W490–W494 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 4 May 2017*
doi: 10.1093/nar/gkx376

ConTra v3: a tool to identify transcription factor binding sites across species, update 2017

Łukasz Kreft[1,†], Arne Soete[2,3,†], Paco Hulpiau[2,3], Alexander Botzki[1], Yvan Saeys[2,4] and Pieter De Bleser[2,3,*]

[1]VIB Bioinformatics Core, Rijvischestraat 126 3R, 9052 Zwijnaarde-Ghent, Belgium, [2]VIB-UGent Center for Inflammation Research, Technologiepark 927, 9052 Zwijnaarde-Ghent, Belgium, [3]Department of Biomedical Molecular Biology, Ghent University, Technologiepark 927, 9052 Zwijnaarde-Ghent, Belgium and [4]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, S9, 9000 Gent, Belgium

Received February 20, 2017; Revised April 14, 2017; Editorial Decision April 22, 2017; Accepted April 25, 2017

ABSTRACT

Transcription factors are important gene regulators with distinctive roles in development, cell signaling and cell cycling, and they have been associated with many diseases. The ConTra v3 web server allows easy visualization and exploration of predicted transcription factor binding sites (TFBSs) in any genomic region surrounding coding or non-coding genes. In this updated version, with a completely re-implemented user interface using latest web technologies, users can choose from nine reference organisms ranging from human to yeast. ConTra v3 can analyze promoter regions, 5′-UTRs, 3′-UTRs and introns or any other genomic region of interest. Thousands of position weight matrices are available to choose from for detecting specific binding sites. Besides this visualization option, additional new exploration functionality is added to the tool that will automatically detect TFBSs having at the same time the highest regulatory potential, the highest conservation scores of the genomic regions covered by the predicted TFBSs and strongest co-localizations with genomic regions exhibiting regulatory activity. The ConTra v3 web server is freely available at http://bioit2.irc.ugent.be/contra/v3.

INTRODUCTION

Eukaryotic gene expression is transcriptionally regulated by the coordinated interaction of transcription factors (TF) with arrays of transcription factor binding sites (TFBSs) (1,2), also known as cis-regulatory modules and with each other (3). Knowing by which TFs a gene is regulated, is essential to reconstruct and model transcriptional regulatory networks governing biological processes such as the cell cycle or differentiation. Traditionally, regulation of genes by TFs is predicted by scanning promoter regions with positional weight matrices (PWMs) of known TFs, retaining putative binding sites scoring higher than an arbitrarily chosen cut-off for a given PWM. The results, however, include a large number of false positives due to the short (6–15 nucleotides) and degenerate nature of TFBSs. Phylogenetic footprinting is commonly and successfully used in combination with the PWM model to reduce its rate of false positive predictions. The main difficulty in this approach is to get correct alignments of regulatory elements in promoter regions that might have diverged during evolution (4). Taking into consideration that conservation of a TFBS among several species in a multiple alignment is neither proof nor required for functionality, the ConTra series of tools (5,6) have been designed to properly display predicted TFBSs in several possible alignments aiming to help the biologist seeking to generate or support a hypothesis. In this update, we describe the new features and expansions of the ConTra v3 web server. The ConTra v3 frontend has been completely re-implemented using latest web technologies to meet the required level of interactivity and user involvement. New features include a new layout, a simpler submission form, an on-screen guide and a dynamic TFBS viewer. The simplified design of the website layout facilitates user interaction and brings the main focus on the information provided. Its responsive design allows users of different screen sized devices to use the service without troubles. The form itself was simplified both visually and practically, allowing the user to have a better understanding of the required data and a clearer overview of the provided input. With the help of the on-screen interactive guide, the user is navigated step-by-step through the form submission process and is provided with sample data. Furthermore, the results page now contains not only static TFBS visualization images but also a dynamic TFBS viewer, where the user can select TFs and zoom in on the identified binding sites. With respect to the

*To whom correspondence should be addressed.Tel: +329 3313 693; Fax: +329 3313 609; Email: pieterdb@irc.vib-ugent.be
†These authors contributed equally to the paper as first authors.

