor names in the legend of the SVG display to see those promoters that contain patterns including specific factors.

Input data required

The potential inputs for PromoterPlot are TransFac result files (internet version), FASTA files (local version only, requires a local copy of TransFac) or Affymetrix IDs (local version only, requires a local copy of TransFac and a local promoter database). The FASTA headers should be kept as short as possible, but can contain start of transcription information in the following format: '>Some_Name' then '#transcription_start_position:start_color#'. Multiple starts can be supplied one after the other in this way. The final characters can be a short description of the gene, for example:

>12345_at#1985:gold#1951:silver#2001:bronze#MyGene

In this example, it would draw a promoter for the gene '>12345_at:MyGene' with a start colored 'gold' at position 1985, 'silver' at 1951 and 'bronze' at 2001. If you wish to use other colors, then you can also enter base-10 RGB values instead of the words in the format 'R, G, B'. We recommend that FASTA titles which do not use the above notation should take the format '>MyGeneName' and avoid using non-alphanumeric characters.

Because the results are active web pages containing server-side scripting, they are stored on our server for a maximum of 72 h for 'anonymous' searches (users get a session ID which they can use to access their data during this time). Users who would like to keep results for longer periods are encouraged to register (free) with a username and a password. Searches stored in this way are kept until the user deletes them or 3 months have passed without access.

Display of results

The primary objective of the display is to make each factor type visually distinct while retaining visual similarities between factors with similar names. It is clear that the same factor must always have the same appearance every time the program is run. The factors are represented by a filled box surrounded by a colored boarder (Figure 2). This two-color approach makes the process of discrimination much easier than with a single color. The fill colors are assigned automatically by taking the name of the transcription factor (e.g. STAT5) and converting the first three letters of the name into their corresponding ASCII values to give the red, green and blue color channels. If the name has fewer than three characters, then the missing characters are replaced by the ASCII code 00. The border color is generated in the same way but now the dominance is reversed so that the color is defined by the ASCII values of the final three characters of the name in the order blue, green and red. Thus, the border colors enable us to visually distinguish factors with very similar names.

The results are displayed in a web page that is composed of three frames: analysis, legend and output. Two viewing modes

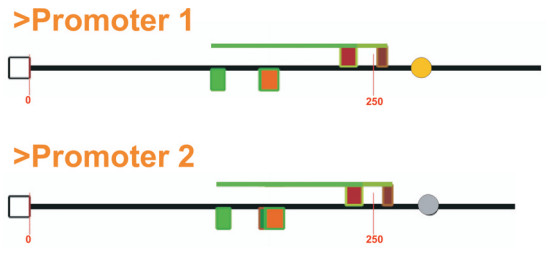

Figure 2. The start(s) of transcription are represented by circles colored gold, silver or bronze. Each factor is colored according to its name with the fill color based on the name stem and the border color based on the name ending. The factors and promoters are clickable and clicking provides information about pattern groups, binding sites and other genes which share the same patterns.

exist: the first mode (pattern view) displays only the patterns discovered by the pattern-finding algorithm. Clicking on a factor or its corresponding entry in the legend will display all of the patterns containing that factor. Additionally, clicking on the box at the 5′ end of each promoter will display all of the patterns found in the selected sequence in any other promoter. The second mode (factor view) displays the individual transcription factors (optionally even those that are not members of patterns). Clicking on a factor will display the location of that factor in all of the promoters. In addition, it will provide the target binding site sequences and positions of these factors in the output window. Thanks to the SVG plug-in, it is possible to zoom and pan the analysis window, facilitating the display of large numbers of sequences. All of the results are stored in a password-protected user folder on our server. In the event that a pattern query takes a very long time to complete, it will continue to run on our server even if the web browser is closed and users may log back in at a later date to view their results.