backend, we updated the PWM libraries to more recent versions including the TRANSFAC database (update 2011.3) (7), the JASPAR core database (update 2016) (8), the cisBP Homo sapiens database (9) and the Taipale motifs collection for visualization (10). PWM libraries that were seldom used according to our web logs, such as the phyloFACTS database (11) and a collection of homeodomain PWMs derived from a protein binding microarray (12) have been removed. The other part of ConTra v3, the exploration part, predicts which TFs are most likely to bind to a given genomic region. In the previous versions of ConTra (5,6), the likelihood score for regulation of a gene by a TF, represented by its PWM, was obtained by an accumulation of the weights of the predicted TFBSs on the reference sequence. The weight of the predicted TFBS was determined by the number of species with a predicted TFBS for the same PWM at about the same position and the conservation extent of that position. The major drawback of the original implementations of the exploration part was the duration of the calculations involved: this could take from hours to days before results were obtained. As a consequence, this feature was not often used. Therefore, the exploration part was completely revised. In ConTra v3, PWM predicted TFBSs are ranked based on regulatory potential (13), conservation score (14) and the degree of overlap with genomic regions coinciding with regions of experimentally validated TF binding obtained from the comprehensive list of TF Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data released by the ReMap project (15). An overall rank for each PWM is calculated using rank product statistical analysis (16). The rank is based on aggregation of the ranked lists scoring the PWM based TFBSs predictions respectively on regulatory potential, degree of conservation and degree of overlap with genomic regions with demonstrated regulatory activity. A selection of up to 20 of these ranked PWMs can then be used as input for visualization analysis. The duration of the calculations involved are reduced to minutes, making this feature a lot more applicable and useful.

INPUT AND OUTPUT

Input

A typical ConTra v3 analysis consists of four steps. First, users have to choose whether they want to visualize or explore a gene of interest. For visualization, it is also necessary to indicate the reference species and the gene of interest. The second step lists a group of available transcripts for genes matching the search terms, from which one can be selected. For every gene, all possible RefSeq and Ensembl transcript variants are listed with a link to the genomic location in the respective genome browser. This way, genes with alternative promoters, UTRs or alternative intronic regions can be analyzed for regulatory differences. In step 3, different genomic regions of the selected transcript can be chosen (upstream, introns, 5′-UTR and 3′-UTR). The final step offers users an extensive choice of PWM motifs: up to 20 PWM motifs can be simultaneously taken into account for analysis. For exploration, one chooses the gene, the transcript, the region of interest and launches the exploration analysis.

Output

For the visualization part, results are split into alignment blocks allowing evaluation of the degree of binding site conservation. In the exploration part, a list of PWMs is given, ranked on the highest regulatory potential, highest conservation scores of the genomic regions covered by their predicted TFBSs and overlap with genomic regions with demonstrated regulatory activity. A selection of these high-scoring PWMs can then be used as input for visualization analysis.

Example

The cytokine interleukin-2 (IL2) is an important signaling protein in the human immune system. Regulation of the IL2 gene has been widely studied (17–19). We validated the ConTra v3 exploration mode by analyzing the IL2 promoter and comparing the results with known regulators from literature. In the first step we selected for exploration of the IL2 gene with *Homo sapiens* as reference species. In the next steps, we choose to analyze the promoter region (500-bp upstream) of the RefSeq transcript NM_000586. Filtering the predicted TFBS with a *q*-value of 0.1 and a PWM information content of at least 5 bits retrieves a list of 10 putative conserved binding sites of which at least for half of them there is experimental support. We selected NF-AT (V$NFAT_Q4_01), ELF1 (V$ELF1_Q6) and OCT (V$OCT_Q6) for visualization with a core and similarity stringency of 0.90 and 0.75 respectively. The results are shown in Figure 1. ConTra v3 successfully predicts the two known regulatory elements consisting of an Octamer (OCT), NF-AT and E26 transformation-specific (ETS) binding site and suggests the presence of a similar, third conserved module further upstream (Figure 1A and B). Fasta and feature color files, available for each alignment block, can be used to produce high quality figures with Jalview as shown in Figure 1C. The UCSC link on the result page in ConTra v3 maps the detected TFBS in the UCSC genome browser (Figure 1D). Also shown are additional ENCODE regulation tracks illustrating the co-localization of the third module with the presence of H3K4Me1 marks and open chromatin.

The online Supplementary Data S2 and 3 contain two case studies that explain step-by-step how to run an exploration and/or visualization analysis including information how users may change parameters and criteria according to their needs.

TECHNICAL DETAILS

Web tool

The web tool is hosted on a Linux CentOS 6.6 server with 16 GB of RAM, an Apache/2.2.15 web server and PHP 5.4.16. ConTra v3 was implemented using the AngularJS engine, the Bootstrap framework and the Bootstrap Material stylesheets. As database storage engine, MySQL was chosen. The on-screen guide is using Intro.js whereas the dynamic TFBS viewer is rendered as SVG. To track user activity Google Analytics was connected to all of the web pages. Each submitted job is queued on a beanstalkd queue

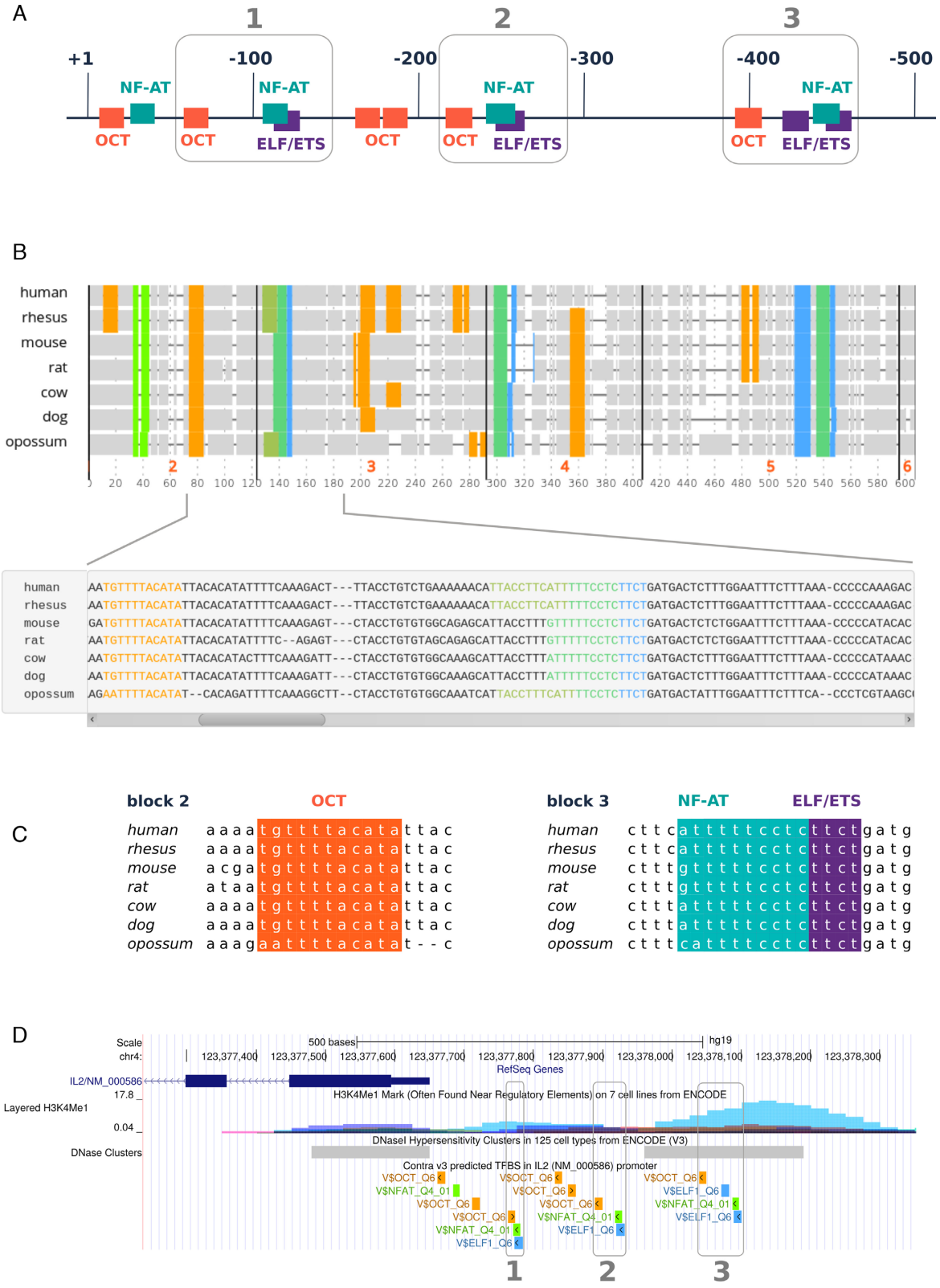

Figure 1. Analysis of the human IL2 promoter with ConTra v3 in exploration mode followed by visualization of a selection of the top-scoring results. (**A**) Overview of conserved binding sites for OCT, NF-AT and ETS/ELF transcription factors (TF) in the promoter region 500 bases upstream of the human interleukin-2 (IL2) gene. Gray boxes show repeats of regulatory regions. Regions 1 and 2 are experimentally supported (18,19). (**B**) Visualization of conserved TFBS across species. A user can choose to show or hide each species and TFBS individually. The alignment region can be zoomed in and out (top) and sites can be inspected at base level (bottom). (**C**) Alignment blocks can be downloaded as FASTA file with a corresponding feature color file to produce figures in several output formats using Jalview. (**D**) The detected binding sites can also be looked at, in a genomic context in the UCSC genome browser. Also shown are additional ENCODE regulation tracks illustrating the co-localization of the third module with the presence of H3K4Me1 marks and open chromatin.

(version 1.9.2). Workers, written in Perl (v5.10.1 × 86_64-linux-thread-multi), take jobs from this queue and process them.

Backend