Assessing the specificity of the results

We have purchased the sequences of 156 000 mammalian promoters from Genomatix. When patterns are predicted by the program, their frequency of occurrence in the test sequence is calculated and compared with their frequency in the control database. If we define a null hypothesis that there is no difference between the two frequencies, then we can test this using a Chi-square test with one degree of freedom (15). Each pattern has a vertical pin associated with it which is clickable. Moving the mouse over the pin brings that pattern to the foreground and makes the others translucent. Clicking on a pin hides all patterns except for selected one and displays the Affymetrix IDs for any of the mammalian promoters from the database which also contain this pattern. The color of the pin indicates the Chi-square result. Those patterns that fail the test have red pins, patterns that pass with a P-value <0.05 have green pins and those that pass with a P-value <0.01 have blue. To partially compensate for small numbers of test promoters, we perform a Yates' correction for discontinuity to reduce the risk of type I errors. Clicking on a gene shows the Affymetrix IDs for all database genes with multiple conserved patterns.

DISCUSSION

Here, we present a new bioinformatic tool (PromoterPlot) for automating the extraction of promoter patterns from microarray-based expression data. Binding sites are predicted using Biobase's 'Match' program from the TransFac suite (16). The pattern prediction process may be filtered to exclude factors that are not believed to be active in the experiment. The patterns identified are displayed in a simple interactive web-based graphical interface and stored on the server for future use in a password-protected user directory. We feel that this may be a useful application for visualizing promoter comparisons and to assist in the identification of regions of potential interest before engaging in time-consuming biochemical characterization. The database hits predicted are also useful for validation. The program may be accessed online at http://promoterplot.fmi.ch.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the Novartis Research Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Figeys,D. (2004) Combining different 'omics' technologies to map and validate protein–protein interactions in humans. *Brief. Funct. Genomic Proteomic*, **2**, 357–365.
2. Werner,T. (1999) Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome*, **10**, 168–175.
3. Werner,T. (2002) Finding and decrypting of promoters contributes to the elucidation of gene function. *In Silico Biol.*, **2**, 23.
4. Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.
5. Perier,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The eukaryotic promoter database. *Nucleic Acids Res*, **28**, 302–303.
6. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 14–17 August, Stanford, CA. AAAI Press, Menlo Park, CA, pp. 28–36.
7. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouzé,P. and Moreau,Y. (2001) A higher order background model improves the detection of regulatory elements by Gibbs Sampling, *Bioinformatics*, **17**, 1113–1122.
8. Werner,T. (2000) Identification and functional modeling of DNA sequence elements of transcription. *Brief Bioinform.*, **1**, 372–380.
9. Frech,K., Quandt,K. and Werner,T. (1997) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.*, **22**, 103–104.
10. Christoffels,V.M., Grange,T., Kaestner,K.H., Cole,T.J., Darlington,G.J., Croniger,C.M. and Lamers,W.H. (1998) Glucocorticoid receptor, C/EBP, HNF3 and protein kinase A coordinately activate the glucocorticoid response unit of the carbamoylphosphate synthase I gene. *Mol. Cell. Biol.*, **18**, 6305–6315.
11. Klingenhoff,A., Frech,K., Quandt,K. and Werner,T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, **15**, 180–186.
12. Klingenhoff,A., Frech,K. and Werner,T. (2000) Regulatory modules shared within gene classes as well as across gene classes can be detected by the same *in silico* approach. *In Silico Biol.*, **2**, S17–S26.
13. Kel,A., Kel-Margoulis,O., Babenko,V. and Wingender,E. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J. Mol. Biol.*, **288**, 353–376.
14. Kel,A., Reymann,S., Matys,V., Nettesheim,P., Wingender,E. and Borlak,J. (2004) A novel computational approach for the prediction of networked transcription factors of aryl hydrocarbon-receptor-regulated genes. *Mol. Pharmacol.*, **66**, 1557–1572.
15. Fisher,S. (1962) Use of chi-square in simple cross-over designs. *J. Ment. Sci*, **108**, 406–410.
16. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.