The backend of ConTra v3 is programmed in a combination of Perl and R (http://www.r-project.org). The visualization part of ConTra v3 relies on the same algorithms implemented in ConTra v2 but uses several updated PWM libraries and multi-species multiple sequence alignments. Furthermore, the framework has been adapted to make inclusion of new PWM collections easier. Users are encouraged to suggest new PWM collections useful for their research.

The exploration part was in the previous versions slow. Therefore, it has been revised completely to make this feature much faster and hence more useful. One caveat remains: it is extremely difficult to predict which TFs regulate a gene of interest. The exploration part is primarily intended to give an idea of which TFs are more likely to bind to the genomic region of interest and to point to PWMs for which visualization of the predicted TFBSs could be interesting. If TFBS prediction is the primary concern, we direct the user to our PhysBinder web application (http://bioit.dmbr.ugent.be/physbinder) (20) that is likely to produce more reliable predictions. For exploration, one chooses the gene, the transcript, the region of interest and launches the exploration analysis. First, using the FIMO (21) application (default *P*-value cut-off: 0.0001) and the combined PWM libraries, TFBS predictions are made for every PWM. Next, the PWMs are ranked independently based on the cumulative scores of the regulatory potential scores of their TFBS predictions (13), the cumulative mean conservation scores of the genomic regions covered by the TFBS predictions (14) and the cumulative regulatory activity scores as a measure of the degree of overlap of the TFBS predictions with genomic regions coinciding with regions of experimentally validated TF binding, contained in the ReMap TF ChIP-Seq dataset (15).

The concept of the regulatory potential of a TF for a target gene was introduced by Tang *et al*. (13) to model the influence of each binding site on gene regulation as a function that decreases monotonically with increasing distance from the transcription start site (TSS) of the gene. Regulatory potential considers both the number of binding sites and their distances to the reported TSS of the putative target gene.

Mean conservation scores of the genomic regions covered by the TFBS predictions are obtained using the bigWigSummary tool from the UCSC genome browser with the phastConsElements100way table of the UCSC Genome Browser (http://genome.ucsc.edu) database. This table contains information about conserved elements identified by phastCons (14), a hidden Markov model-based method that estimates the probability that each nucleotide belongs to a conserved element, based on the multiple alignment. As it considers not only each individual alignment column, but also its flanking columns, PhastCons is effective for identifying conserved elements.

Finally, the regulatory activity scores of the predicted TFBSs are calculated by counting the number of times they intersect with genomic regions coinciding with regions of experimentally validated TF binding, contained in the ReMap TF ChIP-Seq dataset (15).

Rank product analysis (16) is used to select PWMs whose TFBS predictions simultaneously exhibit (i) the highest regulatory potential, (ii) the strongest conservation and (iii) the best overlap with genomic regions with demonstrated regulatory activity.

Using the exploration analysis of the human IL2 promoter region as an example we provide an extensive description of how to use and interpret regulatory potential, conservation score and regulatory activity score rankings in Supplementary Data S1.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: VIB-UGent Center for Inflammation Research, Ghent, Belgium.
Conflict of interest statement. None declared.

REFERENCES

1. Yanez-Cuna,J.O., Dinh,H.Q., Kvon,E.Z., Shlyueva,D. and Stark,A. (2012) Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res.*, **22**, 2018–2030.
2. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
3. Mathelier,A., Shi,W. and Wasserman,W.W. (2015) Identification of altered cis-regulatory elements in human disease. *Trends Genet.*, **31**, 67–76.
4. Fang,F. and Blanchette,M. (2006) FootPrinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.*, **34**, W617–W620.
5. Hooghe,B., Hulpiau,P., van Roy,F. and De Bleser,P. (2008) ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species. *Nucleic Acids Res.*, **36**, W128–W132.
6. Broos,S., Hulpiau,P., Galle,J., Hooghe,B., Van Roy,F. and De Bleser,P. (2011) ConTra v2: a tool to identify transcription factor binding sites across species, update 2011. *Nucleic Acids Res.*, **39**, W74–W78.
7. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
8. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
9. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
10. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
11. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic

discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
12. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Pena-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
13. Tang,Q., Chen,Y., Meyer,C., Geistlinger,T., Lupien,M., Wang,Q., Liu,T., Zhang,Y., Brown,M. and Liu,X.S. (2011) A comprehensive view of nuclear receptor cancer cistromes. *Cancer Res.*, **71**, 6940–6947.
14. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
15. Griffon,A., Barbier,Q., Dalino,J., van Helden,J., Spicuglia,S. and Ballester,B. (2015) Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.*, **43**, e27.
16. Heskes,T., Eisinga,R. and Breitling,R. (2014) A fast algorithm for determining bounds and accurate approximate *P*-values of the rank product statistic for replicate experiments. *BMC Bioinformatics*, **15**, 367–377.
17. Shaw,J.P., Utz,P.J., Durand,D.B., Toole,J.J., Emmel,E.A. and Crabtree,G.R. (1988) Identification of a putative regulator of early T cell activation genes. *Science*, **241**, 202–205.
18. Serfling,E., Avots,A. and Neumann,M. (1995) The architecture of the interleukin-2 promoter: a reflection of T lymphocyte activation. *Biochim. Biophys. Acta*, **1263**, 181–200.
19. Panagoulias,I., Georgakopoulos,T., Aggeletopoulou,I., Agelopoulos,M., Thanos,D. and Mouzaki,A. (2016) Transcription factor Ets-2 acts as a preinduction repressor of interleukin-2 (IL-2) transcription in naive T helper lymphocytes.*J. Biol. Chem.*, **291**, 26707–26721.
20. Broos,S., Soete,A., Hooghe,B., Moran,R., van Roy,F. and De Bleser,P. (2013) PhysBinder: improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties. *Nucleic Acids Res.*, **41**, W531–W534.
21. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.